A supplementary file for 'Estimation of High-Dimensional Mean Regression in Absence of Symmetry and Light-tail Assumptions'

Jianqing Fan, Quefeng Li, Yuyan Wang

October 18, 2015

A.1. Proof of Lemma 1.

First of all, it follows from Lemma 1 of Negahban, *et al.* (2012) that $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha \in \mathbb{C}_{\alpha\eta}$ on the event $\{\lambda_n \geq 2\left\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*_\alpha)\right\|_\infty\}$. Hence, we need to show that the event $\{\lambda_n \geq 2\left\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*_\alpha)\right\|_\infty\}$ holds with high probability. The latter will be established by using Bernstein's inequality along with the union bound.

The gradient of $\mathcal{L}_n$,

$$\nabla\mathcal{L}_n(\boldsymbol{\beta}^*_\alpha) = \frac{1}{n}\sum_{i=1}^n \frac{2}{\alpha}\psi[\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)]\mathbf{x}_i, \tag{1}$$

where $\psi(x) = x$, for $|x| \leq 1$; $\psi(x) = 1$, for $x > 1$; and $\psi(x) = -1$, for $x < -1$. Using $\alpha^{-1}|\psi(\alpha x)| \leq |x|$ and assumption (C3), we have

$$\begin{aligned}
\mathrm{E}\{2\alpha^{-1}\psi[\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)]x_{ij}\}^2 &\leq 4\,\mathrm{E}\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)^2 x_{ij}^2\} \\
&\leq 8\,\mathrm{E}\{(\epsilon_i^2 + |\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)|^2)x_{ij}^2\} \\
&= 8\,\mathrm{E}\{\mathrm{E}(\epsilon_i^2|\mathbf{x})x_{ij}^2 + |\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)|^2 x_{ij}^2\} \\
&\leq v,
\end{aligned}$$

where v is a constant depending on M_2 and κ_0 and the last inequality follows from a similar argument as in the proof of Theorem 1. By (C3) and that $|\psi(x)| \leq 1$, $\psi[\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)]x_{ij}$ is also sub-Gaussian. For any $k \geq 3$, using the relation between the kth moment and the second moment of sub-Gaussian random variables (Rivasplata, 2012),

$$\mathrm{E}\,|\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]x_{ij}|^k \leq \frac{k!}{2} L^{k-2}\,\mathrm{E}\,|\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]x_{ij}|^2,$$

where L is a constant depending on κ_0 only. Hence,

$$\mathrm{E}\,|2\alpha^{-1}\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]x_{ij}|^k \leq \frac{k!}{2}(2L/\alpha)^{k-2} v.$$

By Bernstein inequality (Proposition 2.9 of Massart and Picard (2007)) and note that $\mathrm{E}(\frac{2}{\alpha}\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]\mathbf{x}_i) = \mathbf{0}$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{2}{\alpha}\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]x_{ij}\right| \geq \sqrt{\frac{2vt}{n}} + \frac{2Lt}{\alpha n}\right) \leq 2\exp(-t).$$

Let $t = n\lambda_n^2/(32v)$ and observe that $\frac{2Lt}{\alpha n} \leq \sqrt{\frac{2vt}{n}}$ by the choice of λ_n and α. We have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{2}{\alpha}\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]x_{ij}\right| \geq \frac{\lambda_n}{2}\right) \leq 2\exp\left(-\frac{n\lambda_n^2}{32v}\right).$$

It then follows from union inequality that

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \frac{2}{\alpha}\psi[\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\alpha^*)]\mathbf{x}_i\right\|_\infty > \frac{\lambda_n}{2}\right) \leq 2\exp\left(-\frac{n\lambda_n^2}{32v} + \log p\right) \leq 2\exp(-c_0 n),$$

where $c_0 = \kappa_\lambda^2/(32v) - 1$ and without loss of generality we assume $\log p \leq n$. This completes the proof. □

A.2. Proof of Lemma 2.

Define set $A := \{(\boldsymbol{\beta}, \boldsymbol{\Delta}) : \|\boldsymbol{\beta}\|_2 \leq 4\rho_2 \text{ and } \|\boldsymbol{\Delta}\|_2 \leq 8\rho_2\}$, we first show that for any $(\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A$,

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \geq \frac{1}{n}\sum_{i=1}^n \varphi_{\tau\|\boldsymbol{\Delta}\|_2}(\mathbf{x}_i^T \boldsymbol{\Delta} I(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \leq T)), \tag{2}$$

for all $\alpha \le 1/(T + 8\tau\rho_2)$, where the thresholding function

$$\varphi_t(u) = u^2 I(|u| \le t/2) + (t - |u|)^2 I(t/2 \le |u| \le t), \tag{3}$$

$I(\cdot)$ is the indicator function and the thresholds T and τ will be chosen as in (8). From (8), we essentially need $\alpha \le c_u \rho_2^{-1}$, where c_u is a constant depending on the population level quantities κ_0, κ_l and κ_u only. The introduction of the thresholding function $\varphi_t(u)$ is to apply the contraction theorem of Ledoux and Talagrand (1991). Clearly, $\varphi_t(u) \le u^2$ and satisfies the Lipschitz condition with Lipschitz coefficient bounded by $2t$.

To show (2), if $|\mathbf{x}_i^T \mathbf{\Delta}| > \tau \|\mathbf{\Delta}\|_2$ or $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| > T$, the right hand side of (2) is 0. By convexity of the Huber loss function, (2) holds trivially. If $|\mathbf{x}_i^T \mathbf{\Delta}| \le \tau \|\mathbf{\Delta}\|_2$ and $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \le T$, then

$$|y_i - \mathbf{x}_i^T(\boldsymbol{\beta} + \mathbf{\Delta})| \le |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + |\mathbf{x}_i^T \mathbf{\Delta}| \le T + \tau \|\mathbf{\Delta}\|_2 \le T + 8\tau\rho_2 \le 1/\alpha,$$

and $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \le T \le 1/\alpha$. Since $\ell_\alpha(x) = x^2$ for $|x| \le 1/\alpha$, we have

$$\ell_\alpha(y_i - \mathbf{x}_i^T(\boldsymbol{\beta} + \mathbf{\Delta})) - \ell_\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - [\ell'_\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta})](\mathbf{x}_i^T \mathbf{\Delta}) = (\mathbf{x}_i^T \mathbf{\Delta})^2 \ge \varphi_{\tau\|\mathbf{\Delta}\|_2}(\mathbf{x}_i^T \mathbf{\Delta} I(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \le T)).$$

Therefore, (2) holds in any case. Using (2), to prove the lemma, it suffices to show that for any $(\boldsymbol{\beta}, \mathbf{\Delta}) \in A$, with high probability

$$\frac{1}{n\|\mathbf{\Delta}\|_2^2} \sum_{i=1}^n \varphi_{\tau\|\mathbf{\Delta}\|_2}(\mathbf{x}_i^T \mathbf{\Delta} I(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \le T)) \ge \kappa_1 - \kappa_1 \kappa_2 \sqrt{(\log p)/n} \frac{\|\mathbf{\Delta}\|_1}{\|\mathbf{\Delta}\|_2}.$$

From the definition (3), for any $d > 0$ and $z \in \mathbb{R}$, we have $\varphi_d(dz) = d^2 \varphi_1(z)$. Therefore, it is equivalent to show that for any $(\boldsymbol{\beta}, \mathbf{\Delta}) \in A' := \{(\boldsymbol{\beta}, \mathbf{\Delta}) : \|\boldsymbol{\beta}\|_2 \le 4\rho_2 \text{ and } \|\mathbf{\Delta}\|_2 = 1\}$, with high probability

$$\frac{1}{n} \sum_{i=1}^n \varphi_\tau(\mathbf{x}_i^T \mathbf{\Delta} I(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \le T)) \ge \kappa_1 - \kappa_1 \kappa_2 \sqrt{(\log p)/n} \|\mathbf{\Delta}\|_1. \tag{4}$$

To establish (4), let us consider its complementary event. Define

$$f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Delta} I(|y - \mathbf{x}^T \boldsymbol{\beta}| \le T), \quad g(\mathbf{x}) = \varphi_\tau(f(\mathbf{x})), \quad \text{and} \quad \mathbb{P}_n[g(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i).$$

Let $\mathbb{S}_2(1)$ be the unit sphere with L_2-radius one, and $\mathbb{S}_1(t)$ be the unit sphere with L_1-radius t, which is to be chosen later. The complementary event of (4) is given by

$$\Big\{\mathbb{P}_n[g(\mathbf{x})] < \kappa_1\{1 - \kappa_2\sqrt{(\log p)/n}\|\boldsymbol{\Delta}\|_1\}, \text{ for some } (\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A'\Big\}.$$

Our goal is to show that the probability of this event is very small, which is demonstrated through the following three steps.

(a) First, we show that with the choice of truncation T and τ as in (8), for any fixed $(\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A'$, we have

$$\mathrm{E}[g(\mathbf{x})] \ge \kappa_l/2. \tag{5}$$

(b) Second, with $Z(t) = \sup_{(\boldsymbol{\beta},\boldsymbol{\Delta})\in A' \cap \boldsymbol{\Delta} \in \mathbb{S}_1(t)} |\mathbb{P}_n[g(\mathbf{x})] - \mathrm{E}[g(\mathbf{x})]|$, we prove the tail probability of $Z(t)$ is bounded by

$$P(Z(t) \ge \kappa_l/4 + 40\tau^2\kappa_0 t\sqrt{(\log p)/n}) \le \exp(-c_1'' n - c_2'' t^2 \log p), \tag{6}$$

for each given t.

(c) Finally, we use a standard peeling argument (Alexander, 1987; Van de Geer, 2000) to establish

$$P\Big\{\exists(\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A' : Z(\|\boldsymbol{\Delta}\|_1) \ge \kappa_l/4 + 40\tau^2\kappa_0\|\boldsymbol{\Delta}\|_1\sqrt{(\log p)/n}\Big\} \le \exp(-c_1' n - c_2' \log p).$$

The result (c) together with (5) show that the probability of the complementary event of (4) with $\kappa_1 = \kappa_l/4$ and $\kappa_2 = 40\tau^2\kappa_0\kappa_1^{-1}$ is bounded by $\exp(-c_1' n - c_2' \log p)$, which completes the proof.

We first prove (5). In fact, by condition (C2), for any $(\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A'$, $\mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2] \ge \kappa_l\|\boldsymbol{\Delta}\|_2^2 = \kappa_l$.

So, it suffices to show that $\mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 - g(\mathbf{x})] \le \kappa_l/2$.

Note that, $g(\mathbf{x}) = (\mathbf{x}^T\boldsymbol{\Delta})^2$ for all $\mathbf{x}$ such that $|y - \mathbf{x}^T\boldsymbol{\beta}| \le T$ and $|\mathbf{x}^T\boldsymbol{\Delta}| \le \tau/2$. Therefore, we have

$$\mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 - g(\mathbf{x})] \le \mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 I(|y - \mathbf{x}^T\boldsymbol{\beta}| > T)] + \mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 I(|\mathbf{x}^T\boldsymbol{\Delta}| > \tau/2)]. \tag{7}$$

To bound the first term on the right hand side of (7), it follows from the Cauchy-Schwartz inequality that

$$\mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 I(|y - \mathbf{x}^T\boldsymbol{\beta}| > T)] \le [\mathrm{E}(\mathbf{x}^T\boldsymbol{\Delta})^4]^{1/2}[P(|y - \mathbf{x}^T\boldsymbol{\beta}| > T)]^{1/2}.$$

Since $\mathbf{x}^T\boldsymbol{\Delta}$ is sub-Gaussian with parameter at most κ_0^2 by assumption (C3), we have $\mathrm{E}(\mathbf{x}^T\boldsymbol{\Delta})^4 \le 16\kappa_0^4$. Meanwhile, it follows from the Chebyshev inequality that for any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2 \le 4\rho_2$,

$$\begin{aligned}
T^2 P(|y - \mathbf{x}^T\boldsymbol{\beta}| > T) &\le \mathrm{E}[(y - \mathbf{x}^T\boldsymbol{\beta})^2] \\
&\le 2\,\mathrm{E}\,\epsilon^2 + 2\,\mathrm{E}[\mathbf{x}^T(\boldsymbol{\beta}^* - \boldsymbol{\beta})]^2 \\
&\le 2\sqrt{M_2} + 34\kappa_u\rho_2^2 \\
&\le 36\kappa_u\rho_2^2.
\end{aligned}$$

where in the last inequality, we assume without loss of generality $\rho_2 \ge M_2^{1/4}\kappa_u^{-1/2}$. To bound the second term on the right hand side of (7), by the concentration inequality of sub-Gaussian variables, we have

$$P(|\mathbf{x}^T\boldsymbol{\Delta}| > \tau/2) \le 2\exp\{-\tau^2/(8\kappa_0^2)\}.$$

Then, by choosing T and τ as

$$T = 96\kappa_0^2\kappa_l^{-1}\kappa_u^{1/2}\rho_2 \qquad \text{and} \qquad \tau = \max\{4\kappa_0\log^{1/2}(12\kappa_l^{-1}\kappa_0^2), 1\}, \tag{8}$$

we have

$$\mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 I(|y - \mathbf{x}^T\boldsymbol{\beta}| \ge T)] \le \frac{\kappa_l}{4} \qquad \text{and} \qquad \mathrm{E}[(\mathbf{x}^T\boldsymbol{\Delta})^2 I(|\mathbf{x}^T\boldsymbol{\Delta}| \ge \tau/2)] \le \frac{\kappa_l}{4}.$$

Hence, (5) follows.

Next, we give the tail bound as in (b). Indeed, for any $(\boldsymbol{\beta}, \boldsymbol{\Delta}) \in A'$, we have $\|g\|_\infty \le \tau^2$. Therefore, by Massart concentration inequality (Theorem 14.2 of Bühlmann and Van De Geer (2011)), for any $z > 0$, we have $P(Z(t) \ge \mathrm{E}\, Z(t) + z) \le \exp(-\frac{nz^2}{32\tau^4})$. By choosing $z = \kappa_l/4 + 16\tau^2\kappa_0 t\sqrt{(\log p)/n}$, we have

$$P(Z(t) \ge \mathrm{E}\, Z(t) + z) \le \exp\Big(-\frac{n\kappa_l^2}{512\tau^4} - 8\kappa_0^2 t^2 \log p\Big). \tag{9}$$

Next, we bound $\mathrm{E}\, Z(t)$. Let $\{\omega_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables. A symmetrization theorem (Theorem 14.3 of Bühlmann and Van De Geer (2011)) yields

$$\mathrm{E}[Z(t)] \le 2\,\mathrm{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\Delta})\in A'\cap\boldsymbol{\Delta}\in\mathbb{S}_1(t)} \Big|\frac{1}{n}\sum_{i=1}^n \omega_i g(\mathbf{x}_i)\Big|\right] = 2\,\mathrm{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\Delta})\in A'\cap\boldsymbol{\Delta}\in\mathbb{S}_1(t)} \Big|\frac{1}{n}\sum_{i=1}^n \omega_i \varphi_\tau(f(\mathbf{x}_i))\Big|\right].$$

By definition, the function φ_τ is Lipschitz with parameter at most $2\tau \le 2\tau^2$ and $\varphi_\tau(0) = 0$. Therefore, by the Ledoux-Talagrand contraction theorem (Ledoux and Talagrand (1991), p.112), we have

$$\begin{aligned}
\mathrm{E}[Z(t)] &\le 8\tau^2\,\mathrm{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\Delta})\in A'\cap\boldsymbol{\Delta}\in\mathbb{S}_1(t)} \Big|\frac{1}{n}\sum_{i=1}^n \omega_i f(\mathbf{x}_i)\Big|\right] \\
&= 8\tau^2\,\mathrm{E}\left[\sup_{(\boldsymbol{\beta},\boldsymbol{\Delta})\in A'\cap\boldsymbol{\Delta}\in\mathbb{S}_1(t)} \Big|\frac{1}{n}\sum_{i=1}^n \omega_i \mathbf{x}_i^T\boldsymbol{\Delta} I(|y_i - \mathbf{x}_i^T\boldsymbol{\beta}| \le T)\Big|\right] \\
&\le 8\tau^2 t\,\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \omega_i \mathbf{x}_i h(y_i, \mathbf{x}_i)\right\|_\infty,
\end{aligned}$$

where $h(y_i, \mathbf{x}_i) = \sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2 \le 4\rho_2} I(|y_i - \mathbf{x}_i^T\boldsymbol{\beta}| \le T)$. Since the variables $\{x_{ij}\}_{i=1}^n$ are zero-mean i.i.d. sub-Gaussian with parameter at most κ_0^2, ω_i and $h(y_i, \mathbf{x}_i)$ are bounded, $\{\omega_i x_{ij} h(y_i, \mathbf{x}_i)\}_{i=1}^n$ is also sub-Gaussian. Since $\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \omega_i \mathbf{x}_i h(y_i, \mathbf{x}_i)\right\|_\infty$ is the maxima of p such terms, known bounds on the expectation of sub-Gaussian maxima (e.g. see Ledoux and Talagrand (1991), p.79) yield

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \omega_i \mathbf{x}_i h(y_i, \mathbf{x}_i)\right\|_\infty \le 3\kappa_0\sqrt{(\log p)/n}.$$

Hence,

$$\mathrm{E}[Z(t)] \leq 24\tau^2\kappa_0 t\sqrt{(\log p)/n}. \tag{10}$$

Combining (9) and (10), we have

$$P\Big(Z(t) \geq \kappa_l/4 + 40\tau^2\kappa_0 t\sqrt{(\log p)/n}\Big) \leq \exp(-c_1''n - c_2''t^2\log p),$$

where constants c_1'' and c_2'' depends on κ_l and κ_0 only. This result holds for each given t.

Next, we furnish the peeling argument in (c). Let $h(\|\mathbf{\Delta}\|_1) = \kappa_l/8 + 20\tau^2\kappa_0\|\mathbf{\Delta}\|_1\sqrt{(\log p)/n}$ and $B = \{\exists(\boldsymbol{\beta}, \mathbf{\Delta}) \in A' : Z(\|\mathbf{\Delta}\|_1) \geq 2h(\|\mathbf{\Delta}\|_1)\}$. Since $h(\|\mathbf{\Delta}\|_1) \geq \kappa_l/8$, the set can be covered by partition $\{B_m\}_{m=1}^{\infty}$ with $B_m = \{(\boldsymbol{\beta}, \mathbf{\Delta}) \in A' : 2^{m-4}\kappa_l \leq h(\|\mathbf{\Delta}\|_1) \leq 2^{m-3}\kappa_l\}$. Thus, by union bound,

$$\begin{aligned}
P(B) &\leq \sum_{m=1}^{\infty} P(\mathbf{\Delta} \in B_m \text{ such that } Z(\|\mathbf{\Delta}\|_1) \geq 2h(\|\mathbf{\Delta}\|_1)) \\
&\leq \sum_{m=1}^{\infty} P(Z(\|\mathbf{\Delta}\|_1) \geq 2^{m-3}\kappa_l)
\end{aligned}$$

since $h(\|\mathbf{\Delta}\|_1) \geq 2^{m-4}\kappa_l$ for $\mathbf{\Delta} \in B_m$. By letting $2^{m-3}\kappa_l = \kappa_l/4 + 40\tau^2\kappa_0 t\sqrt{(\log p)/n}$ as in (6) and solving for t, by (6), we obtain

$$\begin{aligned}
P(B) &\leq \sum_{m=1}^{\infty} \exp\left(-c_1''n - \frac{c_2''\kappa_l^2(2^{m-1}-1)^2 n}{\tau^4\kappa_0^2}\right) \\
&\leq \exp(-c_1''n) + \sum_{m=2}^{\infty} \exp\left(-c_1''n - \frac{c_2''n\kappa_l^2 2^{2m-4}}{\tau^4\kappa_0^2}\right) \\
&\leq c_1' \exp(-c_2'n),
\end{aligned}$$

where the last inequality follows from sum of geometric series. □

A.3. Proof of Lemma 3.

Note that,

$$R_q \geq \sum_{j=1}^{p} |\beta^*_{\alpha,j}|^q \geq \sum_{j \in S_{\alpha\eta}} |\beta^*_{\alpha,j}|^q \geq \eta^q |S_{\alpha\eta}|. \tag{11}$$

Therefore, $|S_{\alpha\eta}| \leq \eta^{-q} R_q$. Let $S^c_{\alpha\eta} = \{1, 2, \ldots, p\} \backslash S_{\alpha\eta}$, we have

$$\|\boldsymbol{\beta}^*_{S^c_{\alpha\eta}}\|_1 = \sum_{j \in S^c_{\alpha\eta}} |\beta^*_{\alpha,j}| = \sum_{j \in S^c_{\alpha\eta}} |\beta^*_{\alpha,j}|^q |\beta^*_{\alpha,j}|^{1-q} \leq R_q \eta^{1-q}. \tag{12}$$

Hence, for any $\boldsymbol{\Delta} \in \mathbb{C}_{\alpha\eta}$, we have

$$\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_{S_{\alpha\eta}}\|_1 + \|\boldsymbol{\Delta}_{S^c_{\alpha\eta}}\|_1 \leq 4\|\boldsymbol{\Delta}_{S_{\alpha\eta}}\|_1 + 4\|\boldsymbol{\beta}^*_{\alpha, S^c_{\alpha\eta}}\|_1.$$

By the Cauchy-Schwartz inequality and (12), we can bound further that

$$\|\boldsymbol{\Delta}\|_1 \leq 4\sqrt{|S_{\alpha\eta}|}\|\boldsymbol{\Delta}\|_2 + 4R_q\eta^{1-q} \leq 4R_q^{1/2}\eta^{-q/2}\|\boldsymbol{\Delta}\|_2 + 4R_q\eta^{1-q}.$$

From Theorem 1, $\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2 \leq d_1\alpha^{k-1}$. As we finally need α to be small, without loss of generality, we assume $\|\boldsymbol{\beta}^*_\alpha\|_2 \leq 4\rho_2$. In addition, we assume $\rho_2 \geq 1/8$. It then follows from Lemma 2 that

$$\begin{aligned}
\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}^*_\alpha) &\geq \kappa_1\|\boldsymbol{\Delta}\|_2\{\|\boldsymbol{\Delta}\|_2 - \kappa_2\sqrt{(\log p)/n}[4R_q^{1/2}\eta^{-q/2}\|\boldsymbol{\Delta}\|_2 + 4R_q\eta^{1-q}]\} \\
&= \left(\kappa_1 - 4\kappa_1\kappa_2R_q^{1/2}\eta^{-q/2}\sqrt{(\log p)/n}\right)\|\boldsymbol{\Delta}\|_2^2 - 4\kappa_1\kappa_2R_q\eta^{1-q}\sqrt{(\log p)/n}.
\end{aligned}$$

With $\lambda_n = \kappa_\lambda\sqrt{(\log p)/n}$ and $\eta = \lambda_n$, it holds that

$$4\kappa_1\kappa_2R_q^{1/2}\eta^{-q/2}\sqrt{\frac{\log p}{n}} = 4\kappa_1\kappa_2R_q^{1/2}{\kappa_\lambda}^{-q/2}\left(\frac{\log p}{n}\right)^{(1-q)/2},$$

which is no larger than $\kappa_1/2$ under assumption (2.7). On the other hand,

$$4R_q\kappa_1\kappa_2\eta^{1-q}\sqrt{\frac{\log p}{n}} = 4R_q\kappa_1\kappa_2{\kappa_\lambda}^{1-q}\left(\frac{\log p}{n}\right)^{1-(q/2)}.$$

Therefore, RSC holds with $\kappa_{\mathcal{L}} = \frac{\kappa_1}{2}$ and $\tau^2_{\mathcal{L}} = 4R_q\kappa_1\kappa_2{\kappa_\lambda}^{1-q}(\frac{\log p}{n})^{1-(q/2)}$. □

A.4. Proof of Lemma 4.

It follows from Lemma 2 that

$$\delta\mathcal{L}_n(\boldsymbol{\Delta},\boldsymbol{\beta}) \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_1\kappa_2\|\boldsymbol{\Delta}\|_2\|\boldsymbol{\Delta}\|_1\sqrt{(\log p)/n}.$$

Using the fact that $ab \leq (a^2+b^2)/2$, we conclude that

$$\delta\mathcal{L}_n(\boldsymbol{\Delta},\boldsymbol{\beta}) \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \left(\frac{1}{2}\kappa_1\|\boldsymbol{\Delta}\|_2^2 + \frac{1}{2}\kappa_1\kappa_2^2\|\boldsymbol{\Delta}\|_1^2\left(\frac{\log p}{n}\right)\right).$$

Therefore, (3.2) holds with $\gamma_l = \kappa_1$ and $\tau_l = \kappa_1\kappa_2^2(\log p)/(2n)$. Meanwhile, we have

$$\delta\mathcal{L}_n(\boldsymbol{\Delta},\boldsymbol{\beta}) \leq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\boldsymbol{\Delta})^2.$$

Under the sub-Gaussianity assumption (C3), it follows from some existing work (e.g. page 18 of Loh and Wainwright (2013)) that, with probability greater than $1-c_1\exp(-c_2 n)$, it holds that

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\boldsymbol{\Delta})^2 \leq \kappa_u\left(\frac{3}{2}\|\boldsymbol{\Delta}\|_2^2 + \frac{\log p}{n}\|\boldsymbol{\Delta}\|_1^2\right),$$

where c_1 and c_2 are some generic constants. Hence, (3.3) holds with $\gamma_u = 3\kappa_u$ and $\tau_u = \kappa_u(\log p)/n$.

□

A.5. Proof of Theorem 4.

We prove the theorem by the following two steps:

(a) We first show that, for any $\delta^2 \geq \varepsilon^2/(1-\kappa)$, $\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}}) \leq \delta^2$, for all t greater than the right hand side of (15), where $\kappa \in [0,1)$ is a contraction constant and ε is a tolerance parameter, which will be given in (16) and (17), respectively.

(b) We use RSC condition (3.2) to transform the upper bound of $\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}})$ into the upper bound of $\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2$.

For step (a), by the choice of initial value, we have $\|\widehat{\boldsymbol{\beta}}^0 - \widehat{\boldsymbol{\beta}}\|_2 \leq \|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}^*\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 2\rho_2$, where we assume the sample size n is large enough to guarantee $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \rho_2$. It then follows from Lemma 2 of Loh and Wainwright (2013) that $\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2 \leq 2\rho_2$ for all $t \geq 0$. Therefore, $\|\widehat{\boldsymbol{\beta}}^t\|_2 \leq \|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + \|\boldsymbol{\beta}^*\|_2 \leq 4\rho_2$. Hence, Lemma 4 guarantees that RSC/RSM conditions hold for all $\widehat{\boldsymbol{\beta}}^t$, $t \geq 0$. Since our loss function is convex, we apply Theorem 2 of Agarwal, Negahban, and Wainwright (2012). In order for our proof to be self-contained, we cite their theorem as the follows:

[Theorem 2 of Agarwal, Negahban, and Wainwright (2012)] Suppose for any data set Z_1^n, the loss function $\mathcal{L}_n(\cdot, Z_1^n)$ is convex and differentiable and the regularizer $\mathcal{R}$ is a norm. Consider the optimization problem of $\widehat{\theta} = \operatorname{argmin}_{\mathcal{R}(\theta)\leq\rho}\{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n\mathcal{R}(\theta)\}$ for a radius ρ such that θ^* is feasible, where $\theta^* = \operatorname{argmin} \mathrm{E}\,\mathcal{L}_n(\theta; Z_1^n)$, and a regularization parameter λ_n satisfying bound

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\theta^*)), \tag{13}$$

where $\mathcal{R}^*$ is the dual norm of the regularizer. In addition, suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters (γ_l, τ_l) and (γ_u, τ_u), respectively. Let $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$ be any $\mathcal{R}$-decomposable pair of subspaces such that

$$\kappa = \left\{1 - \frac{\bar{\gamma}_l}{4\gamma_u} + \frac{64\Psi^2(\bar{\mathcal{M}})\tau_u}{\bar{\gamma}_l}\right\}\xi \in [0,1) \qquad \text{and} \qquad \frac{32\rho}{1-\kappa}\xi\chi \leq \lambda_n, \tag{14}$$

where $\Psi(\bar{\mathcal{M}}) = \sup_{\theta\in\bar{\mathcal{M}}\setminus\{0\}} \mathcal{R}(\theta)/\|\theta\|_2$, $\bar{\gamma}_l = \gamma_l - 64\tau_l\Psi^2(\bar{\mathcal{M}})$, $\xi = (1 - 64\tau_u\bar{\gamma}_l^{-1}\Psi^2(\bar{\mathcal{M}}))^{-1}$, and $\chi = 2\left(\bar{\gamma}_l/(4\gamma_u) + 128\tau_u\bar{\gamma}_l^{-1}\Psi^2(\bar{\mathcal{M}})\right)\tau_l + 8\tau_u + 2\tau_l$. Denote $\varepsilon^2 = 8\xi\chi\left(6\Psi(\bar{\mathcal{M}})\|\widehat{\theta} - \theta^*\|_2 + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\right)^2$, where $\Pi_{\mathcal{M}^\perp}(\theta^*)$ is the projection of θ^* onto $\mathcal{M}^\perp$. Then for any $\delta^2 \geq \varepsilon^2/(1-\kappa)$, we have $\phi_n(\widehat{\theta}^t) - \phi_n(\widehat{\theta}) \leq \delta^2$ for all

$$t \geq \frac{2\log((\phi_n(\theta^0) - \phi_n(\widehat{\theta}))/\delta^2)}{\log(1/\kappa)} + \log_2\log_2\left(\frac{\rho\lambda_n}{\delta^2}\right)\left(1 + \frac{\log 2}{\log(1/\kappa)}\right), \tag{15}$$

where $\phi_n(\theta) = \mathcal{L}_n(\theta; Z_1^n) + \lambda_n\mathcal{R}(\theta)$, $\widehat{\theta}^t$ is the solution by the gradient descent algorithm after t^{th} iteration, and θ^0 is the initial value of θ.

In fact, Theorem 2 of Agarwal, Negahban, and Wainwright (2012) is a deterministic statement for all choices of pairs $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$. From Lemma 1 and Lemma 4, we have shown that with our choice of λ_n, the RA-quadratic loss function satisfy (13) and RSC/RSM with probability at least $1 - c_1 \exp(-c_2 n)$. Hence, Theorem 2 of Agarwal, Negahban, and Wainwright (2012) applies to our problem with high probability. We further choose the pair $(\mathcal{M}, \bar{\mathcal{M}}^{\perp}) = (S_{\alpha\eta}, S_{\alpha\eta}^c)$ and give the explicit expression of constants for our problem as the follows:

$$\kappa = \left\{1 - \frac{\bar{\gamma}_l}{4\gamma_u} + \frac{64\kappa_u |S_{\alpha\eta}| \frac{\log p}{n}}{\bar{\gamma}_l}\right\} \left(1 - \frac{64\kappa_u |S_{\alpha\eta}| \frac{\log p}{n}}{\bar{\gamma}_l}\right)^{-1}, \tag{16}$$

$$\varepsilon^2 = 8\xi\chi \left(6\sqrt{|S_{\alpha\eta}|}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2 + 8\|\boldsymbol{\beta}^*_{S^c_{\alpha\eta}}\|_1\right)^2, \tag{17}$$

where $\bar{\gamma}_l = \kappa_1 - 32\kappa_1\kappa_2^2 |S_{\alpha\eta}|(\log p)/n$, $\xi = \{1 - 64\kappa_u |S_{\alpha\eta}|(\log p)/(n\bar{\gamma}_l)\}^{-1}$, and $\chi = 2\{\bar{\gamma}_l/(4\gamma_u) + 128\tau_u |S_{\alpha\eta}|/\bar{\gamma}_l + 1\}\tau_l + 8\tau_u$. It remains to check (14). By (16), $\kappa \in [0, 1)$ is equivalent to requiring

$$|S_{\alpha\eta}| \frac{\log p}{n} < \frac{\bar{\gamma}_l^2}{1536\kappa_u^2}. \tag{18}$$

With $\eta = \lambda_n$, it follows from (11) that

$$|S_{\alpha\eta}| \frac{\log p}{n} \leq R_q \eta^{-q} \frac{\log p}{n} \leq \kappa_\lambda^{-q} R_q \left(\frac{\log p}{n}\right)^{1-(q/2)}.$$

Hence, (18) holds when n is sufficiently large. Moreover, from (14) we need

$$\lambda_n \geq \frac{32\rho}{1-\kappa} \left(1 - \frac{64\kappa_u |S_{\alpha\eta}| \frac{\log p}{n}}{\bar{\gamma}_l}\right)^{-1} \left[1 + \kappa_1 \kappa_2^2 \left(\frac{\bar{\gamma}_l}{12\kappa_u} + \frac{128\kappa_u |S_{\alpha\eta}| \frac{\log p}{n}}{\bar{\gamma}_l}\right) + 8\kappa_u\right] \frac{\log p}{n},$$

which is satisfied under the stated assumption. It then follows from Theorem 2 of Agarwal, Negahban, and Wainwright (2012) that, for any $\delta^2 \geq \varepsilon^2/(1-\kappa)$, $\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}}) \leq \delta^2$, for all iterations t greater than the right hand side of (15).

For step (b), it follows from the RSC condition that

$$\mathcal{L}_n(\widehat{\boldsymbol{\beta}}^t) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) - [\nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}})]^T(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}) \geq \frac{\gamma_l}{2}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 - \tau_l\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1^2.$$

Then we have

$$\begin{aligned}\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}}) &= \mathcal{L}_n(\widehat{\boldsymbol{\beta}}^t) - \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda_n(\|\widehat{\boldsymbol{\beta}}^t\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \\ &\geq [\nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}})]^T(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}) + \lambda_n(\|\widehat{\boldsymbol{\beta}}^t\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) + \frac{\gamma_l}{2}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 - \tau_l\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1^2.\end{aligned}$$

Since $\widehat{\boldsymbol{\beta}}$ is the minimizer of $\phi(\boldsymbol{\beta})$, by the first-order condition, $[\nabla \mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda_n \nabla \|\widehat{\boldsymbol{\beta}}\|_1]^T(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}) \geq 0$. Therefore,

$$\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}}) \geq -\lambda_n[\nabla\|\widehat{\boldsymbol{\beta}}\|_1]^T(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}) + \lambda_n(\|\widehat{\boldsymbol{\beta}}^t\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) + \frac{\gamma_l}{2}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 - \tau_l\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1^2.$$

By the convexity of the L_1-norm, $\|\widehat{\boldsymbol{\beta}}^t\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 - [\nabla\|\widehat{\boldsymbol{\beta}}\|_1]^T(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}) \geq 0$. Hence,

$$\phi(\widehat{\boldsymbol{\beta}}^t) - \phi(\widehat{\boldsymbol{\beta}}) \geq \frac{\gamma_l}{2}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 - \tau_l\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1^2. \tag{19}$$

Next, we bound $\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1$. It follows from Lemma 3 of Agarwal, Negahban, and Wainwright (2012) that

$$\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1 \leq 2\left(2\sqrt{S_{\alpha\eta}}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2 + 4\sqrt{|S_{\alpha\eta}|}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2 + 4\|\boldsymbol{\beta}^*_{\alpha, S^c_{\alpha\eta}}\|_1 + \delta^2/\lambda_n\right),$$

where δ is defined as in (a). Then, by the Cauchy-Schwartz inequality,

$$\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_1^2 \leq 16\left(4|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 + 16|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2^2 + 16\|\boldsymbol{\beta}^*_{\alpha, S^c_{\alpha\eta}}\|_1^2 + \delta^4/\lambda_n^2\right). \tag{20}$$

Equations (19) and (20) together with results in (a) imply that,

$$\delta^2 \geq \frac{\gamma_l}{2}\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 - 16\tau_l\left(4|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 + 16|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2^2 + 16\|\boldsymbol{\beta}^*_{\alpha, S^c_{\alpha\eta}}\|_1^2 + \delta^4/\lambda_n^2\right).$$

Letting $\tilde{\gamma}_l = \gamma_l/2 - 64\tau_l|S_{\alpha\eta}|$, we have

$$\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 \leq \frac{1}{\tilde{\gamma}_l}\left(\delta^2 + \frac{16\tau_l\delta^4}{\lambda_n^2}\right) + \frac{256\tau_l}{\tilde{\gamma}_l}(|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + \|\boldsymbol{\beta}^*_{\alpha,S^c_{\alpha\eta}}\|_1^2). \tag{21}$$

We now bound the second term in (21). By (11) and (12), we have

$$\begin{aligned}
|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + \|\boldsymbol{\beta}^*_{S^c_{\alpha\eta}}\|_1^2 &\leq R_q\eta^{-q}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + R_q^2\eta^{2-2q} \\
&\leq R_q\kappa_\lambda^{-q}\left(\frac{\log p}{n}\right)^{-q/2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + \kappa_\lambda^{-q}R_q^2\left(\frac{\log p}{n}\right)^{1-q} \\
&\leq \kappa_\lambda^{-q}R_q\left(\frac{\log p}{n}\right)^{-q/2}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}\right].
\end{aligned} \tag{22}$$

Meanwhile, from (a) we have

$$\begin{aligned}
\delta^2 &= \frac{\varepsilon^2}{1-\kappa} = \frac{8\xi\chi}{1-\kappa}\left(6\sqrt{|S_{\alpha\eta}|}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2 + 8\|\boldsymbol{\beta}^*_{\alpha,S^c_{\alpha\eta}}\|_1\right)^2 \\
&\leq \frac{8\xi\chi}{1-\kappa}(72|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + 128\|\boldsymbol{\beta}^*_{\alpha,S^c_{\alpha\eta}}\|_1^2) \\
&\leq \frac{1024\xi\chi}{1-\kappa}(|S_{\alpha\eta}|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + \|\boldsymbol{\beta}^*_{\alpha,S^c_{\alpha\eta}}\|_1^2).
\end{aligned} \tag{23}$$

Since $\bar{\gamma}_l \asymp 1$, $\kappa \asymp 1$, $\xi \asymp 1$, $\chi \asymp \frac{\log p}{n}$, and $\tau_l \asymp \frac{\log p}{n}$, it follows from (21), (22) and (23) that

$$\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}\|_2^2 \leq d_3R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2^2 + R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}\right],$$

where d_3 is a generic positive constant depending on M_k, κ_l, κ_u, κ_0 and κ_λ. □

A.6. Proof of Theorem 6.

First, we prove that the approximation error has $\|\boldsymbol{\beta}_\alpha^{c*} - \boldsymbol{\beta}^*\|_2 \leq d_4\alpha^{k-1}$, where $\boldsymbol{\beta}_\alpha^{c*} = \mathrm{argmin}_{\boldsymbol{\beta}}\,\mathrm{E}\,\ell_\alpha^c(y - \mathbf{x}^T\boldsymbol{\beta}^*)$ is the population minimizer under the Catoni loss. Let $g_\alpha(x) = \ell(x) - \ell_\alpha^c(x) = \int_0^x[2t - \frac{2}{\alpha}\psi_c(\alpha t)]\mathrm{d}t$. It follows from (A.2) in the Appendix of the main paper that

$$\mathrm{E}[\ell(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*) - \ell(y - \mathbf{x}^T\boldsymbol{\beta}^*)] \leq \mathrm{E}[|g'_\alpha(y - \mathbf{x}^T\tilde{\boldsymbol{\beta}})\mathbf{x}^T(\boldsymbol{\beta}_\alpha^{c*} - \boldsymbol{\beta}^*)|],$$

where $\tilde{\boldsymbol\beta}$ is a vector lying between $\boldsymbol\beta^*$ and $\boldsymbol\beta_\alpha^{c*}$. Since $|(\psi_c)'''| \le 3$, by the second-order Taylor expansion with an integral remainder,

$$|g_\alpha'(x)| = |2x - \frac{2}{\alpha}\psi_c(\alpha x)| = \left|\frac{\alpha^2}{3}\int_0^x (\psi_c)'''(\alpha s)(x-s)^2 ds\right| \le \alpha^2|x|^3. \tag{24}$$

Hence, we have

$$\begin{aligned}
\mathrm{E}\{\ell(y-\mathbf{x}^T\boldsymbol\beta_\alpha^{c*}) - \ell(y-\mathbf{x}^T\boldsymbol\beta^*)\} &\le \alpha^2\,\mathrm{E}\{|y-\mathbf{x}^T\tilde{\boldsymbol\beta}|^3|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\}\\
&\le 4\alpha^2\,\mathrm{E}\{(|\epsilon|^3 + |\mathbf{x}^T(\tilde{\boldsymbol\beta}-\boldsymbol\beta^*)|^3)|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\}\\
&\le 4\alpha^2\left[\mathrm{E}\{|\epsilon|^3|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\} + \mathrm{E}\{|\mathbf{x}^T(\tilde{\boldsymbol\beta}-\boldsymbol\beta^*)|^3|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\}\right].
\end{aligned}$$

Follow a similar proof as in Theorem 1, we have

$$\mathrm{E}\{|\epsilon|^3|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\} \lesssim \|\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*\|_2 \quad \text{and} \quad \mathrm{E}\{|\mathbf{x}^T(\tilde{\boldsymbol\beta}-\boldsymbol\beta^*)|^3|\mathbf{x}^T(\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*)|\} \lesssim \|\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*\|_2.$$

Therefore, $\|\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*\|_2 \le d_4\alpha^2$, for some generic positive constant d_4. If condition (C1) holds for $k=2$, using a first-order Taylor expansion of $g_\alpha'(x)$ and similar argument as in the above, we have $\|\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*\|_2 \le d_4\alpha$. Next, since $(\psi_c)'(0)=1$, by the same argument as in the proof of Lemma 2 and 3, RSC holds for Catoni's loss with probability no less than $1-c_1\exp(-c_2 n)$, given that $\lambda_n = \kappa_\lambda\sqrt{(\log p)/n}$ for sufficiently large κ_λ and $\lambda_n \lesssim \alpha \lesssim \rho_2^{-1}$. Hence, similarly as in Theorem 2, with high probability, $\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_\alpha^{c*}\|_2 \le d_5\sqrt{R_q}[(\log p)/n]^{1/2-q/4}$, for some generic positive constant d_5. This together with $\|\boldsymbol\beta_\alpha^{c*}-\boldsymbol\beta^*\|_2 \le d_4\alpha^{k-1}$ completes the proof. □

A.7. Proof of Theorem 7.

First of all, observe that

$$\begin{aligned}
\widehat\sigma^2 - \sigma^2 &= \frac{1}{J}\sum_{j=1}^J \frac{1}{m}\sum_{i\in\text{fold } j}\left(\epsilon_i - (\mathbf{x}_i^T\widehat{\boldsymbol\beta}^{(-j)} - \mathbf{x}_i^T\boldsymbol\beta^*)\right)^2 - \sigma^2\\
&= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2 - \frac{1}{J}\sum_{j=1}^J\frac{2}{m}\sum_{i\in\text{fold } j}\epsilon_i\mathbf{x}_i^T(\widehat{\boldsymbol\beta}^{(-j)}-\boldsymbol\beta^*) + \frac{1}{J}\sum_{j=1}^J\frac{1}{m}\sum_{i\in\text{fold } j}\{\mathbf{x}_i^T(\widehat{\boldsymbol\beta}^{(-j)}-\boldsymbol\beta^*)\}^2.
\end{aligned}$$

Given that $\mathrm{E}\,\epsilon^4$ exists, by Central Limit Theorem, $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{E}\,\epsilon^4 - \sigma^4)$. Let $z_i = \mathbf{x}_i^T(\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*)$. We now need to prove that the last two terms are negligible. Conditioning on data outside the jth fold,

$$\mathrm{E}\left\{\frac{1}{m}\left(\sum_{i\in\text{fold } k}\epsilon_i z_i\right)^2\right\} = \mathrm{E}\{\mathrm{E}(\epsilon_i^2|\mathbf{x}_i)z_i^2\} \le \left[\mathrm{E}\{\mathrm{E}(\epsilon_i^2|\mathbf{x}_i)\}^2\right]^{1/2}(\mathrm{E}\,z_i^4)^{1/2} \le \sqrt{6M_2}\kappa_0^2\|\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*\|_2^2.$$

Hence, $m^{-1/2}\sum_{i\in\text{fold } k}\epsilon_i\mathbf{x}_i^T(\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*) = O_P\left(\|\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*\|_2\right) = o_P(1)$, where the last equality follows from Theorem 3. By an analogous argument, we have

$$\begin{aligned}\frac{1}{m}\sum_{i\in\text{fold } k}\left(\mathbf{x}_i^T(\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*)\right)^2 &= O_p\left(\|\widehat{\boldsymbol{\beta}}^{(-j)} - \boldsymbol{\beta}^*\|_2^2\right) = O_p(\max\{\alpha^{2(k-1)}, R_q[(\log p)/n]^{1-q/2}\}) \\ &= o(1/\sqrt{n}).\end{aligned}$$

This completes the proof. □

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, **40**, 2452–2482.

Alexander, K. S. (1987). Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, **75**, 379–423.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48**, 1148–1185.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, **105**, 1042-1055.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes.* Springer.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*, 476–484.

Massart, P. and Picard, J. (2007). *Concentration inequalities and model selection.* Springer.

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

Rivasplata, O. (2012). Subgaussian random variables: an expository note.

Van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge university press Cambridge.