Submitted to the Statistical Science

Supplementary material for "A General Theory of Concave Regularization for High Dimensional Sparse Estimation Problems"

Cun-Hui Zhang* **Tong Zhang**†

Department of Statistics and Biostatistics, Rutgers University

APPENDIX A: TECHNICAL PROOFS

This supplementary material contains the technical proofs for results in the main paper. The notations are consistent with the main paper.

We first prove the following two lemmas, which will be useful in the analysis.

LEMMA 1. *If $\widehat{\boldsymbol{\beta}}$ is the global solution of (2) of [4], then $\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty \le \lambda^*$. In particular, $\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty \le \eta\lambda^*$ under the η-NC condition (18) of [4].*

PROOF. The optimality of $\widehat{\boldsymbol{\beta}}$ implies

$$\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\rho(\widehat{\beta}_j;\lambda) \le \|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{x}_j t\|_2^2/(2n)+\rho(\widehat{\beta}_j+t;\lambda)$$

for all real t. Since $\rho(t;\lambda)$ is subadditive in t,

$$t\boldsymbol{x}_j^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n \le t^2\|\boldsymbol{x}_j\|_2^2/(2n)+\rho(\widehat{\beta}_j+t;\lambda)-\rho(\widehat{\beta}_j;\lambda) \le t^2/2+\rho(t;\lambda).$$

Since t is arbitrary, we obtain the desired bound via the definition of λ^* in (13) of [4]. □

LEMMA 2. *Assume the η-NC condition (18) of [4] with $\eta\in(0,1)$. Suppose $\widehat{\boldsymbol{\beta}}\in\mathbb{R}^p$ satisfy*

$$\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\|\rho(\widehat{\boldsymbol{\beta}};\lambda)\|_1 \le \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n)+\|\rho(\boldsymbol{\beta};\lambda)\|_1+\nu$$

with a certain $\nu>0$. Let $\boldsymbol{\Delta}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$, $\xi=(1+\eta)/(1-\eta)$, and $S=supp(\boldsymbol{\beta})$. Then,

$$\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1 \le \xi\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1+\nu/(1-\eta).$$

Department of Statistics and Biostatistics, Rutgers University, Hill Center, Busch Campus, Piscataway, NJ 08854 (e-mail: czhang@stat.rutgers.edu; tzhang@stat.rutgers.edu).

*Research partially supported by the NSF Grants DMS 0906420, DMS-11-06753 and NSA Grant H98230-11-1-0205

†Research partially supported by the following grants: AFOSR-10097389, NSA -AMS 081024, NSF DMS-1007527, and NSF IIS-1016061

PROOF. From the condition of the lemma, we have

$$\begin{aligned} 0 &\le \nu + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \|\rho(\boldsymbol{\beta};\lambda)\|_1 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n) - \|\rho(\widehat{\boldsymbol{\beta}};\lambda)\|_1 \\ &= \nu - \|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) + \boldsymbol{\varepsilon}^\top \boldsymbol{X}\boldsymbol{\Delta}/n + \|\rho(\boldsymbol{\beta};\lambda)\|_1 - \|\rho(\boldsymbol{\beta}+\boldsymbol{\Delta};\lambda)\|_1. \end{aligned}$$

By (18) of [4], $\|\boldsymbol{\varepsilon}/\eta\|_2^2/(2n) \le \|\boldsymbol{\varepsilon}/\eta - t\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) + \|\rho(t\boldsymbol{\Delta};\lambda)\|_1$ for all $t > 0$, which can be written as

$$\boldsymbol{\varepsilon}^\top \boldsymbol{X}\boldsymbol{\Delta}/n \le \eta t \|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) + (\eta/t)\|\rho(t\boldsymbol{\Delta};\lambda)\|_1.$$

The above two displayed inequalities yield

$$(1-\eta t)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) - \nu \le (\eta/t)\|\rho(t\boldsymbol{\Delta};\lambda)\|_1 + \|\rho(\boldsymbol{\beta};\lambda)\|_1 - \|\rho(\boldsymbol{\beta}+\boldsymbol{\Delta};\lambda)\|_1. \tag{A.1}$$

Now let $t = 1$. It follows from (A.1), $\boldsymbol{\beta}_{S^c} = 0$, and then the sub-additivity of $\rho(t;\lambda)$ that

$$\begin{aligned} & (1-\eta)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) - \nu \\ \le\ & \eta\|\rho(\boldsymbol{\Delta};\lambda)\|_1 + \|\rho(\boldsymbol{\beta}_S;\lambda)\|_1 - \|\rho(\boldsymbol{\beta}_S + \boldsymbol{\Delta}_S;\lambda)\|_1 - \|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1 \\ \le\ & (\eta+1)\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1 + (\eta-1)\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1. \qquad \square \end{aligned}$$

A.1 Proof of Proposition 1

Let $t > 0$. By (13) of [4], $\rho(t;\lambda) \ge t(\lambda^* - t/2) \ge t\lambda^*/2$ for $t \le \lambda^*$. For $t > \lambda^*$, $\rho(t;\lambda) \ge \rho(\lambda^*;\lambda) \ge (\lambda^*)^2/2$. This gives the lower bound of $\rho(t;\lambda)$. Let t_0 be the minimizer in (13) of [4] in the sense of $x/2 + \rho(x;\lambda)/x \to \lambda^*$ as $x \to t_0$ (when t_0 is a discontinuity of $\rho(\cdot;\lambda)$) or $x = t_0$. Let $x > 0$ and $q = \lfloor t/x \rfloor$. Since $\rho(t;\lambda)$ is nondecreasing and subadditive in $t > 0$, we have

$$\rho(t;\lambda) \le \rho(qx;\lambda) + \rho(t-qx;\lambda) \le (q+1)\rho(x;\lambda) \le (t+x)\rho(x;\lambda)/x.$$

It follows that (let $x \to t_0$) $\rho(t;\lambda) \le (t+t_0)(\lambda^*-t_0/2) \le \max_{t'\ge 0}(t+t')(\lambda^*-t'/2) = \rho^*(t;\lambda)$. The bound for $\Delta(a,k;\lambda)$ follows similarly from (let $x \to t_0$)

$$\|\rho(\boldsymbol{b};\lambda)\|_1 \le \sum_{j:b_j\ne 0}(|b_j|+x)\rho(x;\lambda)/x \le k(a+x)\rho(x;\lambda)/x \le k(a+t_0)(\lambda^*-t_0/2) \le k\rho^*(a;\lambda).$$

The fact that $\rho^*(a;\lambda) \le \max(a;2\lambda^*)\lambda^*$ can be verified by simple algebra. $\square$

A.2 Proof of Proposition 2

Let $f(t) = t/\rho(t;\lambda)$ and A be the index set of the $|S|$ largest $|u_j|$. Since $\rho(t;\lambda)$ is nondecreasing in $|t|$, $\|\rho(\boldsymbol{u}_{S^c};\lambda)\|_1 < \xi\|\rho(\boldsymbol{u}_S;\lambda)\|_1$ implies $\|\rho(\boldsymbol{u}_{A^c};\lambda)\|_1 < \xi\|\rho(\boldsymbol{u}_A;\lambda)\|_1$. Since $f(t)$ is nondecreasing in t,

$$\|\boldsymbol{u}_{A^c}\|_1 \le \|\rho(\boldsymbol{u}_{A^c};\lambda)\|_1 f(\|\boldsymbol{u}_{A^c}\|_\infty) \le \xi\|\rho(\boldsymbol{u}_A;\lambda)\|_1 f(\|\boldsymbol{u}_{A^c}\|_\infty) \le \xi\|\boldsymbol{u}_A\|_1.$$

This implies (16) of [4]. In the above derivation, the first inequality follows from the definition of $f(t)$ and $\|\boldsymbol{u}_{A^c}\|_\infty \ge |\boldsymbol{u}_j|$ for all $j \in A^c$; the second inequality is due to the condition $\|\rho(\boldsymbol{u}_{A^c};\lambda)\|_1 < \xi\|\rho(\boldsymbol{u}_A;\lambda)\|_1$; the third inequality follows from the definition of $f(t)$ and the condition $\|\boldsymbol{u}_{A^c}\|_\infty \le |\boldsymbol{u}_j|$ for all $j \in A$. $\square$

A.3 Proof of Proposition 3

Since the left-hand side of (18) of [4] is increasing in $\rho(t;\lambda)$, we assume without loss of generality that

$$\rho(t;\lambda) = \min\left(\lambda^2/2, \lambda|t|\right),\ \lambda = (1+\zeta_0)(\sigma/\eta)\lambda_0,\ \lambda_0 = \left(1+\sqrt{2\ln(2p/\delta)}\right)/\sqrt{n}.$$

Since $\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty \le \max_{|A|=1}\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2/\sqrt{n}$ and $\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2 \le \|\boldsymbol{\varepsilon}\|_2$, Assumption 1 implies that

$$\text{(A.2)} \qquad \|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty \le \sigma\lambda_0,\ \|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2 \le \sigma\min\left[\sqrt{|A|n}\lambda_0, \sqrt{2n}\right],\ \forall\ A\subseteq\{1,\ldots,p\},$$

with at least probability $1-\exp(-n(1-1/\sqrt{2})^2) - \sum_{k=1}^n \binom{p}{k}\left(\delta/(2p)\right)^k \ge 2-\delta_n - e^{\delta/2}$.

Let $A = \{j : |b_j| > \lambda/2\}$ and $k = |A|$. It suffices to consider the case where A and $\boldsymbol{b}$ satisfy

$$\boldsymbol{X}_A\boldsymbol{b}_A = \boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta - \boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}),\ \mathrm{rank}(\boldsymbol{P}_A) = |A| = k \le \frac{\|\boldsymbol{\varepsilon}/\eta\|_2^2/(2n)}{\lambda^2/2} \le \frac{2}{(1+\zeta_0)^2\lambda_0^2},$$

since these conditions hold for the global minimum for (2) of [4] with $\boldsymbol{y} = \boldsymbol{\varepsilon}/\eta$ and the capped-ℓ_1 penalty. Under these conditions, we have $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{P}_A\boldsymbol{\varepsilon}/\eta + \boldsymbol{P}_A^\perp\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}$ and

$$\begin{aligned}
\text{(A.3)} \quad & \|\boldsymbol{\varepsilon}/\eta\|_2^2/(2n) - \|\boldsymbol{\varepsilon}/\eta - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2n) - \|\rho(\boldsymbol{b};\lambda)\|_1 \\
= \quad & (\boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{\varepsilon}/\eta)/n - \|\boldsymbol{X}\boldsymbol{b}\|_2^2/(2n) - \|\rho(\boldsymbol{b};\lambda)\|_1 \\
= \quad & \|\boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta)\|_2^2/(2n) + (\boldsymbol{P}_A^\perp\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c})^\top(\boldsymbol{\varepsilon}/\eta)/n - \|\boldsymbol{P}_A^\perp\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}\|_2^2/(2n) - \|\rho(\boldsymbol{b};\lambda)\|_1 \\
\le \quad & \|\boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta)\|_2^2/(2n) + (\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c})^\top(\boldsymbol{\varepsilon}/\eta)/n - (\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c})^\top\boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta)/n - \|\rho(\boldsymbol{b};\lambda)\|_1 \\
\le \quad & \lambda_0^2 k(\sigma/\eta)^2/2 + \lambda_0(\sigma/\eta)\|\boldsymbol{b}_{A^c}\|_1 - (\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c})^\top\boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta)/n - \|\rho(\boldsymbol{b};\lambda)\|_1 \\
< \quad & \Big(\frac{1}{1+\zeta_0} - 1\Big)\|\rho(\boldsymbol{b};\lambda)\|_1 - (\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c})^\top\boldsymbol{P}_A(\boldsymbol{\varepsilon}/\eta)/n.
\end{aligned}$$

In the above derivation, the second inequality uses (A.2) and the third uses the fact that $\|\rho(\boldsymbol{b};\lambda)\|_1 = \lambda^2k/2 + \lambda\|\boldsymbol{b}_{A^c}\|_1$ by the definition of A and $\lambda = (1+\zeta_0)(\sigma/\eta)\lambda_0$.

It follows from the shifting inequality in [1, 3] that

$$\begin{aligned}
\Big|\boldsymbol{\varepsilon}^\top\boldsymbol{P}_A\boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}\Big| &\le \max_{B\cap A=\emptyset,|B|=k}\|\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2\Big(\|\boldsymbol{b}_{A^c}\|_\infty k^{1/2} + \|\boldsymbol{b}_{A^c}\|_1/k^{1/2}\Big) \\
&\le \max_{B\cap A=\emptyset,|B|=k}\|\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2(\lambda k/2 + \|\boldsymbol{b}_{A^c}\|_1)/\sqrt{k}.
\end{aligned}$$

In the above derivation, the first inequality uses the shifting inequality and the second uses the fact that $\|\boldsymbol{b}_{A^c}\|_\infty \le \lambda/2$ due to the definition of A. It follows from (19) of [4] and $\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2 \le \sigma\lambda_0\sqrt{nk}$ of (A.2) that for all $|A| = |B| = k \le 2/\{(1+\zeta_0)^2\lambda_0^2\}$ with $B\cap A = \emptyset$,

$$\|\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2 \le \lambda_{\max}^{1/2}(\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{X}_B)\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|_2 \le (\sigma\lambda_0\sqrt{nk})(\zeta_0\sqrt{n}) = \sigma\zeta_0\lambda_0 n\sqrt{k}.$$

Thus, by combining the above two displayed inequalities, we find

$$\Big|\boldsymbol{\varepsilon}^\top \boldsymbol{P}_A \boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}/(\eta n)\Big| \le \frac{\sigma\zeta_0\lambda_0 n\sqrt{k}}{\eta n}\Big(\frac{\lambda k/2+\|\boldsymbol{b}_{A^c}\|_1}{\sqrt{k}}\Big) = (1+\zeta_0)^{-1}\zeta_0\|\rho(\boldsymbol{b};\lambda)\|_1.$$

due to $\lambda = (1+\zeta_0)(\sigma/\eta)\lambda_0$ and $\|\rho(\boldsymbol{b};\lambda)\|_1 = \lambda^2 k/2 + \lambda\|\boldsymbol{b}_{A^c}\|_1$. This and (A.3) yield the η-NC condition (18) of [4].

It remains to prove that (19) of [4] is an ℓ_2-regularity condition on $\boldsymbol{X}$. Suppose that the rows of $\boldsymbol{X}$ are iid from $N(0,\boldsymbol{\Sigma})$. Let $\boldsymbol{N}_{k,m}$ denote a $k\times m$ matrix with iid $N(0,1)$ entries. We may write $\boldsymbol{X}_B = \boldsymbol{N}_{n,p}(\boldsymbol{\Sigma}^{1/2})_{p\times B}$. Let $\boldsymbol{U}\boldsymbol{U}^T$ and $\boldsymbol{V}\boldsymbol{D}\boldsymbol{W}^T$ be the SVDs of $\boldsymbol{P}_A$ and $(\boldsymbol{\Sigma}^{1/2})_{p\times B}$ respectively. For fixed $\{A,B\}$, the entries of the $k\times k$ matrix $\boldsymbol{U}^T\boldsymbol{N}_{n,p}\boldsymbol{V}$ are uncorrelated $N(0,1)$ variables, so that we can write $\boldsymbol{P}_A\boldsymbol{X}_B = \boldsymbol{U}\boldsymbol{N}_{k,k}\boldsymbol{V}^T(\boldsymbol{\Sigma}^{1/2})_{p\times B}$. Thus, by Theorem II.13 of [2],

$$\begin{aligned} P\{\lambda^{1/2}_{\max}(\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{X}_B) > (2k^{1/2}+t)\lambda^{1/2}_{\max}(\boldsymbol{\Sigma})\} &\le P\{\lambda^{1/2}_{\max}(\boldsymbol{N}_{k,k}^\top\boldsymbol{N}_{k,k}) > 2k^{1/2}+t\} \\ &\le \Phi(-t) \le e^{-t^2/2}/2,\ t>0, \end{aligned}$$

where $\Phi(t)$ is the $N(0,1)$ distribution function. Since there are no more than $\binom{p}{k,k,p-2k}$ choices of $\boldsymbol{P}_A$ with rank k and $|B|=k$ with $A\cap B=\emptyset$,

$$\max_{B\cap A=\emptyset,|B|=|A|=k}\lambda^{1/2}_{\max}(\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{X}_B) \le \lambda^{1/2}_{\max}(\boldsymbol{\Sigma})\Big(2k^{1/2}+\sqrt{8k\ln(2p/\delta)}\Big),\ \forall\ 1\le k\le n, \tag{A.4}$$

with probability no smaller than

$$1-\frac{1}{2}\sum_{k=1}^n\Big(\frac{\delta}{2p}\Big)^{4k}\binom{p}{k,k,p-2k} \ge 1-\frac{1}{2}\sum_{k=1}^n\frac{(\delta^2/(4p))^{2k}}{(k!)^2} \ge 1-\delta^4/(16p^2).$$

In the event (A.4), we have that for all $|A|=|B|=k$ and $k(1+\zeta_0)^2(1+\sqrt{2\ln(2p/\delta)})^2\le 2n$,

$$\lambda^{1/2}_{\max}(\boldsymbol{X}_B^\top\boldsymbol{P}_A\boldsymbol{X}_B/n) \le \frac{\lambda^{1/2}_{\max}(\boldsymbol{\Sigma})\big(2k^{1/2}+\sqrt{8k\ln(2p/\delta)}\big)}{\{k^{1/2}(1+\zeta_0)(1+\sqrt{2\ln(2p/\delta)})\}/\sqrt{2}} = \frac{\sqrt{8}\lambda^{1/2}_{\max}(\boldsymbol{\Sigma})}{1+\zeta_0}.$$

This proves the desired result. □

A.4 Proof of Theorem 1

Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$. Lemma 2 (with $\nu=0$) implies that

$$\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n) + \|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1 \le \xi\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1. \tag{A.5}$$

Thus, (15) of [4] gives

$$\|\boldsymbol{\Delta}\|_q \le \|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}\|_\infty|S|^{1/q}/\{n\mathrm{RIF}_q(\xi,S)\}. \tag{A.6}$$

It follows from Lemma 1 that $\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty \le \lambda^*$ and $\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty \le \eta\lambda^*$. Thus, we have $\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty = \|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{\varepsilon})/n\|_\infty \le (1+\eta)\lambda^*$. This and (A.6) yield (20) of [4].

Now by combining the definition of $\Delta(a,|S|;\lambda)$ and $\|\boldsymbol{\Delta}\|_1 \le (1+\eta)\lambda^*|S|/\mathrm{RIF}_1(\xi,S)$, which follows from (20) of [4], we obtain an estimate of $\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1$ in (A.5), which leads to the first inequality in (21) of [4]. The second inequality in (21) of [4] then follows from Proposition 1. $\square$

A.5 Proof of Theorem 2

Let $\widehat{S}_1 = \{j \in \widehat{S}\setminus S : |\widehat{\beta}_j| \ge t_0\}$ and $\widehat{S}_2 = \{j \in \widehat{S}\setminus S : |\widehat{\beta}_j| < t_0\}$. As in the proof of (21) of [4], it follows from the ℓ_1 error bound (20) of [4] and the definition of $\Delta(a_1,|S|;\lambda)$ in (14) of [4] that $\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1 \le \Delta(a_1,|S|;\lambda)$ with the given a_1. Thus,

$$|\widehat{S}_1| \le \|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1/\rho(t_0;\lambda) \le \xi\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1/\rho(t_0;\lambda) \le \xi\Delta(a_1,|S|;\lambda)/\rho(t_0;\lambda). \tag{A.7}$$

Let $\lambda_2 > \sqrt{2\xi\kappa_+(m_0)\Delta(a_1\lambda^*,|S|;\lambda)/m_0}$ satisfying $\lambda_2+\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty \le \inf_{0<s<t_0}\dot{\rho}(s;\lambda)$. The first order optimality condition implies that for all $j\in\widehat{S}$, $\boldsymbol{x}_j^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = \dot{\rho}(t;\lambda)\big|_{t=\widehat{\beta}_j}$. For $j\in\widehat{S}_2$, $|\widehat{\beta}_j|\in(0,t_0)$, so that $|\boldsymbol{x}_j^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n| \ge (\lambda_2+\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty)$ by (22) of [4]. Thus, for any set $A\subset\widehat{S}_2$ with $|A|\le m_0$,

$$(\lambda_2+\|\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n\|_\infty)|A| \le \|\boldsymbol{X}_A^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_1 \le \|\boldsymbol{X}_A^\top\boldsymbol{\varepsilon}/n\|_\infty|A|+|A|^{1/2}\|\boldsymbol{X}_A/\sqrt{n}\|_2\|\boldsymbol{X}\boldsymbol{\Delta}\|_2/\sqrt{n}.$$

Since $\|\boldsymbol{X}_A/\sqrt{n}\|_2^2 \le \kappa_+(m_0)$, $\lambda_2|A| \le |A|^{1/2}\sqrt{\kappa_+(m_0)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/n}$. It follows from Theorem 1 that $|A| \le \kappa_+(m_0)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(n\lambda_2^2) \le 2\xi\kappa_+(m_0)\Delta(a_1,|S|;\lambda)/\lambda_2^2 < m_0$. Thus,

$$\max_{A\subset\widehat{S}_2,|A|\le m_0}|A| < m_0,$$

which implies that $|\widehat{S}_2| < m_0$. Combine this estimate with (A.7), we obtain the desired bound. $\square$

A.6 Proof of Theorem 3

It follows from the assumption of the theorem that for all $\boldsymbol{b}\in\mathbb{R}^p$,

$$\|\boldsymbol{X}\boldsymbol{b}-\boldsymbol{\varepsilon}/\eta\|_2^2+\lambda^2 n\|\boldsymbol{b}\|_0-\|\boldsymbol{\varepsilon}/\eta\|_2^2 = \|\boldsymbol{X}\boldsymbol{b}\|_2^2+(2/\eta)\boldsymbol{\varepsilon}^\top X\boldsymbol{b}+\lambda^2 n\|\boldsymbol{b}\|_0$$

is bounded from below by $\|\boldsymbol{X}\boldsymbol{b}\|_2^2-2\lambda\sqrt{n\|\boldsymbol{b}\|_0}\|\boldsymbol{X}\boldsymbol{b}\|_2+\lambda^2 n\|\boldsymbol{b}\|_0 = (\|\boldsymbol{X}\boldsymbol{b}\|_2^2-\lambda\sqrt{n\|\boldsymbol{b}\|_0})^2 \ge 0$. This implies the null-consistency condition. Moreover, (A.1) with $t=1/\eta$ and $\nu=0$ implies that

$$\|\widehat{\boldsymbol{\beta}}^{(\ell_0)}\|_0-\|\boldsymbol{\beta}\|_0 \le \eta^2\|\widehat{\boldsymbol{\beta}}^{(\ell_0)}-\boldsymbol{\beta}\|_0 \le \eta^2\|\widehat{\boldsymbol{\beta}}^{(\ell_0)}\|_0+\eta^2\|\boldsymbol{\beta}\|_0,$$

which leads to the first bound of the theorem. The second bound is a direct consequence of Theorem 1, since $\Delta(\xi,|S|;\lambda)=\lambda^2|S|/2$ by (14) of [4]. $\square$

A.7 Proof of Theorem 4

For simplicity, let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(\ell_0)}$, $\widehat{S} = \mathrm{supp}(\widehat{\boldsymbol{\beta}})$, and $S = \mathrm{supp}(\boldsymbol{\beta})$. We know that $\|\widehat{\boldsymbol{\beta}}\|_0 \le (1+\eta^2)/(1-\eta^2)\|\boldsymbol{\beta}\|_0$ and thus $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\|_0 \le s$. Similar to the proof of Theorem 3, we have

$$
\begin{aligned}
0 \ge & \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 + 2(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o - \boldsymbol{y})^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o) + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & \kappa_-(s) n \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\|_2^2 - \sqrt{2\kappa_-(s)}\lambda n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_1 + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & \kappa_-(s) n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_S\|_2^2 + \kappa_-(s) n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_2^2 \\
& \quad - 2\sqrt{0.5\lambda^2 n|\widehat{S} - S|}\sqrt{\kappa_-(s)n}\|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_2 + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & \kappa_-(s) n \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2^2 - 0.5\lambda^2 n|\widehat{S} - S| + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & 2\lambda^2 n(|S - \widehat{S}| - \delta^o) - 0.5\lambda^2 n|\widehat{S} - S| + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & \lambda^2 n(|S - \widehat{S}| + 0.5|\widehat{S} - S| - 2\delta^o).
\end{aligned}
$$

The first inequality uses the same derivation of a similar result in the proof of Theorem 3. The second inequality uses the assumption of the theorem, $(\boldsymbol{P}_S\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon})^\top \boldsymbol{X} = (\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o - \boldsymbol{y})^\top \boldsymbol{X}$, and the fact that $(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o - \boldsymbol{y})^\top \boldsymbol{X}_S = 0$. The forth inequality uses $b^2 - 2ab \ge -a^2$ and $\|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_S\|_2 \ge \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2$. The fifth inequality uses

$$
\begin{aligned}
\kappa_-(s) n \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2^2 \ge & \kappa_-(s) n \sum_{j \in S - \widehat{S}; |\widehat{\beta}_j|^2 \ge 2\lambda/\kappa_-(s)} (\widehat{\boldsymbol{\beta}}^o)_j^2 \\
\ge & 2\lambda^2 n \left| \{ j \in S - \widehat{S}; (\widehat{\beta}_j^o)^2 \ge 2\lambda^2/\kappa_-(s) \} \right| \ge 2\lambda^2 n(|S - \widehat{S}| - \delta^o).
\end{aligned}
$$

The last inequality uses the derivation $\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0 \ge |\widehat{S}| - |S| = |\widehat{S} - S| - |S - \widehat{S}|$ and simple algebra. This proves the first desired bound. Similarly, we have

$$
\begin{aligned}
0 \ge & \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 - \sqrt{2\kappa_-(s)}\lambda n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_1 + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & 0.5\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 + 0.5\kappa_-(s) n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_S\|_2^2 + 0.5\kappa_-(s) n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_2^2 \\
& \quad - \sqrt{2\kappa_-(s)|\widehat{S} - S|}\lambda n \|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_{\widehat{S}-S}\|_2 + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & 0.5\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 + 0.5\kappa_-(s) n \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2^2 - \lambda^2 n|\widehat{S} - S| + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & 0.5\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 + \lambda^2 n(|S - \widehat{S}| - \delta^o) - \lambda^2 n|\widehat{S} - S| + \lambda^2 n[\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0] \\
\ge & 0.5\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 - \lambda^2 n\delta^o.
\end{aligned}
$$

The second inequality uses the definition of $\kappa_-(s)$. The third inequality uses $0.5b^2 - \sqrt{2}ab \ge -a^2$ and $\|(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o)_S\|_2 \ge \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2$. The fourth inequality uses the previously derived inequality $\kappa_-(s) n \|(\widehat{\boldsymbol{\beta}}^o)_{S-\widehat{S}}\|_2^2 \ge 2\lambda^2 n(|S - \widehat{S}| - \delta^o)$. The last inequality uses the derivation $\|\widehat{\boldsymbol{\beta}}\|_0 - \|\widehat{\boldsymbol{\beta}}^o\|_0 \ge |\widehat{S}| - |S| = |\widehat{S} - S| - |S - \widehat{S}|$ and simple algebra. This leads to the second desired bound. □

A.8 Proof of Theorem 5

Since $\widetilde{\boldsymbol{\beta}}^{(j)}$ are approximate local solutions with excess $\nu^{(j)}$, (26) of [4] gives

$$\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}/n+\dot\rho(\widetilde{\boldsymbol{\beta}}^{(1)};\lambda)-\dot\rho(\widetilde{\boldsymbol{\beta}}^{(2)};\lambda)\|_2\le(\nu^{(1)})^{1/2}+(\nu^{(2)})^{1/2}\le\sqrt{\nu}.$$

Let $E=\widetilde{S}^{(1)}\cup\widetilde{S}^{(2)}$. Since $|E|\le m+k$ and $\kappa<\kappa_-(m+k)$, it follows that

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/n\le&-\boldsymbol{\Delta}^\top(\dot\rho(\widetilde{\boldsymbol{\beta}}^{(1)};\lambda)-\dot\rho(\widetilde{\boldsymbol{\beta}}^{(2)};\lambda))+\sqrt{\nu}\|\boldsymbol{\Delta}\|_2\\ \le&\kappa\|\boldsymbol{\Delta}\|_2^2+\left|(\boldsymbol{\Delta}^\top\theta(|\widetilde{\boldsymbol{\beta}}^{(1)}|,\kappa)\right|+\sqrt{\nu}\|\boldsymbol{\Delta}\|_2\\ \le&\kappa\|\boldsymbol{\Delta}\|_2^2+\left(\|\theta(|\widetilde{\boldsymbol{\beta}}^{(1)}_E|,\kappa)\|_2+\sqrt{\nu}\right)\|\boldsymbol{\Delta}\|_2\end{aligned}$$

Since $\|\boldsymbol{\Delta}\|_2^2\le\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/\{n\kappa_-(m+k)\}$, (27) of [4] follows.

Let $E_1:=\{j:|\widetilde{\beta}_j^{(1)}-\widetilde{\beta}_j^{(2)}|\ge\lambda_0/\sqrt{\kappa_-(m+k)}\}$. We have $\lambda_0^2|E_1|\le\kappa_-(m+k)\|\boldsymbol{\Delta}\|_2^2\le\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/n$. Since $j\in S\setminus\widetilde{S}^{(2)}$ implies $\widetilde{\beta}_j^{(1)}-\widetilde{\beta}_j^{(2)}=\widetilde{\beta}_j^{(1)}$, (28) of [4] follows.

Let $E_2:=\widetilde{S}^{(2)}\setminus S$ and $\lambda_0'=\dot\rho(0+;\lambda)-\|\boldsymbol{X}_{S^c}^\top(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{(1)}-\boldsymbol{y})/n\|_\infty$. For $j\in E_2$,

$$\begin{aligned}\lambda_0'&\le\dot\rho(0+;\lambda)+\mathrm{sgn}(\widetilde{\boldsymbol{\beta}}^{(2)})\boldsymbol{x}_j^\top(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{(1)}-\boldsymbol{y})/n\\ &\le\{\dot\rho(0+;\lambda)-\mathrm{sgn}(\widetilde{\boldsymbol{\beta}}^{(2)})\dot\rho(\widetilde{\beta}_j^{(2)};\lambda)\}+|\boldsymbol{x}_j^\top(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{(2)}-\boldsymbol{y})/n+\dot\rho(\widetilde{\beta}_j^{(2)};\lambda)|+|\boldsymbol{x}_j^\top\boldsymbol{X}\boldsymbol{\Delta}/n|.\end{aligned}$$

Since $\theta(0+,\kappa)=0$ means $\dot\rho(0+;\lambda)-\mathrm{sgn}(t)\dot\rho(t;\lambda)=\dot\rho(0+;\lambda)-\dot\rho(|t|;\lambda)\le\kappa|t|$ for $t\ne0$ and $\widetilde{\boldsymbol{\beta}}^{(2)}_{E_2}=-\boldsymbol{\Delta}_{E_2}$,

$$\begin{aligned}|E_2|\lambda_0'&\le\kappa\|\boldsymbol{\Delta}_{E_2}\|_1+\|\boldsymbol{X}_{E_2}^\top(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{(2)}-\boldsymbol{y})/n+\dot\rho(\widetilde{\boldsymbol{\beta}}^{(2)}_{E_2};\lambda)\|_1+\|\boldsymbol{X}_{E_2}^\top\boldsymbol{X}\boldsymbol{\Delta}/n\|_1\\ &\le\sqrt{|E_2|}\Big\{\kappa\|\boldsymbol{\Delta}\|_2+\sqrt{\nu^{(2)}}+\|\boldsymbol{X}_{E_2}^\top\boldsymbol{X}\boldsymbol{\Delta}/n\|_2\Big\}\end{aligned}$$

Since $\|\boldsymbol{\Delta}\|_2^2\le\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/\{n\kappa_-(m+k)\}$ and $\|\boldsymbol{X}_{E_2}^\top\boldsymbol{X}\boldsymbol{\Delta}/n\|_2^2\le\kappa_+(m)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/n$, (29) of [4] follows. $\square$

A.9 Proof of Theorem 6

We note that $\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)=\boldsymbol{X}^\top(\boldsymbol{\varepsilon}-\boldsymbol{P}_S\boldsymbol{\varepsilon})=\boldsymbol{X}^\top\boldsymbol{P}_S^\perp\boldsymbol{\varepsilon}$.

(i) Since $\boldsymbol{x}_j^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o-\boldsymbol{y})/n+\dot\rho(\widehat{\beta}_j^o;\lambda)=\dot\rho(\widehat{\beta}_j^o;\lambda)$ for $j\in S$ and $\boldsymbol{x}_j^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o-\boldsymbol{y})/n+\dot\rho(\widehat{\beta}_j^o;\lambda)=0$ for $j\notin S$, $\nu=\|\dot\rho(\widehat{\boldsymbol{\beta}}_S^o;\lambda)\|^2$. Let $\widehat{\boldsymbol{\beta}}^{(2)}$ be the global solution of (2) of [4]. Under the additional conditions, $\widehat{\boldsymbol{\beta}}^{(2)}=\widehat{\boldsymbol{\beta}}^o$ by Theorems 2 and (27) of [4] with $\widehat{\boldsymbol{\beta}}^{(1)}=\widehat{\boldsymbol{\beta}}^o$.

(ii) Since $\min_{j\in S}|\widehat{\beta}_j^o|\ge\theta_1\lambda^*$, the map $\boldsymbol{b}_S\to\widehat{\boldsymbol{\beta}}_S^o-(\boldsymbol{X}_S^\top\boldsymbol{X}_S/n)^{-1}\dot\rho(\boldsymbol{b}_S;\lambda)$ is continuous and closed in the rectangle $B=\{\boldsymbol{v}:\|\boldsymbol{v}_S-\widehat{\boldsymbol{\beta}}_S^o\|_\infty\le\theta_1\lambda^*,\boldsymbol{v}_{S^c}=0\}$. Thus, the Brouwer fixed point theorem implies a $\widehat{\boldsymbol{\beta}}\in B$ satisfying $\mathrm{sgn}(\widehat{\boldsymbol{\beta}})=\mathrm{sgn}(\boldsymbol{\beta})$ and

$$\boldsymbol{X}_S^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})=(\boldsymbol{X}_S^\top\boldsymbol{X}_S/n)(\widehat{\boldsymbol{\beta}}^o-\widehat{\boldsymbol{\beta}})_S=\dot\rho(\widehat{\boldsymbol{\beta}}_S;\lambda).$$

Since $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}=\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{o}-\boldsymbol{X}_S(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{o})_S=\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{o}-\boldsymbol{X}_S(\boldsymbol{X}_S^\top\boldsymbol{X}_S/n)^{-1}\dot\rho(\widehat{\boldsymbol{\beta}}_S;\lambda)$,

$$\|\boldsymbol{X}_{S^c}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty\le\|\boldsymbol{X}_{S^c}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{o})/n\|_\infty+\theta_2\lambda^*\le\lambda^*,$$

so that $\widehat{\boldsymbol{\beta}}$ is a local solution of (2) of [4]. The proof of global optimality of $\widetilde{\boldsymbol{\beta}}^{o}$ is the same as (i). $\square$

A.10 Proof of Theorem 7

The proof is similar to that of Theorem 2. As intermediate results, we will prove lemmas that are analogous to Lemma 1 and Theorem 1. In the following, we assume that the conditions of the theorem hold. We also let $\boldsymbol{\Delta}=\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}$.

LEMMA 3. *Let* $\lambda_1^*:=\sup_{t\ge0}|\dot\rho(t;\lambda)|$. *We have* $\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}})/n\|_\infty\le\lambda_1^*$.

PROOF. A local solution satisfies $|\boldsymbol{x}_j^\top(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}-\boldsymbol{y})/n|=|\dot\rho(\widetilde\beta_j;\lambda)|\le\lambda_1^*$ for all j. $\square$

LEMMA 4. *We have* $\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1\le b$ *with* $\boldsymbol{\Delta}=\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}$.

PROOF. We consider two situations: the first is $\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1\le\nu$, and the second is $\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1>\nu$. In the first situation, we obtain directly from Lemma 2 that

$$\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1\le2\nu/(1-\eta)=\xi'\nu.$$

In the second situation, we obtain from Lemma 2 that $\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1\le\xi'\|\rho(\boldsymbol{\Delta}_S;\lambda)\|_1$. Therefore (15) of [4] gives $\|\boldsymbol{\Delta}\|_1\le\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}\|_\infty|S|/\{n\mathrm{RIF}_1(\xi',S)\}$. It follows from Lemma 3 that $\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty\le\lambda_1^*$. Similarly, $\|\boldsymbol{X}^\top(\boldsymbol{\varepsilon}/\eta)/n\|_\infty\le\lambda^*=\lambda$ due to (18) of [4]. Since $\lambda=\lambda^*\le\inf_t|\rho(t;\lambda)/t|\le\lambda_1^*$, we have $\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty=\|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{\varepsilon})/n\|_\infty\le(1+\eta)\lambda_1^*$. This implies that $\|\boldsymbol{\Delta}\|_1\le a_1'\lambda_1^*|S|$, where $a_1'=(1+\eta)/\mathrm{RIF}_1(\xi',S)$. This can be combined with Lemma 2 and the definition of $\Delta(a,|S|;\lambda)$ to obtain

$$\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1\le\xi'\Delta(a_1'\lambda_1^*,|S|;\lambda).$$

Combine the two situations, we obtain the lemma. $\square$

We are now ready to prove the theorem.

(i) Let $\boldsymbol{\Delta}^{(\ell_1)}=\widehat{\boldsymbol{\beta}}^{(\ell_1)}-\boldsymbol{\beta}$. Since $|\boldsymbol{\varepsilon}^\top\boldsymbol{X}\boldsymbol{\Delta}^{(\ell_1)}/n|\le\|\boldsymbol{X}\boldsymbol{\Delta}^{(\ell_1)}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}^{(\ell_1)};\lambda)\|_1$ by (18) of [4],

$$\begin{aligned}\nu &= \|\boldsymbol{X}\boldsymbol{\Delta}^{(\ell_1)}\|_2^2/(2n)-\boldsymbol{\varepsilon}^\top\boldsymbol{X}\boldsymbol{\Delta}^{(\ell_1)}/n+\|\rho(\widehat{\boldsymbol{\beta}}^{(\ell_1)};\lambda)\|_1-\|\rho(\boldsymbol{\beta};\lambda)\|_1\\ &\le 2\Big\{\|\boldsymbol{X}\boldsymbol{\Delta}^{(\ell_1)}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}^{(\ell_1)};\lambda)\|_1\Big\}\\ &\le 2\Big\{\xi a_1\lambda^2|S|+\Delta(a_1\lambda|S|/m,m;\lambda)\Big\}\end{aligned}$$

$$\leq \quad 2\Big\{\xi a_1 \lambda^2 |S| + \lambda^2 \max(a_1|S|, 2m)\Big\} = O(\lambda^2|S|).$$

(ii) Let $\widehat{S}_1 = \{j \in \widehat{S} \setminus S : |\widehat{\beta}_j| \geq t_0\}$, $\widehat{S}_2 = \{j \in \widehat{S} \setminus S : |\widehat{\beta}_j| < t_0\}$, and $\lambda_2 > \sqrt{2\kappa_+(m_0)b/m_0}$ satisfying $\lambda_2 + \|\boldsymbol{X}^\top \boldsymbol{\varepsilon}/n\|_\infty < \inf_{0<s<t_0} \dot{\rho}(s;\lambda)$. Just as in the proof of Theorem 2, we have $|\widehat{S}_1| \leq \|\rho(\boldsymbol{\Delta}_{S^c};\lambda)\|_1/\rho(t_0;\lambda)$, and for any $A \subset \widehat{S}_2$ with $|A| \leq m_0$, $|A| \leq \kappa_+(m_0)\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2/(n\lambda_2^2)$. We apply Lemma 4 to obtain $|\widehat{S}_1| \leq b/\rho(t_0;\lambda)$ and $|A| \leq 2\kappa_+(m_0)b/\lambda_2^2 < m_0$. Thus, $\max_{A\subset \widehat{S}_2, |A|\leq m_0} |A| < m_0$, which implies that $|\widehat{S}_2| < m_0$. The theorem follows. $\square$

REFERENCES

[1] T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58:1300–1308, 2010.

[2] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook on the Geometry of Banach Spaces*, volume 1. 2001.

[3] F. Ye and C.-H. Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

[4] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. 2012.