A supplementary file for "A High-dimensional M-estimator Framework for Bi-level Variable Selection"

Bin Luo · Xiaoli Gao

S1 Mean squared prediction error for the cross-validation

Compared with the mean squared prediction error for the cross-validation, the trimmed version is robust to outliers in validation sets and provides a better selection of tuning parameters. To illustrate this point, we re-run the simulation using the mean squared prediction error for the cross-validation. We report the results for Example 2(i) in Table S.1. Note that the results with the trimmed mean squared prediction error are displayed in Table 2 in our paper. In the light-tailed setting ($N(0, 1)$), with similar estimation errors, it is not surprising that the mean squared prediction error for the cross-validation yields slightly better group/variable selection performance than the trimmed version, as there are not any outliers in the dataset. However, in the heavy-tailed settings (t_1 and Mix Cauchy), we can clearly see that the robust GMCP and GMCP-HT using the trimmed mean squared prediction error perform better in both parameter estimation and group/variable selection. In particular, the robust GMCP-HT method with the trimmed version is able to largely reduce the false negative rates in group/variable selection while maintaining competitive false positive rates.

S2 Proofs

To prove Theorem 1, we need the following Lemma 1.

Bin Luo
Duke University
E-mail: bin.luo2@duke.edu

Xiaoli Gao
The University of North Carolina at Greensboro

		MCP		GMCP		GMCP-HT	
		Huber	Cauchy	Huber	Cauchy	Huber	Cauchy
N(0,1)	ℓ_2 error	7.23	7.34	1.69	1.67	1.58	1.57
	ℓ_1 error	30.35	30.85	7.55	7.49	6.81	6.74
	MS	24.66	22.58	39.76	38.82	29.24	28.82
	GS	16.79	15.44	8.96	8.73	7.7	7.61
	FPR	2.77	2.41	4.71	4.52	2.53	2.45
	FNR	33.71	35.76	0	0	0	0
	GFPR	11.56	10.11	3.15	2.9	1.81	1.71
	GFNR	1.33	1	0	0	0	0
t_1	ℓ_2 error	11.33	11.36	5.15	4.53	4.99	4.44
	ℓ_1 error	46.55	47.09	22.72	19.32	22.32	19.47
	MS	11.33	10.65	37.31	34.87	32.02	31.33
	GS	9.16	9.11	7.73	7.14	8.17	8.73
	FPR	1.05	0.92	4.38	3.82	3.35	3.17
	FNR	63.24	63.59	4.88	3.53	6.71	5.71
	GFPR	4.22	4.09	2.34	1.56	2.85	3.36
	GFNR	13.5	12.17	7.83	5.5	8.5	7.17
MixCauchy	ℓ_2 error	8.65	9.14	2.92	2.73	2.84	2.7
	ℓ_1 error	36.59	38.91	12.95	11.82	12.4	11.35
	MS	19.17	16.15	35.32	34.46	27.69	26.63
	GS	13.8	12.56	7.38	7.24	7.21	7.09
	FPR	2.06	1.62	3.83	3.69	2.28	2.1
	FNR	45.76	51	1	2.24	2	2.88
	GFPR	8.44	7.12	1.56	1.51	1.45	1.4
	GFNR	2.17	2.17	1.5	3	2.5	3.83

Table S.1 Simulation results under the model with bi-level sparsity in Example 2(i), with the mean squared prediction error for the cross-validation. The mean ℓ_2 error, ℓ_1 error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed.

Lemma 1 *Suppose $\mathcal{L}_n$ in (5) satisfies Assumption 2 and the random errors and covariates satisfy Assumption 3. For any $t \in (0, n)$, we have*

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \le C_0 \sqrt{\frac{t}{n}}$$

with probability at least $1 - 2p\exp(-t)$.

Proof. The gradient of $\mathcal{L}_n$ is

$$\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i)\mathbf{x}_i l'(\epsilon_i v(\mathbf{x}_i)).$$

If Assumption 3(ii) (a) holds, then

$$\begin{aligned} E[w(\mathbf{x}_i)\mathbf{x}_i l'(\epsilon_i v(\mathbf{x}_i))] &= E[w(\mathbf{x}_i)\mathbf{x}_i l'(\epsilon_i)] \\ &= E[w(\mathbf{x}_i)\mathbf{x}_i]E[l'(\epsilon_i)] \\ &= \mathbf{0}, \end{aligned} \tag{S2.1}$$

where the second equality follows from $\epsilon_i \perp\!\!\!\perp \mathbf{x}_i$. If Assumption 3(ii) (b) is satisfied instead, we obtain

$$E[w(\mathbf{x}_i)\mathbf{x}_i l'(\epsilon_i v(\mathbf{x}_i))] = E[w(\mathbf{x}_i)\mathbf{x}_i E[l'(\epsilon_i v(\mathbf{x}_i))|\mathbf{x}_i] = \mathbf{0}. \tag{S2.2}$$

Therefore, $E[\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)] = \mathbf{0}$ under Assumption 3(ii).

Let $\mu_j = E[w(\mathbf{x}_i)x_{ij}]$, $j = 1, 2, \ldots, p$. Then we have

$$\begin{aligned} E|w(\mathbf{x}_i)x_{ij}|^m =& E|w(\mathbf{x}_i)x_{ij} - \mu_j + \mu_j|^m \\ \leq& E[2^{m-1}(|w(\mathbf{x}_i)\mathbf{x}_{ij} - \mu_j|^m + |\mu_j|^m)] \\ \leq& 2^{m-1}[E|w(\mathbf{x}_i)x_{ij} - \mu_j|^m + \tau^m] \\ \leq& 2^{m-1}[m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}) + \tau^m], \end{aligned} \tag{S2.3}$$

where $\max_{1\leq j\leq p}|\mu_j| < \tau < \infty$ and the last inequality follows from Assumption 3(i), by which $w(\mathbf{x}_i)x_{ij}$ is sub-Gaussian hence for $m > 0$ (Rivasplata (2012))

$$E|w(\mathbf{x}_i)x_{ij} - \mu_j|^m \leq m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}).$$

Next we bound the $E|w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))|^m$ from the above. By Assumption 2(i) and the bound in (S2.3), we have

$$\begin{aligned} E|w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))|^m &\leq k_1^m E|w(\mathbf{x}_i)x_{ij}|^m \\ &\leq k_1^m 2^{m-1}[m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}) + \tau^m]. \end{aligned} \tag{S2.4}$$

By taking $m = 2$ in (S2.4), we obtain

$$E|w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))|^2 \leq l_1, \tag{S2.5}$$

where $l_1 = k_1^2(8k_0^2 + 2\tau^2)$. For all integer $m \geq 3$, by equation (S2.4) we have

$$\begin{aligned} E|w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))|^m &\leq k_1^m 2^{m-1}[m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}) + \tau^m] \\ &\leq \frac{m!}{2} k_1^{m-2}(2\tau + \sqrt{2}k_0)^{m-2}[k_1^2(8k_0^2 + 2\tau^2)] \\ &= \frac{m!}{2} l_2^{m-2} l_1, \end{aligned} \tag{S2.6}$$

where $l_2 = k_1(2\tau + \sqrt{2}k_0)$. By Bernstein inequality (Proposition 2.9 of Massart (2007)) we have

$$P(|\frac{1}{n}\sum_{i=1}^n w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i)) - \frac{1}{n}\sum_{i=1}^n E[w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))]| \geq \sqrt{\frac{2l_1 t}{n}} + \frac{l_2 t}{n}) \leq 2\exp(-t).$$

Together with equation (S2.1), we have

$$P(|\frac{1}{n}\sum_{i=1}^n w(\mathbf{x}_i)x_{ij}l'(\epsilon_i v(\mathbf{x}_i))| \geq C_0\sqrt{\frac{t}{n}}) \leq 2\exp(-t)$$

for $t \in (0, n]$, where $C_0 = \sqrt{2l_1} + l_2$. It then follows from union inequality that

$$P(\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \geq C_0\sqrt{\frac{t}{n}}) \leq 2p\exp(-t).$$

□

Proof of Theorem 1

By letting $t = (1 + C_2)\log p$ in Lemma 1, we have

$$P(\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \le C_1\sqrt{\frac{\log p}{n}}) \le 1 - 2\exp(-C_2 \log p)$$

as desired for $n \ge (1 + C_2)\log p$, where $C_1 = C_0\sqrt{(1 + C_2)}$. Next we provide the proof of Theorem 1 (ii). We first suppose the existence of stationary points in the local RSC region and will establish this fact at the end of the proof. Suppose $\hat{\boldsymbol{\beta}}$ is a stationary point of program (4) such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le r$. Since $\hat{\boldsymbol{\beta}}$ is a stationary point and $\hat{\boldsymbol{\beta}}$ is feasible, we have the inequality

$$\langle \nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}) + \lambda\mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle \ge \mathbf{0}, \tag{S2.7}$$

where $\mathbf{D} := \mathrm{diag}((\sqrt{d_1}\mathbf{1}_{d_1}^T, \cdots, \sqrt{d_J}\mathbf{1}_{d_J}^T)^T)$ denotes a $p \times p$ diagonal matrix, $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{z}}_1^T, \cdots, \tilde{\boldsymbol{z}}_J^T)^T$ and $\tilde{\boldsymbol{z}}_j \in \partial\|\hat{\boldsymbol{\beta}}_j\|_2$. Recall

$$\partial\|\hat{\boldsymbol{\beta}}_j\|_2 = \begin{cases} \frac{\hat{\boldsymbol{\beta}}_j}{\|\hat{\boldsymbol{\beta}}_j\|_2} & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 \ne 0, \\ \{\boldsymbol{z} : \|\boldsymbol{z}\|_2 \le 1, \boldsymbol{z} \in \mathbb{R}^{d_j}\} & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 = 0, \end{cases}$$

for $j = 1, 2, \cdots, J$. By the convexity of $\frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2 - q_\lambda(\boldsymbol{\beta})$, we have

$$\langle \nabla q_\lambda(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle \ge q_\lambda(\boldsymbol{\beta}^*) - q_\lambda(\hat{\boldsymbol{\beta}}) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2. \tag{S2.8}$$

So together with inequality (S2.7) we obtain

$$\langle \nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}) + \lambda\mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle \ge q_\lambda(\boldsymbol{\beta}^*) - q_\lambda(\hat{\boldsymbol{\beta}}) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2.$$

Since $\langle \lambda\mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle \le \sum_{j=1}^J \sqrt{d_j}\lambda\|\boldsymbol{\beta}_j^*\|_2 - \sum_{j=1}^J \sqrt{d_j}\lambda\|\hat{\boldsymbol{\beta}}_j\|_2$, this means

$$\langle \nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle \ge \rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2. \tag{S2.9}$$

Let $\tilde{\boldsymbol{\nu}} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. From the RSC inequality (6), we have

$$\langle \nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla\mathcal{L}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \ge \gamma\|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau\frac{\log p}{n}\|\tilde{\boldsymbol{\nu}}\|_1^2. \tag{S2.10}$$

Combining inequality (S2.10) with inequality (S2.9), we then have

$$(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau\frac{\log p}{n}\|\tilde{\boldsymbol{\nu}}\|_1^2 + (\rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*)) \le \langle \nabla\mathcal{L}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\rangle. \tag{S2.11}$$

So by Holder's inequality, we conclude that

$$(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau\frac{\log p}{n}\|\tilde{\boldsymbol{\nu}}\|_1^2 + (\rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*)) \le \|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty\|\tilde{\boldsymbol{\nu}}\|_1. \tag{S2.12}$$

Assume $\lambda \geq 4\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty$ and $\lambda \geq 8\tau R\frac{\log p}{n}$, we have

$$\begin{aligned}
(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 &\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + (2R\tau\frac{\log p}{n} + \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty)\|\tilde{\boldsymbol{\nu}}\|_1 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \sum_{j=1}^{J}\sqrt{d_j}(2R\tau\frac{\log p}{n} + \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty)\|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \frac{1}{2}\sum_{j=1}^{J}\sqrt{d_j}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \frac{1}{2}(\rho_\lambda(\tilde{\boldsymbol{\nu}}) + \frac{\mu}{2}\|\tilde{\boldsymbol{\nu}}\|_2^2),
\end{aligned}$$

implying that

$$0 \leq (\gamma - \frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \leq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}). \tag{S2.13}$$

Recall $S \subseteq \{1, \cdots, J\}$ includes all indexes of important groups and $|S| = s$. By the assumption 1 for ρ, we have

$$\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) = \rho_\lambda(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_S) \geq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}_S),$$

where $\hat{\boldsymbol{\beta}}_S$ denotes the zero-padded vector in $\mathbb{R}^p$ with zeros on groups in S^c. Then starting from inequality (S2.13), we have

$$\begin{aligned}
0 &\leq (\gamma - \frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \\
&\leq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&= \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}_S) - \rho_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&\leq \rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&= \frac{3}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) \\
&= \frac{3}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}).
\end{aligned} \tag{S2.14}$$

Let A denote the group index set of the first s groups of $\tilde{\boldsymbol{\nu}}$ with largest ℓ_2 norm. Recall $d_a = \max_{1\leq j\leq J} d_j$, $d_b = \min_{1\leq j\leq J} d_j$, $d = \sqrt{\frac{d_a}{d_b}}$. By assumption 1(i) and (iv) we have

$$\begin{aligned}
0 \leq 3\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) &\leq 3\sum_{j\in S}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda) - \sum_{j\in S^c}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda) \\
&\leq 3\sum_{j\in A}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda) - \sum_{j\in A^c}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda).
\end{aligned} \tag{S2.15}$$

Let $c := \max_{j\in A^c} \|\tilde{\boldsymbol{\nu}}_j\|_2$ and define $f(t,\lambda) := \frac{t\lambda}{\rho(t,\lambda)}$ for $t,\lambda > 0$. By assumption on ρ, for any fixed $\lambda \in \mathbb{R}^+$, function $t \mapsto f(t,\lambda)$ is non-decreasing on $\mathbb{R}^+$. Thus

$$\begin{aligned}\sum_{j\in A} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda)\cdot f(c,\sqrt{d_a}\lambda) &\le \sum_{j\in A} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda)\cdot f(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda)\\ &\le \sum_{j\in A} \sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2.\end{aligned} \tag{S2.16}$$

Similarly we also obtain

$$\begin{aligned}\sum_{j\in A^c} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda)\cdot f(c,\sqrt{d_b}\lambda) &\ge \sum_{j\in A^c} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda)\cdot f(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda)\\ &\ge \sum_{j\in A^c} \sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2.\end{aligned} \tag{S2.17}$$

Combining inequality (S2.15) with (S2.16) and (S2.17) we have

$$\begin{aligned}0 &\le 3\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c})\\ &\le \frac{1}{f(c,\sqrt{d_a}\lambda)}\Big(3\sum_{j\in A}\sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{f(c,\sqrt{d_a}\lambda)}{f(c,\sqrt{d_b}\lambda)}\sum_{j\in A^c}\sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2\Big)\\ &\le 3\sum_{j\in A}\sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{f(c,\sqrt{d_a}\lambda)}{f(c,\sqrt{d_b}\lambda)}\sum_{j\in A^c}\sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2\\ &= \sqrt{d_a}\lambda\Big(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{\rho(c,\sqrt{d_b}\lambda)}{\rho(c,\sqrt{d_a}\lambda)}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2\Big)\\ &\le \sqrt{d_a}\lambda\Big(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - g(d)^{-1}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2\Big),\end{aligned} \tag{S2.18}$$

where the third inequality follows from

$$f(c,\sqrt{d_a}\lambda) \ge \lim_{r\to 0^+} f(r,\sqrt{d_a}\lambda) = \lim_{r\to 0^+} \frac{(r-0)\sqrt{d_a}\lambda}{\rho(r,\sqrt{d_a}\lambda)-\rho(0,\sqrt{d_a}\lambda)} = 1,$$

and the last inequality follows from assumption 1(ii). Hence,

$$3g(d)\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \ge \sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2,$$

implying that

$$
\begin{aligned}
\|\tilde{\boldsymbol{\nu}}\|_1 &\leq \sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_1 + \sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_1 \\
&\leq \sum_{j\in A}\sqrt{d_a}\|\tilde{\boldsymbol{\nu}}_j\|_2 + \sum_{j\in A^c}\sqrt{d_a}\|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq \sqrt{d_a}(1+3g(d))\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq \sqrt{d_a s}(1+3g(d))\|\tilde{\boldsymbol{\nu}}\|_2.
\end{aligned} \tag{S2.19}
$$

Combing inequalities (S2.14) and (S2.18) then gives

$$
(\gamma - \frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \leq \frac{1}{2}\sqrt{d_a}\lambda(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - g(d)^{-1}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2) \leq \frac{3}{2}\sqrt{d_a}\lambda\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \leq \frac{3}{2}\sqrt{d_a s}\lambda\|\tilde{\boldsymbol{\nu}}\|_2,
$$

from which we conclude that

$$
\|\tilde{\boldsymbol{\nu}}\|_2 \leq \frac{6\sqrt{d_a}\lambda\sqrt{s}}{4\gamma - 3\mu} \tag{S2.20}
$$

as wanted. Combining the ℓ_2-bound with inequality (S2.19) yields the ℓ_1 bound

$$
\|\tilde{\boldsymbol{\nu}}\|_1 \leq \frac{6(1+3g(d))d_a\lambda s}{4\gamma - 3\mu}. \tag{S2.21}
$$

Finally, in order to establish the existence of local stationary points, we simply define $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that

$$
\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq r, \|\boldsymbol{\beta}\|_1 < R} \left\{\mathcal{L}_n(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta})\right\}. \tag{S2.22}
$$

Then $\hat{\boldsymbol{\beta}}$ is a stationary point of program (S2.22). Therefore, we have

$$
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{\frac{d_a s\log p}{n}}.
$$

Provided $n > Cr^{-2}d_a s\log p$, the point $\hat{\boldsymbol{\beta}}$ will lie in the interior of the sphere of radius r around $\boldsymbol{\beta}^*$. Hence, $\hat{\boldsymbol{\beta}}$ is also a stationary point of the original program (4), guaranteeing the existence of such local stationary points. □

To prove Theorem 2, we need the following result adopted directly from the Lemma 1 in Loh (2017).

Lemma 2 *Suppose $\mathcal{L}_n$ satisfies the local RSC condition (4) and $n \geq \frac{2\tau}{\gamma}k\log p$. Then $\mathcal{L}_n$ is strongly convex over the region $S_r := \{\boldsymbol{\beta} \in \mathbb{R}^p : supp(\boldsymbol{\beta}) \subseteq I_S, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$.*

Proof. The proof is similar to the proof of Lemma 1 in Loh (2017). □

Proof of Theorem 2
The proof is an adaptation of the arguments of Theorem 2 in the paper Loh (2017). We use the following three steps of the primal-dual witness (PDW) construction:

(i) Optimize the restricted program

$$\hat{\boldsymbol{\beta}}_{I_S} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{I_S}:\|\boldsymbol{\beta}\|_1\le R}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j\in S} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda) \right\}, \tag{S2.23}$$

and establish that $\|\hat{\boldsymbol{\beta}}_{I_S}\|_1 < R$.

(ii) Recall $q_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^J \sqrt{d_j}\lambda\|\boldsymbol{\beta}_j\|_2 - \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2\sqrt{d_j}\lambda)$ defined in Section 2. Define $\hat{\boldsymbol{z}}_j \in \partial\|\hat{\boldsymbol{\beta}}_j\|_2$ and let $\hat{\boldsymbol{z}}_{I_S} = (\hat{\boldsymbol{z}}_j^T, j \in S)^T$, and choose $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_{I_S}^T, \hat{\boldsymbol{z}}_{I_S^c}^T)^T$ to satisfy the zero-subgradient condition

$$\nabla\mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}) + \lambda\mathbf{D}\hat{\boldsymbol{z}} = \mathbf{0}, \tag{S2.24}$$

where $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_{I_S}, \mathbf{0}_{I_S^c})$ and $\mathbf{D} = \operatorname{diag}((\sqrt{d_1}\mathbf{1}_{d_1}^T, \cdots, \sqrt{d_J}\mathbf{1}_{d_J}^T)^T)$. Show that $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}_{I_S}^{\mathcal{O}}$ and establish strict dual feasibility: $\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2 < 1$.

(iii) Verify via second order conditions that $\hat{\boldsymbol{\beta}}$ is a local minimum of program (4) and conclude that all stationary points $\hat{\boldsymbol{\beta}}$ satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le r$ are supported on I_S and agree with $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$.

Proof of Step (i) : By applying Theorem 1 to the restricted program (S2.23), we have

$$\|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \le \frac{6(1+3g(d))d_a\lambda s}{4\gamma - 3\mu},$$

and thus

$$\|\hat{\boldsymbol{\beta}}_{I_S}\|_1 \le \|\boldsymbol{\beta}^*\|_1 + \|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \le \frac{R}{2} + \|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \le \frac{R}{2} + \frac{6(1+3g(d))d_a\lambda s}{4\gamma - 3\mu} < R,$$

under the assumption of the theorem. This complete step (i) of the PDW construction. □

To prove step (ii), we need the following Lemma 3 and 4:

Lemma 3 *Under the conditions of Theorem 2, we have the bound*

$$\|\hat{\boldsymbol{\beta}}_{I_S}^{\mathcal{O}} - \boldsymbol{\beta}^*_{I_S}\|_2 \le C_5\sqrt{\frac{\log p}{kn}}$$

and $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}_{I_S}^{\mathcal{O}}$ *with probability at least* $1 - 2\exp(-C_4\log p/k^2)$.

Proof. Recall $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}, \mathbf{0}_{I_S^c})$. By the optimality of the oracle estimator, we have

$$\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}) \le \mathcal{L}_n(\boldsymbol{\beta}^*). \tag{S2.25}$$

Consider $n \ge \frac{2\tau}{\gamma} k \log p$. By Lemma 2 $\mathcal{L}_n(\boldsymbol{\beta})$ is strongly convex over restricted region S_r. Hence,

$$\mathcal{L}_n(\boldsymbol{\beta}^*) + \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^* \rangle + \frac{\gamma}{4} \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 \le \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}). \tag{S2.26}$$

Together with inequality (S2.25) we obtain

$$\begin{aligned}\tfrac{\gamma}{4} \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 &\le \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathcal{O}} \rangle \le \|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_1 \\ &\le \sqrt{k} \|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2,\end{aligned}$$

implying that

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \le \frac{4\sqrt{k}}{\gamma} \|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty. \tag{S2.27}$$

By applying Lemma 1 to the restricted program (S2.23), we have

$$P(\|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*_{I_S}))\|_\infty \le C_0 \sqrt{\frac{t}{n}}) \ge 1 - 2k \exp(-t).$$

Let $t = C_3 \log p / k^2$. Then we obtain

$$P(\|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*_{I_S}))\|_\infty \le C_0 \sqrt{C_3} \sqrt{\frac{\log p}{k^2 n}}) \ge 1 - 2 \exp(-C_4 \log p / k^2), \tag{S2.28}$$

where we require $k^2 \log k = \mathcal{O}(\log p)$. Combining inequality (S2.27) and (S2.28), we obtain

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \le C_5 \sqrt{\frac{\log p}{kn}} \tag{S2.29}$$

as desired, where $C_5 = 4C_0 \sqrt{C_3}/\gamma$.

Next we show $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$. When $n > C_5^2 / r^2 \log p / k$, we have $\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 < r$ and thus $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$ is an interior point of the oracle program in (8), implying

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) = \mathbf{0}. \tag{S2.30}$$

By assumption we have $\lambda = C_6 \sqrt{\frac{\log p}{n}}$ and $\boldsymbol{\beta}^{*G}_{\min} \ge C_8 \sqrt{\frac{d_a \log p}{n}}$, where we choose $C_8 = C_6 \delta + C_5$. Together with inequality (S2.29), we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j\|_2 \ge \|\boldsymbol{\beta}^*_j\|_2 - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j - \boldsymbol{\beta}^*_j\|_2 &\ge \boldsymbol{\beta}^{*G}_{\min} - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \\ &\ge (C_6 \delta + C_5) \sqrt{\tfrac{d_a \log p}{n}} - C_5 \sqrt{\tfrac{\log p}{kn}} \\ &\ge \sqrt{d_a} \delta \lambda.\end{aligned}$$

for all $j \in S$. Together with the assumption that ρ is (μ, δ)-amenable, we have

$$\nabla q_\lambda(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) = \lambda \mathbf{D}_{I_S I_S} \hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S}, \tag{S2.31}$$

where $\hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S} = ((\hat{\boldsymbol{z}}^{\mathcal{O}}_j)^T, j \in S)^T$ and $\hat{\boldsymbol{z}}^{\mathcal{O}}_j \in \partial \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j\|_2$. Combining equation (S2.30) and (S2.31), we obtain

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) + \lambda \mathbf{D}_{I_S I_S} \hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S} = \mathbf{0}. \tag{S2.32}$$

Hence $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$ satisfies the zero-subgradient condition for the restricted program (S2.23). By step (i) $\hat{\boldsymbol{\beta}}_{I_S}$ is an interior point of the program (S2.23), then it must also satisfy the same zero-subgradient condition. Under the strict convexity in Lemma 4, the solution that satisfies the zero-subgradient condition is unique. Thus, we obtain $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$. □

The following lemma guarantees that the program in (S2.23) is strictly convex:

Lemma 4 *Suppose $\mathcal{L}_n$ satisfies the local RSC condition (4) and ρ is μ-amenable with $\gamma > \mu$. Suppose in addition the sample size satisfies $n > \frac{2\tau}{\gamma - \mu} k \log p$, then the restricted program in (S2.23) is strictly convex.*

Proof. This is almost identical to the proof of Lemma 2 in Loh et al. (2017). We refer the reader to the arguments provided in that paper. □

Proof of step (ii) : We rewrite the zero-subgradient condition (S2.24) as

$$\left(\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\right) + \left(\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\right) + \lambda \mathbf{D} \hat{\boldsymbol{z}} = \mathbf{0}.$$

Let $\hat{Q}$ be a $p \times p$ matrix $\hat{Q} = \int_0^1 \nabla^2 \mathcal{L}_n \left(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right) dt$. By the zero-subgradient condition and the fundamental theorem of calculas, we have

$$\hat{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \left(\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\right) + \lambda \mathbf{D} \hat{\boldsymbol{z}} = \mathbf{0},$$

And its block form is

$$\begin{bmatrix} \hat{Q}_{I_S I_S} & \hat{Q}_{I_S I_S^c} \\ \hat{Q}_{I_S^c I_S} & \hat{Q}_{I_S^c I_S^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S} - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S}) \\ \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c} - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S^c}) \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{D}_{I_S I_S} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{I_S^c I_S^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{z}}_{I_S} \\ \hat{\boldsymbol{z}}_{I_S^c} \end{bmatrix} = \mathbf{0}. \tag{S2.33}$$

The selection property implies $\nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S^c}) = \mathbf{0}$. Plugging this result into equation (S2.33) and performing some algebra, we conclude that

$$\mathbf{D}_{I_S^c I_S^c} \hat{\boldsymbol{z}}_{I_S^c} = \frac{1}{\lambda} \left\{ \hat{Q}_{I_S^c I_S} (\boldsymbol{\beta}^*_{I_S} - \hat{\boldsymbol{\beta}}_{I_S}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c} \right\}. \tag{S2.34}$$

Therefore,

$$\begin{aligned}\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2 &\le \max_{j\in S^c}\sqrt{d_j}\|\hat{\boldsymbol{z}}_j\|_\infty\\ &= \|\mathbf{D}_{I_S^cI_S^c}\hat{\boldsymbol{z}}_{I_S^c}\|_\infty\\ &= \tfrac{1}{\lambda}\|\hat{Q}_{I_S^cI_S}(\hat{\boldsymbol{\beta}}_{I_S}-\boldsymbol{\beta}^*_{I_S})-\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty\\ &\le \tfrac{1}{\lambda}\|\hat{Q}_{I_S^cI_S}(\hat{\boldsymbol{\beta}}_{I_S}-\boldsymbol{\beta}^*_{I_S})\|_\infty+\tfrac{1}{\lambda}\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty\\ &\le \tfrac{1}{\lambda}\left\{\max_{j\in I_S^c}\|\mathbf{e}_j^T\hat{Q}_{I_S^cI_S}\|_2\right\}\|(\hat{\boldsymbol{\beta}}_{I_S}-\boldsymbol{\beta}^*_{I_S})\|_2+\tfrac{1}{\lambda}\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty,\end{aligned} \tag{S2.35}$$

where $\mathbf{e}_j$ is a standard unit vector with jth element being 1. Observe that

$$\begin{aligned}[(\mathbf{e}_j^T\hat{Q}_{I_S^cI_S})_m]^2 &\le [\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)\mathbf{x}_{ij}v(\mathbf{x}_i)\mathbf{x}_{im}\int_0^1 l''((y_i-\mathbf{x}_i^T\boldsymbol{\beta}^*-t(\mathbf{x}_i\hat{\boldsymbol{\beta}}-\mathbf{x}_i\boldsymbol{\beta}^*))v(\mathbf{x}_i))dt]^2\\ &\le k_2^2[\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)\mathbf{x}_{ij}\cdot v(\mathbf{x}_i)\mathbf{x}_{im}]^2,\end{aligned}$$

for all $j\in I_S^c$ and $m\in I_S$, where the last inequality follows from assumption 2(ii). By conditions of Theorem 2, the variables $w(\mathbf{x}_i)\mathbf{x}_{ij}$ and $v(\mathbf{x}_i)\mathbf{x}_{im}$ are both sub-Gaussian. Using standard concentration results for i.i.d sums of products of sub-Gaussian variables, we have

$$P([(e_j^T\hat{Q}_{I_S^cI_S})_m]^2\le C_3')\ge 1-C_2'\exp(-C_3'n).$$

It then follows from union inequality that

$$P(\max_{j\in I_S^c}\|e_j^T\hat{Q}_{I_S^cI_S}\|_2\le\sqrt{C_3'k})\ge 1-C_2'\exp(-C_3'n+\log(k(p-k)))\ge 1-C_2'\exp(-\frac{C_3'}{2}n), \tag{S2.36}$$

where $n\ge\frac{2}{C_3'}\log(k(p-k))$. By Lemma 3 we obtain

$$\|\hat{\boldsymbol{\beta}}_{I_S}-\boldsymbol{\beta}^*_{I_S}\|_2\le C_5\sqrt{\frac{\log p}{kn}}. \tag{S2.37}$$

Furthermore, Theorem 1 gives

$$\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty\le\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*))\|_\infty\le C_1\sqrt{\frac{\log p}{n}}. \tag{S2.38}$$

Combining inequality (S2.35), (S2.36), (S2.37) and (S2.38), we have

$$\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2\le\frac{1}{\lambda}C_6'\sqrt{\frac{\log p}{n}},$$

with probability at least $1-C_7\exp(-C_4\log p/k^2)$, where $C_6'=\sqrt{C_3'}C_5+C_1$. In particular, for $\lambda=C_6\sqrt{\frac{\log p}{n}}$ for some $C_6>C_6'$, we conclude at last that the strict dual feasibility condition $\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2<1$ holds, completing step (ii) of the PDW construction.

Step (iii) : Since the proof for this step is almost identical to the proof in Step (iii) of Theorem 2 in Loh (2017), except for the slightly different notations. We refer the reader to the arguments provided in that paper. □

Proof of Theorem 3

By the condition that $\boldsymbol{\beta}^{*I}_{\min} \geq C_5\sqrt{\frac{\log p}{kn}} + \theta$, we have

$$\begin{aligned}|\hat{\beta}^{\mathcal{O}}_j| \geq |\beta^*_j| - |\hat{\beta}^{\mathcal{O}}_j - \beta^*_j| &\geq \boldsymbol{\beta}^{*I}_{\min} - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_\infty \\ &\geq (C_5\sqrt{\tfrac{\log p}{kn}} + \theta) - C_5\sqrt{\tfrac{\log p}{kn}} \\ &= \theta.\end{aligned} \tag{S2.39}$$

for all $j \in I_0$, where the second inequality follows from Lemma 3. For $j \in I_S - I_0$,

$$|\hat{\beta}^{\mathcal{O}}_j| \leq \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_\infty \leq C_5\sqrt{\frac{\log p}{kn}} < \theta, \tag{S2.40}$$

where the second inequality follows from Lemma 3 and the last inequality follows from the condition in Theorem 3. Recall $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}, \mathbf{0}_{I_S^c})$. By Theorem 2 we have $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}$ with probability at least $1 - C_7\exp(-C_4\log p/k^2)$. Together with (S2.39) and (S2.40), we have

$$\hat{\boldsymbol{\beta}}^h(\theta) = \hat{\boldsymbol{\beta}} \cdot I(|\hat{\boldsymbol{\beta}}| \geq \theta) = \hat{\boldsymbol{\beta}}^{\mathcal{O}} \cdot I(|\hat{\boldsymbol{\beta}}^{\mathcal{O}}| \geq \theta) = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_0}, \mathbf{0}_{I_0^c}),$$

as desired. It then gives the result

$$\|\hat{\boldsymbol{\beta}}^h(\theta) - \boldsymbol{\beta}^*\|_2 \leq \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 \leq C_5\sqrt{\frac{\log p}{kn}},$$

where the last inequality follows from Lemma 3. □

References

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L., Wainwright, M. J., et al. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482.

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

Rivasplata, O. (2012). Subgaussian random variables: An expository note. *Internet publication, PDF.*