Supporting Information for "Building Generalized Linear Models with Ultrahigh Dimensional Features: A Sequentially Conditional Approach" by Qi Zheng, Hyokyoung G. Hong, and Yi Li

A: Proofs of main theorems

The proofs of the main theorems and corollaries are contained in this section.

Proof of Theorem 3.1: Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d_1) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leqslant d_1/(K\sqrt{s})\}$ and $\mathcal{B}_{r,S}^1(d_2) = \{\beta_r : |\beta_r - \beta_{r|S}^*| \leqslant d_2/K\}$, where $d_1 = A_4\sqrt{\rho^3 \log p/n}$ and $d_2 = A_6\sqrt{\rho^3 \log p/n}$ with A_4 and A_6 defined as in Lemma 6.

We first define an event

$$
\begin{aligned}
\Omega_3 := \Big\{ & \sup_{|S|\leqslant\rho, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1)} \left|\mathbb{G}_n \left\{ l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right\}\right| \leqslant 2A_3 d_1\sqrt{\rho\log p}, \\
& \sup_{|S|<\rho, r\in S^c, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1), \beta_r\in\mathcal{B}_{r,S}^1(d_2)} \Big|\mathbb{G}_n \Big\{ l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_r X_r\right) \\
& \qquad\qquad - l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta_{r|S}^* X_r\right)\Big\}\Big| \leqslant 2A_3(d_1+d_2)\sqrt{\rho\log p}, \\
& \max_{|S|\leqslant\rho} \left|\mathbb{G}_n \left\{ l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right\}\right| \leqslant 7(A_2KL + b_{\max})\sqrt{\rho\log p}, \\
& \max_{|S|<\rho, r\in S^c} \left|\mathbb{G}_n \left\{ l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta_{r|S}^* X_r, Y)\right\}\right| > 7(2A_2KL + b_{\max})\sqrt{\rho\log p} \Big\},
\end{aligned}
$$

where A_2 and A_3 are defined as in Lemma 4. By Lemma 4, $P(\Omega_3) \geqslant 1 - 24\exp(-6\rho\log p)$. In the rest of the proof, we consider the sample points in Ω_3.

In the proof of Lemma 6, we show that $\max_{|S|\leqslant\rho} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leqslant A_4 K^{-1}(\rho^2\log p/n)^{1/2}$ almost surely given Ω_3. Given an index set S and $\boldsymbol{\beta}_S$ such that $|S| < \rho, \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leqslant$

$A_4K^{-1}(\rho^2\log p/n)^{1/2}$, and for any $j\in S^c$,

$$\begin{aligned}
&\ell_{S\cup\{j\}}(\beta^*_{j|S}|\boldsymbol{\beta}_S)-\ell_S(\boldsymbol{\beta}_S)\\
&=n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\beta^*_{j|S}X_j,Y)-l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S+\beta^*_{j|S}X_j,Y)\right\}\\
&\quad+n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S+\beta^*_{j|S}X_j,Y)\right\}+E\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\beta^*_{j|S}X_j,Y)\right\}-E\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S,Y)\right\}\\
&\quad\quad-n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S,Y)-l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y)\right\}-n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y)\right\}\\
&\geqslant-2A_3A_4\rho^2\log p/n-7(2A_2KL+b_{\max})\sqrt{\rho\log p/n}+E\left[\ell_{S\cup\{j\}}(\beta^*_{j|S}|\boldsymbol{\beta}_S)\right]-E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\\
&\quad\quad\quad-7(A_2KL+b_{\max})\sqrt{\rho\log p/n}-2A_3(A_4+A_6)\rho^2\log p/n\\
&\geqslant-\sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}|\beta^*_{j|S}|+\sigma_{\min}\beta^{*2}_{j|S}/2\\
&\quad\quad\quad-7(3A_2KL+2b_{\max})\sqrt{\rho\log p/n}-2A_3(2A_4+A_6)\rho^2\log p/n,
\end{aligned}$$

where the first inequality follows from the definition of Ω_3 and the last inequality follows from part (iii) of Lemma 5. Thus,

$$\begin{aligned}
&\ell_{S\cup\{j\}}(\beta^*_{j|S}|\widehat{\boldsymbol{\beta}}_S)-\ell_S(\widehat{\boldsymbol{\beta}}_S)\geqslant\inf_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\leqslant A_4K^{-1}(\rho^2\log p/n)^{1/2}}\ell_{S\cup\{j\}}(\beta^*_{j|S}|\boldsymbol{\beta}_S)-\ell_S(\boldsymbol{\beta}_S)\\
&\geqslant-\sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}|\beta^*_{j|S}|+\sigma_{\min}\beta^{*2}_{j|S}/2\\
&\quad\quad-7(3A_2KL+2b_{\max})\sqrt{\rho\log p/n}-2A_3(2A_4+A_6)\rho^2\log p/n.
\end{aligned}$$

By Lemma 1, if $\mathcal{M}\not\subseteq S$, $\exists r\in S^c\cap\mathcal{M}$, such that $|\beta^*_{r|S}|\geqslant C\sigma_{\max}^{-1}n^{-\alpha}$. Thus, there exists some constant C_1 that does not depend on n such that

$$\begin{aligned}
&\max_{j\in S^c}\ell_{S\cup\{j\}}(\beta^*_{j|S}|\widehat{\boldsymbol{\beta}}_S)-\ell_S(\widehat{\boldsymbol{\beta}}_S)\\
&\geqslant C^2\sigma_{\min}\sigma_{\max}^{-2}n^{-2\alpha}/2-\sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}C\sigma_{\max}^{-1}n^{-\alpha}\\
&\quad\quad-7(3A_2KL+2b_{\max})\sqrt{\rho\log p/n}-2A_3(2A_4+A_6)\rho^2\log p/n\geqslant C_1n^{-2\alpha}
\end{aligned}$$

provided $\rho n^{-1+4\alpha}\log p\to 0$. Moreover, we obtain that

$$\begin{aligned}
&\min_{|S|<\rho,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}\{(\widehat{\boldsymbol{\beta}}_S^{\mathrm{T}},\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S))^T\}-\ell_S(\widehat{\boldsymbol{\beta}}_S)=\min_{|S|<\rho,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}\{\widehat{\beta}_{j|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\}-\ell_S(\widehat{\boldsymbol{\beta}}_S)\\
&\geqslant\min_{|S|<\rho,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}(\beta^*_{j|S}|\widehat{\boldsymbol{\beta}}_S)-\ell_S(\widehat{\boldsymbol{\beta}}_S)\geqslant C_1n^{-2\alpha},
\end{aligned}$$

where the inequality follows from $\widehat{\boldsymbol{\beta}}_{j|S}(\widehat{\beta}_S)$ being the maximizer of $\ell_{S\cup\{j\}}(\beta_j|\widehat{\boldsymbol{\beta}}_S)$.

Withdrawing the restriction to Ω_3, we obtain that

$$P\left[\min_{|S|<\rho,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}\{\widehat{\beta}_{j|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\}-\ell_S(\widehat{\boldsymbol{\beta}}_S)\geqslant C_1n^{-2\alpha}\right]\geqslant 1-24\exp(-6\rho\log p).$$

This completes the proof of Theorem 3.1. $\square$

Proof of Corollary 3.1: Define

$$\Omega_4:=\left\{\min_{|S|<\rho,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}\{\widehat{\beta}_{j|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\}-\ell_S(\widehat{\boldsymbol{\beta}}_S)\geqslant C_1n^{-2\alpha}\right\},$$
$$\Omega_5:=\left\{\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},Y)\right\}\right|\leqslant(\sqrt{2}M+2\mu_{\max})\tau KL+b_{\max}\right\}.$$

By Theorem 3.1 and Lemma 3, the event $\Omega_4\cap\Omega_5$ holds with probability at least $1-26\exp(-6\rho\log p)$. We thus restrict our attention to the event $\Omega_4\cap\Omega_5$.

Given any S such that $|S|<\rho,\mathcal{M}\not\subseteq S$, let r be the index selected by SC. Then given $\Omega_4\cap\Omega_5$, $\ell_{S\cup\{r\}}(\widehat{\boldsymbol{\beta}}_{S\cup\{r\}})-\ell_S(\widehat{\boldsymbol{\beta}}_S)\geqslant C_1n^{1-2\alpha}$. If $\rho n^{-1+4\alpha}\log p\to 0$, then $n^{-1}(\log n+2\eta\log p)=o(n^{-2\alpha})$ and thus,

$$\begin{aligned}&\mathrm{EBIC}(S\cup\{r\})-\mathrm{EBIC}(S)\\&=-2\ell_{S\cup\{r\}}(\widehat{\boldsymbol{\beta}}_{S\cup\{r\}})+(|S|+1)(\log n+2\eta\log p)/n-\{-2\ell_S(\widehat{\boldsymbol{\beta}}_S)+|S|(\log n+2\eta\log p)/n\}\\&\leqslant-2C_1n^{-2\alpha}+(\log n+2\eta\log p)/n<0,\end{aligned}$$

when n is sufficiently large. Therefore, our proposed SC does not stop when $\mathcal{M}\not\subseteq S_k$ and $|S_k|<\rho$. Noting that

$$2(\sqrt{2}M+2\mu_{\max})\tau KL+2b_{\max}\geqslant\sup_{\boldsymbol{\beta}\in\mathbb{B}}\mathbb{E}_n\left\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},Y)\right\}-\inf_{\boldsymbol{\beta}\in\mathbb{B}}\mathbb{E}_n\left\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},Y)\right\}$$
$$\geqslant\ell_{S_k}(\widehat{\boldsymbol{\beta}}_{S_k})-\ell_{S_0}(\widehat{\boldsymbol{\beta}}_{S_0})\geqslant\sum_{1\leqslant t\leqslant k}\{\ell_{S_t}(\widehat{\boldsymbol{\beta}}_{S_t})-\ell_{S_{t-1}}(\widehat{\boldsymbol{\beta}}_{S_{t-1}})\}\geqslant kC_1n^{-2\alpha},$$

we have that $\mathcal{M}\not\subseteq S_N$ implies $2C_1^{-1}\big\{(\sqrt{2}M+2\mu_{\max})\tau KL+b_{\max}\big\}n^{2\alpha}>N$, which contradicts the definition of N. Hence, we have some $k\leqslant N$ such that $\mathcal{M}\subset S_k$ with probability at least $1-26\exp(-6\rho\log p)$. This completes the proof of Corollary 3.1. $\square$

Proof of Theorem 3.2: In the proof of Corollary 3.1, we have shown that, with probability going to 1, SC will not stop when $\mathcal{M} \not\subseteq S$ and $|S| < \rho$.

For any $r \in S^c \cap \mathcal{M}^c$, $\beta^*_{r|S}$ is the maximizer of $E\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}^*_S)\}$. Hence, by the concavity of $E\left[\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}^*_S)\right]$, $\beta^*_{r|S}$ is the unique solution to the equation $E\left[\left\{Y - \mu\left(\boldsymbol{\beta}_S^{*T}\mathbf{X}_S + \beta_r X_r\right)\right\}X_r\right]$ $= 0$. By the mean value theorem,

$$
\begin{aligned}
&E\left[\left\{Y - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right\}X_r\right] = E\left[\left\{\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right\}X_r\right]\\
&= E\left[\left\{\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right\}X_r\right] - E\left[\left\{\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r\right)\right\}X_r\right]\\
&= \beta^*_{r|S}E\left\{\sigma\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r\right)X_r^2\right\},
\end{aligned}
$$

where $\widetilde{\beta}_r$ is some point between 0 and $\beta^*_{r|S}$.

By Conditions (A) and (B), $\left|\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r\right| \leqslant \|\boldsymbol{\beta}^*_S\|_1\|\mathbf{X}_S\|_\infty + |\widetilde{\beta}_r||X_r| \leqslant 2KL$. Thus, $|\sigma(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r)| \geqslant \sigma_{\min}$ and

$$
o(n^{-\alpha}) = \left|E\left[\left\{Y - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right\}X_r\right]\right| = \left|\beta^*_{r|S}E\left\{\sigma\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r\right)X_r^2\right\}\right| \geqslant \sigma_{\min}\left|\beta^*_{r|S}\right|.
$$

Therefore, $|\beta^*_{r|S}| = o(n^{-\alpha})$ and consequently $\max_{S:|S|\leqslant\rho, r\in S^c\cap\mathcal{M}^c}|\beta^*_{r|S}| = o(n^{-\alpha})$.

Under Ω_3 that is defined in Theorem 3.1, $\max_{|S|\leqslant\rho}\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\| \leqslant A_4K^{-1}(\rho^2\log p/n)^{1/2}$ almost surely. For any $r \in S^c$,

$$
\begin{aligned}
&\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S) - \ell_S(\boldsymbol{\beta}_S)\\
&= n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r, Y) - l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r, Y)\right\}\\
&\quad + n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r, Y)\right\} + E\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r, Y)\right\} - E\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y)\right\}\\
&\qquad - n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y) - l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right\} - n^{-1/2}\mathbb{G}_n\left\{l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right\}\\
&\leqslant 2A_3A_4\rho^2\log p/n + 7(2A_2KL + b_{\max})\sqrt{\rho\log p/n} + E\left[\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S)\right] - E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\\
&\qquad\qquad + 7(A_2KL + b_{\max})\sqrt{\rho\log p/n} + 2A_3(A_4 + A_6)\rho^2\log p/n\\
&\leqslant \sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}|\beta^*_{r|S}| + \sigma_{\min}\beta^{*2}_{r|S}/2\\
&\qquad\qquad + 7(3A_2KL + 2b_{\max})\sqrt{\rho\log p/n} + 2A_3(2A_4 + A_6)\rho^2\log p/n,
\end{aligned}
$$

where the first inequality follows from the definition of Ω_3 and the second inequality follows from part (iii) of Lemma 5. Thus,

$$\ell_{S\cup\{r\}}(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S) - \ell_S(\widehat{\boldsymbol{\beta}}_S) \leqslant \sup_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\leqslant A_4K^{-1}(\rho^2\log p/n)^{1/2}} \ell_{S\cup\{j\}}(\beta^*_{j|S}|\boldsymbol{\beta}_S) - \ell_S(\boldsymbol{\beta}_S)$$

$$\leqslant \sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}|\beta^*_{r|S}| + \sigma_{\min}\beta^{*2}_{r|S}/2$$
$$+ 7(3A_2KL + 2b_{\max})\sqrt{\rho\log p/n} + 2A_3(2A_4 + A_6)\rho^2\log p/n.$$

Since $\max_{S:|S|<\rho,r\in S^c\cap\mathcal{M}^c}|\beta^*_{r|S}| = o(n^{-\alpha})$ and $\rho n^{-1+4\alpha}\log p \to 0$,

$$\max_{r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S) - \ell_S(\widehat{\boldsymbol{\beta}}_S)$$
$$\leqslant \sigma_{\max}\lambda_{\max}A_4K^{-1}(\rho^2\log p/n)^{1/2}o(n^{-\alpha}) + \sigma_{\min}o(n^{-2\alpha})/2$$
$$+ 7(3A_2KL + 2b_{\max})\sqrt{\rho\log p/n} + 2A_3(2A_4 + A_6)\rho^2\log p/n \leqslant C_1n^{-2\alpha}/3.$$

By Part (ii) of Lemma 6, with probability at least $1 - 12\exp(-6\rho\log p)$,

$$\max_{|S|<\rho,r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}\left\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\right\} - \ell_S\left(\widehat{\boldsymbol{\beta}}_S\right)$$
$$\leqslant \max_{|S|<\rho,r\in S^c\cap\mathcal{M}^c}\left|\ell_{S\cup\{r\}}\left\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\right\} - \ell_{S\cup\{r\}}\left(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S\right)\right|$$
$$+ \max_{|S|<\rho,r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S) - \ell_S(\widehat{\boldsymbol{\beta}}_S)$$
$$\leqslant A_7\rho^2\log p/n + C_1n^{-2\alpha}/3 \leqslant C_1n^{-2\alpha}/2.$$

Withdrawing the restriction on Ω_3, we obtain that with probability at least $1-36\exp(-6\rho\log p)$,

$$\max_{|S|<\rho,r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\} - \ell_S(\widehat{\boldsymbol{\beta}}_S) \leqslant C_1n^{-2\alpha}/2.$$

Therefore, if $\mathcal{M}\not\subseteq S$, SC would select a noise variable with probability less than $36\exp(-4\rho\log p)$.

For $k > |\mathcal{M}|$, $\mathcal{M}\not\subseteq S_k$ implies that at least $k-|\mathcal{M}|$ noise variables are selected within the k steps. Then for $k = C_2|\mathcal{M}|$ with $C_2 > 1$,

$$P\left(\mathcal{M}\not\subseteq S_k\right) \leqslant \sum_{j=k-|\mathcal{M}|}^{k}\binom{k}{j}\left\{36\exp(-4\rho\log p)\right\}^j \leqslant |\mathcal{M}|k^{|\mathcal{M}|}\left\{36\exp(-4\rho\log p)\right\}^{k-|\mathcal{M}|}$$
$$\leqslant 36\exp(-4\rho\log p + \log|\mathcal{M}| + |\mathcal{M}|\log k) \leqslant 36\exp(-3\rho\log p).$$

Therefore, $\mathcal{M}\subset S_{C_2|\mathcal{M}|}$ with probability at least $1 - 36\exp(-3\rho\log p)$. This completes the proof of Theorem 3.2. $\square$

Proof of Theorem 3.3: As shown in Corollary 3.1, SC will not stop when $\mathcal{M} \not\subseteq S$ and $|S| < \rho$ with probability converging to 1. Also, by Corollary 3.1 or Theorem 3.2, $\mathcal{M}$ will be included in S_k for some $k < \rho$ with probability going to 1. Therefore, SC stops at the kth step if $\text{EBIC}(S_{k+1}) > \text{EBIC}(S_k)$.

On the other hand, it is easy to see that $\text{EBIC}(S_{k+1}) > \text{EBIC}(S_k)$ if and only if $2\ell_{S_{k+1}}(\widehat{\boldsymbol{\beta}}_{S_{k+1}}) - 2\ell_{S_k}(\widehat{\boldsymbol{\beta}}_{S_k}) \leqslant (\log n + 2\eta \log p)/n$. By Lemma 7, conditions (A5) and (A6) in Chen and Chen (2012) are satisfied with probability tending to 1. Thus, following the proof of Equation (3.2) in Chen and Chen (2012) with $|S_{k+1}| - |S_k| = 1$, we can show that with probability tending to 1,

$$2\ell_{S_{k+1}}(\widehat{\boldsymbol{\beta}}_{S_{k+1}}) - 2\ell_{S_k}(\widehat{\boldsymbol{\beta}}_{S_k}) < (\log n + 2\eta \log p)/n,$$

for all $\eta > 0$. Thus, with probability tending to 1, the procedure stops at the kth step. This completes the proof of Theorem 3.3. □

B: Additional lemmas and proofs

We state and prove several needed lemmas.

Lemma 1: *Given a model S such that $|S| < \rho, \mathcal{M} \not\subseteq S$, under Condition (E),*

(i) $\exists r \in S^c \cap \mathcal{M}$, such that $\beta^*_{r|S} \neq 0$.

(ii) in addition, if Conditions (A) and (B) hold, then $\exists r \in S^c \cap \mathcal{M}$, such that $|\beta^*_{r|S}| \geqslant C\sigma^{-1}_{\max} n^{-\alpha}$.

Proof: As $\beta^*_{j|S}$ is the maximizer of $E\{\ell_{S\cup\{j\}}(\beta_j|\boldsymbol{\beta}^*_S)\}$, by the concavity of $E\left[\ell_{S\cup\{j\}}(\beta_j|\boldsymbol{\beta}^*_S)\right]$, $\beta^*_{j|S}$ is the solution to the equation $E\left[\left\{Y - \mu\left(\boldsymbol{\beta}^{*T}_S \mathbf{X}_S + \beta_j X_j\right)\right\}X_j\right] = 0$.

(i): Suppose that $\beta^*_{j|S} = 0, \forall j \in S^c \cap \mathcal{M}$. Then,

$$0 = E\left[\left\{Y - \mu\left(\boldsymbol{\beta}^{*T}_S \mathbf{X}_S + \beta^*_{j|S} X_j\right)\right\}X_j\right] = E\left[\left\{\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*T}_S \mathbf{X}_S\right)\right\}X_j\right]$$
$$\Rightarrow \max_{j \in S^c \cap \mathcal{M}} \left|E\left[\left\{\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*T}_S \mathbf{X}_S\right)\right\}X_j\right]\right| = 0,$$

which contradicts Condition (E). Thus, $\exists r \in S^c \cap \mathcal{M}$, such that $\beta^*_{r|S} \neq 0$.

(ii): By the mean value theorem,

$$
\begin{aligned}
&E\left[\left\{Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right\}X_r\right] = E\left[\left\{\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right\}X_r\right] \\
&= E\left[\left\{\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right\}X_r\right] - E\left[\left\{\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta^*_{r|S}X_r)\right\}X_r\right] \\
&= \beta^*_{r|S} E\left\{\sigma(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r)X_r^2\right\},
\end{aligned}
$$

where $\widetilde{\beta}_r$ is some point between 0 and $\beta^*_{r|S}$.

By Conditions (A) and (B), $\left|\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r\right| \leqslant \|\boldsymbol{\beta}_S^*\|_1 \|\mathbf{X}_S\|_\infty + |\widetilde{\beta}_r||X_r| \leqslant 2KL$. Thus, $|\sigma(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r)| \leqslant \sigma_{\max}$ and

$$
Cn^{-\alpha} \leqslant \left|E\left[\left\{\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right\}X_r\right]\right| = \left|\beta^*_{r|S} E\left\{\sigma(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \widetilde{\beta}_r X_r)X_r^2\right\}\right| \leqslant \sigma_{\max}\left|\beta^*_{r|S}\right|.
$$

Therefore, $|\beta^*_{r|S}| \geqslant C\sigma_{\max}^{-1} n^{-\alpha}$. This completes the proof of Lemma 1. □

Lemma 2: *Let $\xi_i, i = 1, \ldots, n$ be n i.i.d random variables such that $|\xi_i| \leqslant B$ for a constant $B > 0$. Under Conditions (A), (B), and (C), we have $E\left(|Y_i\xi_i - E\left[Y_i\xi_i\right]|^m\right) \leqslant m!(2B(\sqrt{2}M + \mu_{\max}))^m$, for every $m \geqslant 1$.*

Proof: By Conditions (A) and (B), $|\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i| \leqslant KL$, $\forall i \geqslant 1$. Thus, $\left|\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)\right| \leqslant \mu_{\max}$ and consequently, $E\left(|Y_i|\right) \leqslant E\left\{|Y_i - \mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)| + \left|\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)\right|\right\} \leqslant E\left[|\epsilon_i|\right] + \mu_{\max} \leqslant E\left(\epsilon_i^2\right)^{1/2} + \mu_{\max} \leqslant \sqrt{2}M + \mu_{\max}$, where the last inequality follows from Condition (C). Then

$$
\begin{aligned}
&E(|Y_i|^m) = E\{|\epsilon_i + \mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)|^m\} \leqslant E\left\{\sum_{t=0}^m \binom{m}{t}|\epsilon_i|^t\left|\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)\right|^{m-t}\right\} \\
&\leqslant \sum_{t=0}^m \binom{m}{t} E\left(|\epsilon_i|^t\right)\mu_{\max}^{m-t} \leqslant \sum_{t=0}^1 \binom{m}{t} E\left(|\epsilon_i|^t\right)\mu_{\max}^{m-t} + \sum_{t=2}^m \binom{m}{t} E\left(|\epsilon_i|^t\right)\mu_{\max}^{m-t} \\
&\leqslant \mu_{\max}^m + mE\left(|\epsilon_i|\right)\mu_{\max}^{m-1} + \sum_{t=2}^m t!\binom{m}{t} M^t \mu_{\max}^{m-t} \\
&\leqslant m!\left\{\mu_{\max}^m + \sqrt{2}M\mu_{\max}^{m-1} + \sum_{t=2}^m \binom{m}{t} M^t\mu_{\max}^{m-t}\right\} \leqslant m!(\sqrt{2}M + \mu_{\max})^m,
\end{aligned}
$$

for every $m \geqslant 1$. By the same arguments, it can be shown that, for every $m \geqslant 1$,

$$E\{|Y_i\xi_i - E[Y_i\xi_i]|^m\} \leqslant E\{(|Y_i\xi_i| + |E[Y_i\xi_i]|)^m\} \leqslant E\left\{\sum_{t=0}^{m}\binom{m}{t}|Y_i\xi_i|^t|E[Y_i\xi_i]|^{m-t}\right\}$$
$$\leqslant \sum_{t=0}^{m}\binom{m}{t}E\left(|Y_i|^t\right)B^tE\left(|Y_i|\right)^{m-t}B^{m-t} \leqslant m!\{2B(\sqrt{2}M + \mu_{\max})\}^m.$$

This completes the proof of Lemma 2. $\square$

Lemma 3: *Under Conditions (A) – (C), when n is sufficiently large such that $28\sqrt{\rho\log p/n} < 1$, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leqslant (\sqrt{2}M + 2\mu_{\max})\tau KL + b_{\max}$, with probability $1 - 2\exp(-8\rho\log p)$.*

Proof: By Conditions (B), $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right| \leqslant \tau KL$. Thus,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leqslant \sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left(\left|Y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right|\right)\right| + \sup_{\boldsymbol{\beta}\in\mathbb{B}}\mathbb{E}_n\left\{\left|b(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right|\right\}$$
$$\leqslant \mathbb{E}_n\left(|Y|\right)\tau KL + b_{\max} \leqslant \left[\left|\mathbb{E}_n\left\{|Y| - E\left(|Y|\right)\right\}\right| + E\left(|Y|\right)\right]\tau KL + b_{\max}$$
$$\leqslant \left[\left|\mathbb{E}_n\left\{|Y| - E\left(|Y|\right)\right\}\right|\right]\tau KL + (\sqrt{2}M + \mu_{\max})\tau KL + b_{\max},$$

where the last inequality follows from Lemma 2.

Taking $\xi_i = 1\{Y_i > 0\} - 1\{Y_i < 0\}$ in Lemma 2, we have $E\left[\left||Y_i| - E\left[|Y_i|\right]\right|^m\right] \leqslant m!(2(\sqrt{2}M + \mu_{\max}))^m$. Let $A_1 = 2(\sqrt{2}M + \mu_{\max})$. Applying Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner (1996)) yields that

$$P\left[\left|\sum_{i=1}^{n}\{|Y_i| - E(|Y_i|)\}\right| > 7A_1\sqrt{n\rho\log p}\right] \leqslant 2\exp\left(-\frac{49A_1^2n\rho\log p}{4nA_1^2 + 14A_1^2\sqrt{n\rho\log p}}\right) \quad (1)$$
$$\leqslant 2\exp(-8\rho\log p),$$

when n is sufficiently large such that $28\sqrt{\rho\log p/n} < 1$. Thus,

$$P\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\}\right| \geqslant 2(\sqrt{2}M + \mu_{\max})\tau KL + b_{\max}\right]$$
$$\leqslant P\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\{l(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\}\right| \geqslant (7A_1\sqrt{\rho\log p/n} + \sqrt{2}M + \mu_{\max})\tau KL + b_{\max}\right]$$
$$\leqslant 2\exp(-8\rho\log p).$$

This completes the proof of Lemma 3. $\square$

Lemma 4: *Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d_1) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leqslant d_1/(K\sqrt{s})\}$ and $\mathcal{B}_{r,S}^1(d_2) = \{\beta_r : |\beta_r - \beta_{r|S}^*| \leqslant d_2/K\}$, where $d_1, d_2 < KL$ and $s = |S|$. Under Conditions (A) – (C), when n is sufficiently large such that $28\sqrt{\rho \log p/n} < 1$, we have*

(*i*) $|\mathbb{G}_n\left[l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]| \leqslant 2A_3 d_1\sqrt{\rho \log p}$, *uniformly over* $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)$ *and* $|S| \leqslant \rho$, *with probability at least* $1 - 6\exp(-6\rho\log p)$, *where* $A_3 := 7(2\sqrt{2}M + 3\mu_{\max})$.

(*ii*) $|\mathbb{G}_n\left[l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_r X_r\right) - l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta_{r|S}^* X_r\right)\right]| \leqslant 2A_3(d_1+d_2)\sqrt{\rho\log p}$, *uniformly over* $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1), \beta_r \in \mathcal{B}_{r,S}^1(d_2), r \in S^c$ *and* $|S| < \rho$, *with probability at least* $1-6\exp(-6\rho\log p)$,

(*iii*) $|\mathbb{G}_n\left[l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| \leqslant 7(A_2KL+b_{\max})\sqrt{\rho\log p}$, *uniformly over* $|S| \leqslant \rho$, *with probability at least* $1 - 6\exp(-6\rho\log p)$, *where* $A_2 := 2(\sqrt{2}M + \mu_{\max})$.

(*iv*) $|\mathbb{G}_n\left[l(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta_{r|S}^* X_r, Y)\right| \leqslant 7(2A_2KL + b_{\max})\sqrt{\rho\log p}$, *uniformly over* $r \in S^c$ *and* $|S| < \rho$, *with probability at least* $1 - 6\exp(-6\rho\log p)$.

Proof: (*i*): Let $\mathcal{R}_s(d_1)$ denote a ball with dimensionality s and radius $d_1/(K\sqrt{s})$. Then $\mathcal{B}_S^0(d_1) = \mathcal{R}_s(d_1) + \boldsymbol{\beta}_S^*$. Let $\mathcal{C}_s := \{\mathcal{C}(\boldsymbol{\xi}_k)\}$ be a collection of cubes that cover the ball $R_s(d_1)$, where $\mathcal{C}(\boldsymbol{\xi}_k)$ is a cube containing $\boldsymbol{\xi}_k$ with sides of length $d_1/(K\sqrt{s}n^2)$, and $\boldsymbol{\xi}_k$ is some point in $\mathcal{R}_s(d_1)$. Since the volume of $\mathcal{C}(\boldsymbol{\xi}_k)$ is $\{d_1/(K\sqrt{s}n^2)\}^s$ and the volume of $\mathcal{R}_s(d_1)$ is less than $\{2d_1/(K\sqrt{s})\}^s$, we need no more than $(4n^2)^s$ cubes to cover $\mathcal{R}_s(d_1)$. Thus, we can assume $|\mathcal{C}_s| \leqslant (4n^2)^s$ without loss of generality. For any $\boldsymbol{\xi} \in \mathcal{C}(\boldsymbol{\xi}_k)$, $\|\boldsymbol{\xi} - \boldsymbol{\xi}_k\| \leqslant d_1/(Kn^2)$. In addition, let $T_{1S}(\boldsymbol{\xi}) := \mathbb{E}_n[Y\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_S]$, $T_{2S}(\boldsymbol{\xi}) := \mathbb{E}_n\left[b\{(\boldsymbol{\beta}_S^* + \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\} - b(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right]$, and $T_S(\boldsymbol{\xi}) := T_{1S}(\boldsymbol{\xi}) - T_{2S}(\boldsymbol{\xi})$.

Given any $\boldsymbol{\xi} \in \mathcal{R}_s(d_1)$, we can find some $\mathcal{C}(\boldsymbol{\xi}_k) \in \mathcal{C}_s$ containing $\boldsymbol{\xi}$. It is easy to see that

$$
\begin{aligned}
|T_S(\boldsymbol{\xi}) - E\{T_S(\boldsymbol{\xi})\}| &\leqslant |T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_k)| + |T_S(\boldsymbol{\xi}_k) - E[T_S(\boldsymbol{\xi}_k)]| + |E[T_S(\boldsymbol{\xi})] - E[T_S(\boldsymbol{\xi}_k)]| \\
&=: I + II + III.
\end{aligned}
$$

We deal with III first. By the mean value theorem,

$$E\left\{T_S(\boldsymbol{\xi}_k)\right\} - E\left\{T_S(\boldsymbol{\xi})\right\} = E\left[Y(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S + b\left\{(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_k)^{\mathrm{T}}\mathbf{X}_S\right\} - b\left\{(\boldsymbol{\beta}_S^* + \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right\}\right]$$
$$= E\left\{Y(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right\} + E\left[\mu\left\{\left(\boldsymbol{\beta}_S^* + \widetilde{\boldsymbol{\xi}}\right)^{\mathrm{T}}\mathbf{X}_S\right\}(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right],$$

where $\widetilde{\boldsymbol{\xi}}$ is some point between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_k$. We bound the two items separately.

$$\left|E\left\{Y(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right\}\right| \leqslant E\left(|Y|\right) d_1/(Kn^2)\sqrt{s}K \leqslant (\sqrt{2}M + \mu_{\max})d_1\sqrt{s}/n^2, \tag{2}$$

where the first inequality follows from the fact $\boldsymbol{\xi} \in \mathcal{C}(\boldsymbol{\xi}_k)$ and Condition (B), and the second inequality follows from Lemma 2. On the other hand, $\left|E\left[\mu\left\{\left(\boldsymbol{\beta}_S^* + \widetilde{\boldsymbol{\xi}}\right)^{\mathrm{T}}\mathbf{X}_S\right\}(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right]\right| \leqslant \mu_{\max}d_1\sqrt{s}/n^2$. This, coupled with (2), yields that

$$\left|E\left\{T_S(\boldsymbol{\xi}_k)\right\} - E\left\{T_S(\boldsymbol{\xi})\right\}\right| \leqslant (\sqrt{2}M + 2\mu_{\max})d_1\sqrt{s}/n^2. \tag{3}$$

Next, we evaluate II. Since $|\mathbf{X}_{iS}^{\mathrm{T}}\boldsymbol{\xi}| \leqslant d_1$ for all $\boldsymbol{\xi} \in \mathcal{R}_s(d_1)$, by Lemma 2,

$$E\left\{\left|Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S - E\left(Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S\right)\right|^m\right\} \leqslant m!\{2(\sqrt{2}M + \mu_{\max})d_1\}^m = m!(A_2d_1)^m.$$

By Bernstein's inequality,

$$P\left[\max_{1\leqslant k\leqslant (4n^2)^s} n\left|T_{1S}(\boldsymbol{\xi}_k) - E\left\{T_{1S}(\boldsymbol{\xi}_k)\right\}\right| > 7A_2d_1\sqrt{n\rho\log p}\right]$$
$$\leqslant \left(4n^2\right)^s 2\exp\left(-\frac{1}{2}\frac{49(A_2d_1)^2\rho\log p}{2(A_2d_1)^2 + 7(A_2d_1)^2\sqrt{\rho\log p/n}}\right) \leqslant 2\exp(-8\rho\log p), \tag{4}$$

when n is sufficiently large such that $28\sqrt{\rho\log p/n} \leqslant 1$.

As $|b\{(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_k)^{\mathrm{T}}\mathbf{X}_S\} - b(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)| \leqslant \mu_{\max}d_1$, applying Bernstein's inequality again yields that

$$P\left[\max_{1\leqslant k\leqslant (4n^2)^s} n\left|T_{2S}(\boldsymbol{\xi}_k) - E\left\{T_{2S}(\boldsymbol{\xi}_k)\right\}\right| > 7\mu_{\max}d_1\sqrt{n\rho\log p}\right] \leqslant 2\exp(-8\rho\log p). \tag{5}$$

Combining (4) and (5) together

$$P\left[\max_{1\leqslant k\leqslant (4n^2)^s} n\left|T_S(\boldsymbol{\xi}_k) - E\left\{T_S(\boldsymbol{\xi}_k)\right\}\right| > A_3d_1\sqrt{n\rho\log p}\right] \leqslant 4\exp(-8\rho\log p), \tag{6}$$

where $A_3 := 7(2\sqrt{2}M + 3\mu_{\max})$.

We now assess I. Following the same arguments as used for Lemma 3,

$$P\left\{\sup_{\boldsymbol{\xi}\in\mathcal{C}(\boldsymbol{\xi}_k)}|T_S(\boldsymbol{\xi})-T_S(\boldsymbol{\xi}_k)|>(2\sqrt{2}M+3\mu_{\max})d_1\sqrt{s}/n^2\right\}\leqslant 2\exp(-8\rho\log p). \qquad (7)$$

Combining (3), (6), and (7) together yields that

$$\begin{aligned}
&P\left[\sup_{\boldsymbol{\xi}\in\mathcal{R}_s(d_1)}|T_S(\boldsymbol{\xi})-E\left\{T_S(\boldsymbol{\xi})\right\}|\geqslant 2A_3d_1\sqrt{\rho\log p/n}\right]\\
&\leqslant P\left[\sup_{\boldsymbol{\xi}\in\mathcal{R}_s(d_1)}|T_S(\boldsymbol{\xi})-E\left\{T_S(\boldsymbol{\xi})\right\}|\geqslant A_3d_1\sqrt{\rho\log p/n}+(2\sqrt{2}M+3\mu_{\max})d_1\sqrt{s}/n^2\right]\\
&\leqslant 6\exp(-8\rho\log p).
\end{aligned}$$

By the combinatoric inequality $\binom{p}{s}\leqslant(ep/s)^s$, we obtain that

$$\begin{aligned}
&P\left[\sup_{|S|\leqslant\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1)}\left|\mathbb{G}_n\left\{l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S,Y\right)-l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y\right)\right\}\right|\geqslant 2A_3d_1\sqrt{\rho\log p}\right]\\
&\leqslant\sum_{s=1}^{\rho}(ep/s)^s6\exp(-8\rho\log p)\leqslant 6\exp(-6\rho\log p).
\end{aligned}$$

(ii): Let $I(d_2)$ denote the interval $[-d_2/K,d_2/K]$. Then $\mathcal{B}^1_{r,S}(d_2)=\beta^*_{r|S}+I(d_2)$. Let $\mathcal{D}:=\{\mathcal{D}(\nu_t)\}$ be a collection of intervals that cover $I(d_2)$, where $\mathcal{D}(\nu_t)$ is an interval containing ν_t with length $d_2/(Kn^2)$, and ν_t is some point in $I(d_2)$. Then $|\mathcal{D}|\leqslant 4n^2$ and $|\nu_t|\leqslant d/K$. Since the length of $\mathcal{D}(\nu_t)$ is $d_2/(Kn^2)$ and the length of $I(d_2)$ is less than $2d_2/K$, we need no more than $(4n^2)^s$ cubes to cover $\mathcal{R}_s(d_1)$. Thus, we can assume $|\mathcal{C}_s|\leqslant(4n^2)^s$ without loss of generality. For any $\nu\in\mathcal{D}(\nu_t)$, $|\nu-\nu_t|\leqslant d_2/(Kn^2)$.

Let $T_{1Sr}(\boldsymbol{\xi},\nu):=\mathbb{E}_n\left\{Y\left(\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_S+\nu X_r\right)\right\}$, $T_{2Sr}(\boldsymbol{\xi},\nu):=\mathbb{E}_n\big[b\big\{(\boldsymbol{\beta}_S^*+\boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S+(\beta^*_{r|S}+\nu)X_r\big\}-b\big(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S+\beta^*_{r|S}X_r\big)\big]$, and $T_{Sr}(\boldsymbol{\xi},\nu):=T_{1Sr}(\boldsymbol{\xi},\nu)-T_{2Sr}(\boldsymbol{\xi},\nu)$. Given any $(\boldsymbol{\xi}^{\mathrm{T}},\nu)^{\mathrm{T}}\in\mathcal{R}_s(d_1)\times I(d_2)$, we can find a $\mathcal{C}(\boldsymbol{\xi}_k)$ in $\mathcal{C}_s$ containing $\boldsymbol{\xi}$ and a $\mathcal{D}(\nu_t)$ in $\mathcal{D}$ containing ν. Then,

$$\begin{aligned}
|T_{Sr}(\boldsymbol{\xi},\nu)-E\left\{T_{Sr}(\boldsymbol{\xi},\nu)\right\}|&\leqslant|T_{Sr}(\boldsymbol{\xi},\nu)-T_{Sr}(\boldsymbol{\xi}_k,\nu_t)|+|T_{Sr}(\boldsymbol{\xi}_k,\nu_t)-E\left\{T_{Sr}(\boldsymbol{\xi}_k,\nu_t)\right\}|\\
&\quad+|E\left\{T_{Sr}(\boldsymbol{\xi},\nu)\right\}-E\left\{T_{Sr}(\boldsymbol{\xi}_k,\nu_t)\right\}|=:IV+V+VI,
\end{aligned}$$

The items IV, V, and VI can be evaluated by the same arguments as used for I, II, and III, respectively. Thus, we omit the details here. Combining the bounds of the items IV,V,VI

yields that

$$P\left[\sup_{|S|<\rho, r\in S^c, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(d), \beta_r\in\mathcal{B}_{r,S}^1(d)} \left|\mathbb{G}_n\left\{l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\beta_r X_r\right)-l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S+\beta_{r|S}^* X_r\right)\right\}\right| \right.$$
$$\left. \geqslant 2A_3(d_1+d_2)\sqrt{\rho\log p}\right] \leqslant 6\exp(-6\rho\log p).$$

(iii) and (iv): The two parts can be easily proved following the arguments used for Lemma 3. We thus omit the details here. This completes the proof of Lemma 4. □

LEMMA 5: *Given a model S and $r\in S^c$, under Conditions (A), (B), and (D), for any $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|\leqslant L/\sqrt{s}$ and $\beta_r\in[-L,L]$,*

(i) $\sigma_{\min}\lambda_{\min}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2\leqslant E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\}-E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\leqslant\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2.$

(ii) $\sigma_{\min}(\beta_r-\beta_{r|S}^*)^2/2\leqslant E\left\{\ell_{S\cup\{r\}}(\beta_{r|S}^*|\boldsymbol{\beta}_S^*)\right\}-E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S^*)\right\}\leqslant\sigma_{\max}(\beta_r-\beta_{r|S}^*)^2/2.$

(iii)

$$\begin{aligned}
&-\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\||\beta_r-\beta_{r|S}^*|+\sigma_{\min}|\beta_r-\beta_{r|S}^*|^2/2\\
&\leqslant E\left\{\ell_{S\cup\{r\}}(\beta_{r|S}^*|\boldsymbol{\beta}_S)\right\}-E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)\right\}\\
&\leqslant\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\||\beta_r-\beta_{r|S}^*|+\sigma_{\max}|\beta_r-\beta_{r|S}^*|^2/2.
\end{aligned}$$

Proof: (i): For any $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|\leqslant L/\sqrt{s}$, $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|_1\leqslant L$. Then by Taylor's Expansion,

$$\begin{aligned}
&E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}-E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\}\\
&=E\left\{Y\mathbf{X}_S^{\mathrm{T}}-\mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\mathbf{X}_S^{\mathrm{T}}\right\}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)+\frac{1}{2}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)^{\mathrm{T}}E\left\{-\sigma(\widetilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S)\mathbf{X}_S^{\otimes 2}\right\}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)\\
&=-\frac{1}{2}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)^{\mathrm{T}}E\left\{\sigma\left(\widetilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S\right)\mathbf{X}_S^{\otimes 2}\right\}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*),
\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}_S$ is between $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_S^*$. By Condition (D),

$$\sigma_{\min}\lambda_{\min}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2\leqslant E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\}-E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\leqslant\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2.$$

(ii): Similarly, for any $\beta_r\in[-L,L]$, it can be shown that

$$\sigma_{\min}(\beta_r-\beta_{r|S}^*)^2/2\leqslant E\left\{\ell_{S\cup\{r\}}(\beta_{r|S}^*|\boldsymbol{\beta}_S^*)\right\}-E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S^*)\right\}\leqslant\sigma_{\max}(\beta_r-\beta_{r|S}^*)^2/2.$$

(iii): Noting that $E\left[\left\{Y-\mu\big(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S+\beta^*_{r|S}X_r\big)\right\}X_r\right]=0$, it can be shown that

$$
\begin{aligned}
&E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)\right\}-E\left\{\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S)\right\}\\
&=E\left[\left\{Y-\mu\big(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\beta^*_{r|S}X_r\big)\right\}X_r\right](\beta_r-\beta^*_{r|S})-\frac{1}{2}E\left\{\sigma\big(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\widetilde{\beta}_rX_r\big)X_r^2\right\}(\beta_r-\beta^*_{r|S})^2\\
&=-(\beta_r-\beta^*_{r|S})E\left\{\sigma\big(\widetilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S+\beta^*_{r|S}X_r\big)X_r\mathbf{X}_S^{\mathrm{T}}\right\}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)\\
&\qquad\qquad-E\left\{\sigma\big(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\widetilde{\beta}_{r,S}X_r\big)X_r^2\right\}(\beta_r-\beta^*_{r|S})^2/2,
\end{aligned}
$$

where $\widetilde{\boldsymbol{\beta}}_S$ is some point between $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_S^*$ and $\widetilde{\beta}_{r,S}$ is some point between β_r and $\beta^*_{r|S}$.

By Conditions (A) and (B) and the facts that $\boldsymbol{\beta}_S\in\mathbb{B}$ and $\beta_r\in[-L,L]$, simple algebra shows $|\widetilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S+\beta^*_{r|S}X_r|\leqslant 2KL$ and $|\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\widetilde{\beta}_{r,S}X_r|\leqslant 2KL$. By Condition (D) and the Cauchy-Schwartz inequality, we obtain that

$$
\begin{aligned}
&-\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\||\beta_r-\beta^*_{r|S}|-\sigma_{\max}|\beta_r-\beta^*_{r|S}|^2/2\\
&\qquad\leqslant-(\beta_r-\beta^*_{r|S})E\left\{\sigma\big(\widetilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S+\beta_rX_r\big)X_r\mathbf{X}_S^{\mathrm{T}}\right\}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)\\
&\qquad\qquad-E\left\{\sigma\big(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S+\widetilde{\beta}_{r,S}X_r\big)X_r^2\right\}(\beta_r-\beta^*_{r|S})^2/2\\
&\qquad\leqslant\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\||\beta_r-\beta^*_{r|S}|-\sigma_{\min}|\beta_r-\beta^*_{r|S}|^2/2.
\end{aligned}
$$

This completes the proof of Lemma 5. □

LEMMA 6: *Under Conditions (A) – (E),*

(i) There exist some constants A_4 and A_5 that do not depend on n, such that $\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|\leqslant A_4K^{-1}\sqrt{\rho^2\log p/n}$ and $|\ell_S(\widehat{\boldsymbol{\beta}}_S)-\ell_S(\boldsymbol{\beta}_S^*)|\leqslant A_5\rho^2\log p/n$ hold uniformly over $S:|S|\leqslant\rho$, with probability at least $1-6\exp(-6\rho\log p)$.

(ii) There exist some constants A_6 and A_7 that do not depend on n, such that $|\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)-\beta^*_{r|S}|\leqslant A_6K^{-1}\sqrt{\rho^2\log p/n}$ and $|\ell_{S\cup\{r\}}\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\}-\ell_{S\cup\{r\}}(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S)|\leqslant A_7\rho^2\log p/n$ holds, uniformly over $S:|S|<\rho$ and $r\in S^c$, with probability at least $1-12\exp(-6\rho\log p)$.

Proof: Define

$$
\Omega_1(d_1):=\left\{\sup_{|S|\leqslant\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1)}\left|\mathbb{G}_n\left\{l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S,Y\right)-l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y\right)\right\}\right|<2A_3d_1\sqrt{\rho\log p}\right\}.
$$

By Lemma 4, the event $\Omega_1(d_1)$ holds with probability at least $1-6\exp(-6\rho\log p)$. In the rest of the proof of Lemma 6, we restrict our attention on $\Omega_1(d_1)$ with $d_1 = A_4\sqrt{\rho^3\log p/n}$ for some $A_4 > 2(\sigma_{\min}\lambda_{\min})^{-1}K^2A_3$.

(i): If $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_4K^{-1}\sqrt{\rho^2\log p/n}$, then $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leqslant A_4\sqrt{\rho^3\log p/n}/(K\sqrt{s})$ and consequently, $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)$. By Lemma 5 (i),

$$\begin{aligned}
&\ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S)\\
&= \Big(\ell_S(\boldsymbol{\beta}_S^*) - E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\} - \left[\ell_S(\boldsymbol{\beta}_S) - E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\right]\Big) + \left[E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\} - E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\right)\\
&\geqslant \sigma_{\min}\lambda_{\min}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2 - 2A_3d_1\sqrt{\rho\log p/n}\\
&= \sigma_{\min}\lambda_{\min}A_4^2\rho^2\log p/(K^2n) - 2A_3A_4\rho^2\log p/n > 0.
\end{aligned}$$

Thus,

$$\inf_{|S|\leqslant\rho, \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_4K^{-1}\sqrt{\rho^2\log p/n}} \ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S) > 0.$$

By the concavity of $\ell_S(\cdot)$, $\max_{|S|\leqslant\rho}\left\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\right\| \leqslant A_4K^{-1}\sqrt{\rho^2\log p/n}$.

On the other hand, for any $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leqslant A_4K^{-1}\sqrt{\rho^2\log p/n}$,

$$\begin{aligned}
&|\ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S)|\\
&\leqslant \Big|\ell_S(\boldsymbol{\beta}_S^*) - E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\} - \left[\ell_S(\boldsymbol{\beta}_S) - E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\right]\Big| + \left|E\left\{\ell_S(\boldsymbol{\beta}_S^*)\right\} - E\left\{\ell_S(\boldsymbol{\beta}_S)\right\}\right|\\
&\leqslant \sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2 + 2A_3d_1\sqrt{\rho\log p/n} \leqslant A_5\rho^2\log p/n,
\end{aligned}$$

where $A_5 := 4\sigma_{\max}\lambda_{\max}A_4^2K^{-2} + 2A_3A_4$. As $\max_{|S|\leqslant\rho}\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leqslant A_4K^{-1}\sqrt{\rho^2\log p/n}$, we obtain that $\max_{|S|\leqslant\rho}|\ell_S(\widehat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}_S^*)| \leqslant A_5\rho^2\log p/n$. Withdrawing the restriction to $\Omega_1(d_1)$, we complete the proof of part (i).

(ii): Define

$$\begin{aligned}
\Omega_2(d_1, d_2) := \Bigg\{ &\sup_{|S|<\rho, r\in S^c, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(d), \beta_r\in\mathcal{B}_{r,S}^1(d)} \Big|\mathbb{G}_n\left\{l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_rX_r\right)\right.\\
&\left.-l\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S + \beta_{r|S}^*X_r\right)\right\}\Big| < 2A_3(d_1+d_2)\sqrt{\rho\log p}\Bigg\},
\end{aligned}$$

where $d_1 = A_4\sqrt{\rho^3 \log p/n}$ and $d_2 = A_6(\rho^3 \log p/n)^{1/2}$ for some $A_6 > 0$ satisfying $\sigma_{\min} A_6^2 K^{-2} - \sigma_{\max}\lambda_{\max} A_4 A_6 K^{-2} - 2A_3(A_4 + A_6) > 0$.

By Lemma 4, the event $\Omega_1(d_1) \cap \Omega_2(d_1, d_2)$ holds with probability at least $1 - 12\exp(-6\rho \log p)$. Thus we restrict our attention to $\Omega_1(d_1) \cap \Omega_2(d_1, d_2)$.

For any β_r satisfying $|\beta_r - \beta^*_{r|S}| = A_6 K^{-1}(\rho^2 \log p/n)^{1/2}$, $\beta_r \in \mathcal{B}^1_{r,S}(d)$ and given any $\boldsymbol{\beta}_S$ such that $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leqslant A_4 K^{-1}\sqrt{\rho^2 \log p/n}$, by part (iii) in Lemma 5,

$$
\begin{aligned}
&\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S) - \ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)\\
&= \Big(\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S) - E\left\{\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S)\right\} - \left[\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S) - E\left\{\ell_{S\cup\{r\}}(\beta_{r,S}|\boldsymbol{\beta}_S)\right\}\right]\Big)\\
&\qquad + E\left\{\ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S)\right\} - E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)\right\}\\
&\geqslant -\sigma_{\max}\lambda_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\||\beta_r - \beta^*_{r|S}| + \sigma_{\min}|\beta_r - \beta^*_{r|S}|^2/2 - 2A_3(d_1 + d_2)\sqrt{\rho \log p/n}\\
&\geqslant -\sigma_{\max}\lambda_{\max} A_4 A_6 K^{-2}\rho^2 \log p/n + \sigma_{\min} A_6^2 K^{-2}\rho^2 \log p/n\\
&\qquad\qquad - 2A_3(A_4\sqrt{\rho^3 \log p/n} + A_6\sqrt{\rho^3 \log p/n})\sqrt{\rho \log p/n} > 0.
\end{aligned}
$$

Therefore,

$$
\inf_{\substack{|S|<\rho, r\in S^c, |\beta_{r,S} - \beta^*_{r|S}| = A_6 K^{-1}(\rho^2 \log p/n)^{1/2}\\ \|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leqslant A_4 K^{-1}(\rho^2 \log p/n)^{1/2}}} \ell_{S\cup\{r\}}(\beta^*_{r|S}|\boldsymbol{\beta}_S) - \ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S) > 0.
$$

By the concavity of $\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)$,

$$
\sup_{\substack{|S|<\rho, r\in S^c,\\ \|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leqslant A_4 K^{-1}(\rho^2 \log p/n)^{1/2}}} |\widehat{\beta}_{r|S}(\boldsymbol{\beta}_S) - \beta^*_{r|S}| \leqslant A_6 K^{-1}(\rho^2 \log p/n)^{1/2}.
$$

Under $\Omega_1(d_1)$, $\max_{|S|\leqslant\rho} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\| \leqslant A_4 K^{-1}(\rho^2 \log p/n)^{1/2}$. Therefore, $\max_{|S|<\rho, r\in S^c} |\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S) - \beta^*_{r|S}| \leqslant A_6 K^{-1}(\rho^2 \log p/n)^{1/2}$.

Analogous to part (i), it can be shown that

$$
\max_{|S|<\rho, r\in S^c} |\ell_{S\cup\{r\}}\left\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\right\} - \ell_{S\cup\{r\}}\left(\beta^*_{r|S}|\widehat{\boldsymbol{\beta}}_S\right)| \leqslant A_7\rho^2 \log p/n.
$$

Withdrawing the restriction to $\Omega_1(d_1) \cap \Omega_2(d_1, d_2)$ completes the proof of Lemma 6. □

LEMMA 7: *Suppose Conditions (A) – (D) hold and* $b'''_{\max} = \sup_{|t|\leqslant\tau KL} |b'''(t)| < \infty$.

(*i*) *The conditions (A4) and (A5) in Chen and Chen (2012) are satisfied for all* S *such that* $\mathcal{M} \subseteq S$ *and* $|S| \leqslant \rho$*, with probability at least* $1 - 2\exp(-3\rho \log p)$*.*

(*ii*) *There exists some constant* A_{11} *such that* $|\mathbb{E}_n \left[\{Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}} \mathbf{X}_S)\} X_r \right]| < A_{11}\sqrt{\log p/n}$*, uniformly over* $S : \mathcal{M} \subseteq S, |S| \leqslant \rho, r \in S$*, with probability at least* $1 - \exp(-3\log p)$*.*

Proof: Given any index S such that $\mathcal{M} \subseteq S$ and $|S| \leqslant \rho$, then $\boldsymbol{\beta}_{*S}^T \mathbf{X}_S = \boldsymbol{\beta}_{*\mathcal{M}}^T \mathbf{X}_{\mathcal{M}}$, where $\boldsymbol{\beta}_{*S}$ is the subvector of $\boldsymbol{\beta}_*$ corresponding to S. Thus,

$$E\left[\{Y - \mu(\boldsymbol{\beta}_{*S}^{\mathrm{T}} \mathbf{X}_S)\} \mathbf{X}_S\right] = E\left(E\left[\{Y - \mu(\boldsymbol{\beta}_{*\mathcal{M}}^T \mathbf{X}_{\mathcal{M}})\} | \mathbf{X}_S\right] \mathbf{X}_S\right) = 0,$$

which implies $\boldsymbol{\beta}_S^* = \boldsymbol{\beta}_{*S}$.

(*i*): Given any $\boldsymbol{\pi} \in \mathbb{R}^{|S|}$, let $h(\boldsymbol{\pi}, \boldsymbol{\beta}_S) = (\sigma_{\max} K^2 |S|)^{-1} \sigma\left(\boldsymbol{\beta}_S^{\mathrm{T}} \mathbf{X}_S\right) \left(\boldsymbol{\pi}^{\mathrm{T}} \mathbf{X}_S\right)^2$. By Conditions (A) and (B), $h(\boldsymbol{\pi})$ is bounded between -1 and 1 uniformly over $\|\boldsymbol{\pi}\| = 1$ and $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)$. Define the function class $\mathcal{H}_S := \{h(\boldsymbol{\pi}, \boldsymbol{\beta}_S) : \|\boldsymbol{\pi}\| = 1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)\}$. By the arguments used for Lemma 11 in Belloni and Chernozhukov (2011) and Lemmas 2.6.15 and 2.6.17 in van der Vaart and Wellner (1996), there exists some universal constant A_8 such that the class of functions $\mathcal{H}_S$ has a VC index bounded by $A_8 s$ (for the definition of the VC index, we refer to page 85 in van der Vaart and Wellner (1996)). By Theorem 2.6.7 in van der Vaart and Wellner (1996), for any probability measure Q, there exists some universal constant A_9, such that the covering number $\sup_Q N\Big(\epsilon \|\mathcal{H}_S\|_{Q,2}, \mathcal{H}_S, L_2(Q)\Big)$ is bounded by $(A_9/\epsilon)^{2A_8 s}$ for any $\epsilon > 0$ (for the definition of covering numbers, we refer to page 83 in van der Vaart and Wellner (1996)).

Thus, by Theorem 1.1 in Talagrand (1994), there exists some constant A_{10} that depends on A_8 and A_9 only, such that $P\left[\sup_{\|\boldsymbol{\pi}\|=1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)} |\mathbb{G}_n \{h(\boldsymbol{\pi}, \boldsymbol{\beta}_S)\}| \geqslant A_{10}\sqrt{\rho \log p}\right] \leqslant \exp(-5\rho \log p)$ and consequently,

$$\begin{aligned} P\Big[\sup_{|S|=s, \|\boldsymbol{\pi}\|=1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)} & \left| \mathbb{E}_n \left\{ \sigma\left(\mathbf{X}_S^{\mathrm{T}} \boldsymbol{\beta}_S\right) \left(\boldsymbol{\pi}^{\mathrm{T}} \mathbf{X}_S\right)^2 \right\} - E\left\{ \sigma\left(\mathbf{X}_S^{\mathrm{T}} \boldsymbol{\beta}_S\right) \left(\boldsymbol{\pi}^{\mathrm{T}} \mathbf{X}_S\right)^2 \right\} \right| \\ & \geqslant A_{10} K^2 \sqrt{\rho^3 \log p/n} \Big] \leqslant \sum_{s=|\mathcal{M}|}^{\rho} \left(\frac{ep}{s}\right)^s \exp(-5\rho \log p) \leqslant \exp(-3\rho \log p). \end{aligned} \quad (8)$$

By Condition (D), $\sigma_{\min}\kappa_{\min} \leqslant \lambda_{\min}\left[E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\mathbf{X}_S^{\otimes 2}\right\}\right] \leqslant \lambda_{\min}\left[E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\mathbf{X}_S^{\otimes 2}\right\}\right] \leqslant \sigma_{\max}\kappa_{\max}$, for all $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)$ and $S : \mathcal{M} \subseteq S, |S| < \rho$. This, coupled with (8) implies that,

$$\sigma_{\min}\kappa_{\min}/2 \leqslant \lambda_{\min}\left[\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\mathbf{X}_S^{\otimes 2}\right\}\right] \leqslant \lambda_{\max}\left[\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\mathbf{X}_S^{\otimes 2}\right\}\right] \leqslant 2\sigma_{\max}\kappa_{\max},$$

uniformly over all S satisfying $\mathcal{M} \subseteq S$ and $|S| \leqslant \rho$, with probability at least $1-\exp(-3\rho\log p)$. Therefore, the condition (A4) in Chen and Chen (2012) is satisfied with probability at least $1-\exp(-3\rho\log p)$.

Noting that $\forall \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d_1)$,

$$\begin{aligned}
&\left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - \mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
\leqslant &\left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
&\quad + \left|E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
&\qquad + \left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - E\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
\leqslant &\ 2A_{10}K^2\sqrt{\rho^3\log p/n} + \mu_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{*S}\|\sqrt{s}K\lambda_{\max}.
\end{aligned}$$

Then the condition (A5) in Chen and Chen (2012) is satisfied uniformly over all S such that $\mathcal{M} \subseteq S$ and $|S| \leqslant \rho$, with probability at least $1-\exp(-3\rho\log p)$.

(ii): Part (ii) can be proved by slightly modifying the arguments used for (8). We thus omit the details. $\square$

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, 22:555–574.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer: New York.