Adversarial Contamination Meets Hard Thresholding: An Iterative Algorithm with Signal Adaptivity and Minimax Optimality

Shixiang Liu, Hanming Yang*

Renmin University of China

Supplementary Material

This supplementary material provides additional details in support of the main text. Section S1 presents additional simulations that are not included in the main text due to space limitations. Sections S3 to S9 contain the proofs of theorems and corollaries stated in Section 3 of the main paper. Section S10 describes the extension of the two-stage AC-IHT algorithm to generalized linear models and provides the proof of Theorem 6. Section S11 provides the proof of Theorem 7 on heavy-tailed regression discussed in the Discussion section of the main text.

S1 Additional Simulations

In this section, we present additional simulation results that are not included in the main text due to space limitations. These supplementary results provide further support for the theoretical findings reported in the paper.

*Corresponding author. Email: yanghanming@ruc.edu.cn

S1.1 Varying the level of sparsity and contamination

We fix $p = 1000,\ n = 600$, vary sparsity level s from 7 to 35 and contamination level o from 12 to 120. All non-zero signals take a value of 1. The ℓ_2 estimation errors of β^* are shown in Figure 1.

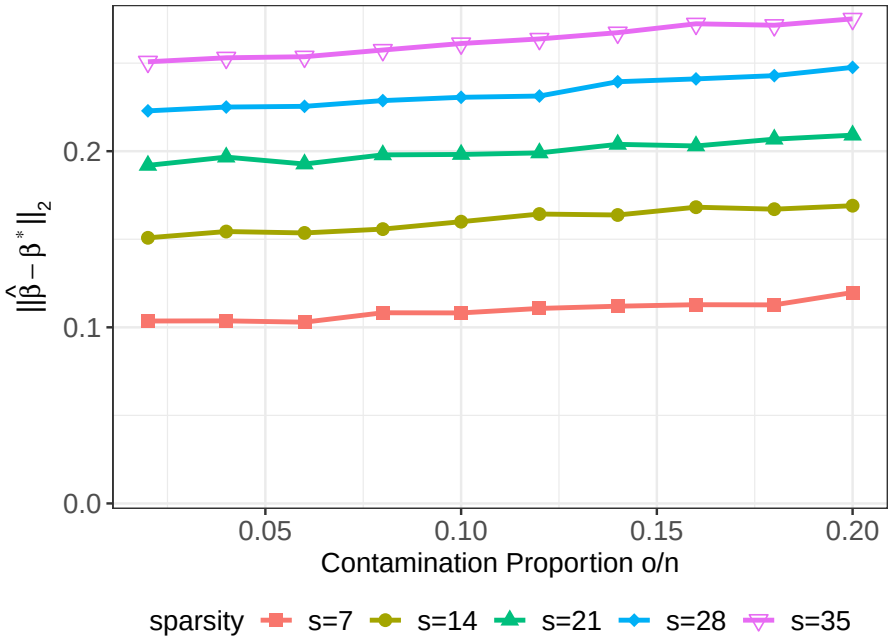

Figure 1: The ℓ_2 estimation error of the two-stage AC-IHT algorithm with varying s and o under 300 replications.

As illustrated in Figure 1, the ℓ_2 estimation error $\|\tilde{\beta} - \beta^*\|_2$ grows linearly with the contamination proportion o/n, yet it grows only on the order of $\sqrt{s}$ of the sparsity level s. This empirical behavior matches our theoretical guarantees (summarized in Table 1 in our manuscript): No matter in which signal cases, for fixed o, the ℓ_2 error bound increases linearly in $\sqrt{s}$; for fixed s, it increases linearly in o. In particular,

when $o/n \leq 1/5$, the ℓ_2 error bound in Theorem 3 can be rewritten as

$$\|\tilde{\beta} - \beta^*\|_2 \asymp \sigma\sqrt{\frac{s}{n-o}} \asymp \sigma\sqrt{\frac{s}{n}}\left(1 + \frac{o}{n}\right),$$

which explains the linear relationship with o.

S1.2 Asymptotic normality

We then consider the asymptotic normality of the estimator obtained from two-stage AC-IHT by constructing the z-score based on Corollary 2 with 300 replications. Specifically, we set the 5th and 6th elements of γ to 1 and all others to 0. We assess the asymptotic normality of AC-IHT, IHT-ℓ_1, and AC-SCAD by using histograms, Q-Q plots, and the R^2 values from the no-intercept linear fit of these Q-Q plots.

As illustrated in Figure 2, AC-IHT exhibits the best asymptotic normality performance: Its histogram closely aligns with the normal density curve, and the points in its Q-Q plot lie almost perfectly along the diagonal, with an $R^2 = 0.9891$. In contrast, IHT-ℓ_1 and AC-SCAD show noticeable deviations and lower R^2 values.

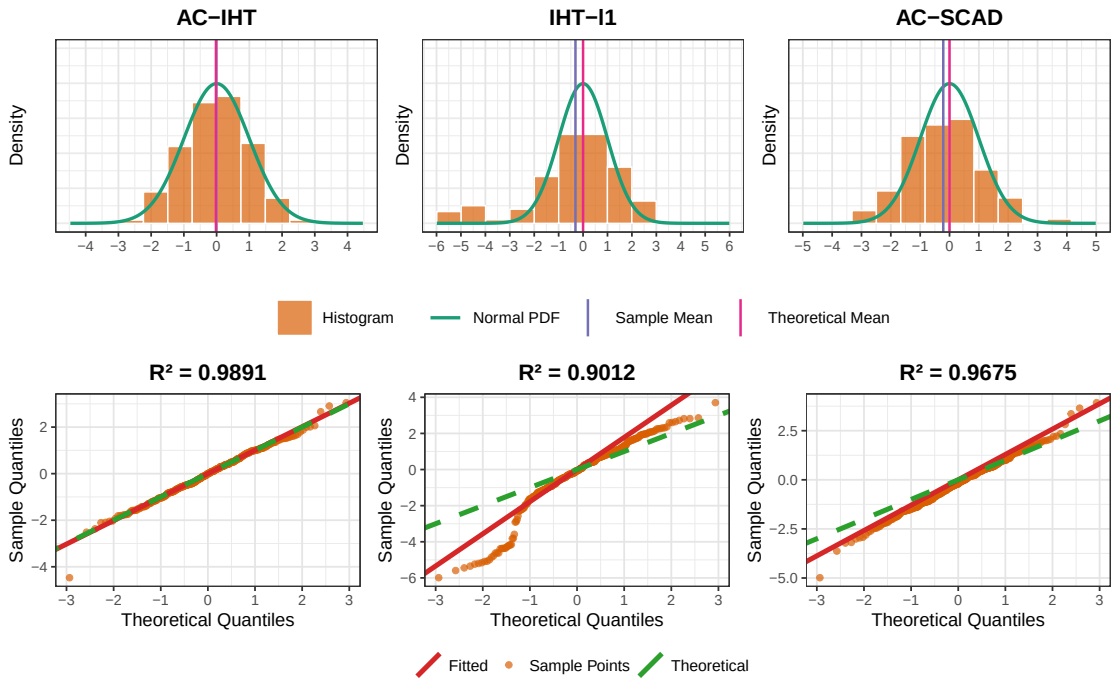

Figure 2: Comparison of asymptotic normality across AC-IHT, IHT-ℓ_1, and AC-SCAD methods.

S1.3 Heavy tailedness

Following the theoretical extension in the discussion section, we conduct supplementary experiments under heavy-tailed noise settings. Specifically, we generate i.i.d. white noise from Student's t-distributions with 2 and 3 degrees of freedom and assume no adversarial contamination, while keeping all other model configurations identical to those in Section 4. And we include the Oracle Huber estimator as an ideal reference, which is the Huber estimator with the known support set. Results over 100 replications are reported in Table 1, with standard errors shown in parentheses. The findings demonstrate that our AC-IHT method achieves favorable performance in both estimation accuracy and support recovery, which aligns with its minimax

near-optimal guarantee in Theorem 7.

Table 1: Comparison of estimation accuracy under heavy-tailed noise.

Method	$\|\beta-\beta^*\|_2$	$\|\beta-\beta^*\|_\infty$	$\|\beta-\beta^*\|_\Sigma$	**MCC**	**Sym_diff**
			t(df=2)		
AC-IHT	0.610 (0.027)	0.378 (0.014)	0.569 (0.024)	0.907 (0.009)	1.940 (0.193)
IHT-ℓ_1	0.757 (0.013)	0.494 (0.005)	0.713 (0.014)	0.895 (0.005)	1.970 (0.083)
Ada-Huber	0.683 (0.011)	0.354 (0.007)	0.770 (0.012)	0.563 (0.006)	21.440 (0.549)
Oracle Huber	0.269 (0.007)	0.162 (0.005)	0.254 (0.006)	1.000 (0.000)	0.000 (0.000)
			t(df=3)		
AC-IHT	0.435 (0.020)	0.301 (0.016)	0.403 (0.018)	0.959 (0.005)	0.850 (0.107)
IHT-ℓ_1	0.624 (0.023)	0.416 (0.015)	0.578 (0.021)	0.926 (0.005)	1.400 (0.096)
Ada-Huber	0.598 (0.010)	0.310 (0.006)	0.683 (0.011)	0.632 (0.005)	15.120 (0.409)
Oracle Huber	0.252 (0.006)	0.153 (0.004)	0.235 (0.005)	1.000 (0.000)	0.000 (0.000)

S1.4 Dynamic of convergence

In this subsection, we examine the iteration-wise convergence behavior of AC-IHT under varying sample sizes and contamination levels. Specifically, we consider sample sizes $n \in \{300, 500, 700\}$ and numbers of contaminated observations $o \in \{10, 30\}$, while keeping all other simulation settings the same as in the beginning of Section 4. Estimation accuracy is assessed using the ℓ_2 error (L_2) and the ℓ_∞ error ($L_{\max}$), and support recovery accuracy is evaluated by the Matthews Correlation Coefficient (MCC). The corresponding iteration trajectories are displayed in Figure 3.

Figure 3 shows a consistent convergence pattern across all configurations. The errors L_2 and $L_{\max}$ decrease during the initial iterations and then level off after about 20 iterations, whereas MCC rises toward one after about 15 iterations and remains stable thereafter, indicating reliable support recovery. Moreover, larger sample sizes

lead to improved estimation and support recovery performance, as evidenced by smaller $L_2/L_{\max}$ and higher MCC. When the contamination level increases from $o = 10$ to $o = 30$, convergence becomes slightly slower, and the terminal errors are mildly larger, particularly for smaller n, which aligns with our theoretical results.

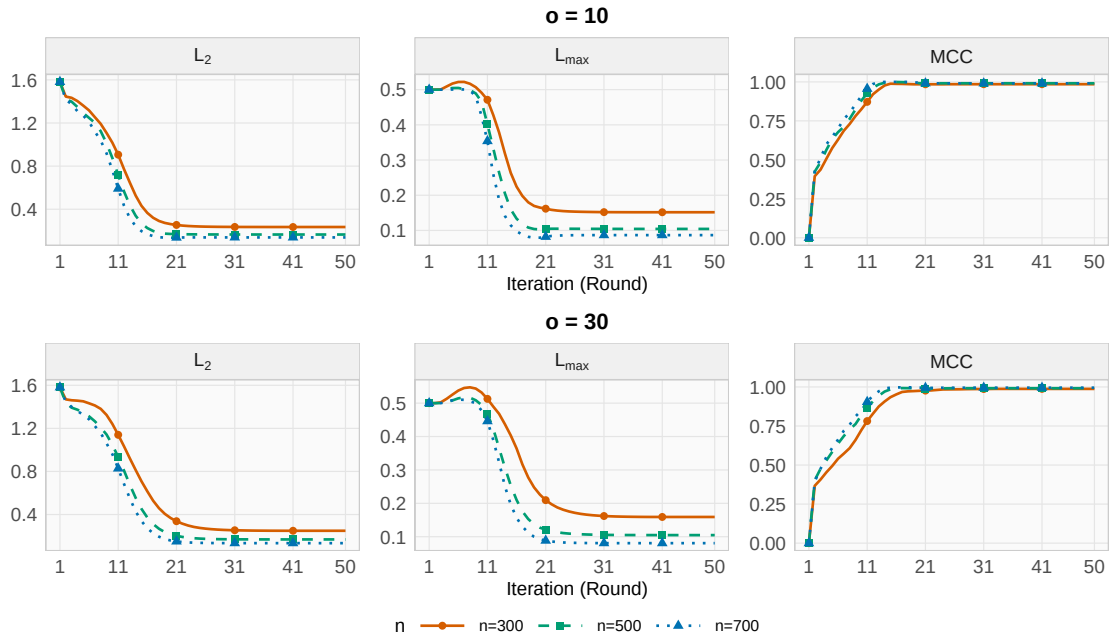

Figure 3: Convergence dynamics of AC-IHT over iterations for different values of n and o. Each (o, n) setting undergoes 100 repeated simulations.

S1.5 Relationship between signal strength and sparsity.

Assumption 4 indicates that the required signal strength (for θ^*) increases as o decreases. The following simulation validates this phenomenon. Following the simulation settings in Section 4, we fix the non-zero entries of θ^* at 0.2, and examine the support recovery of θ^* as o increases under standard Gaussian noise. The simulation result in Figure 4 (with 100 replications) shows that as $o(= \|\theta^*\|_0)$

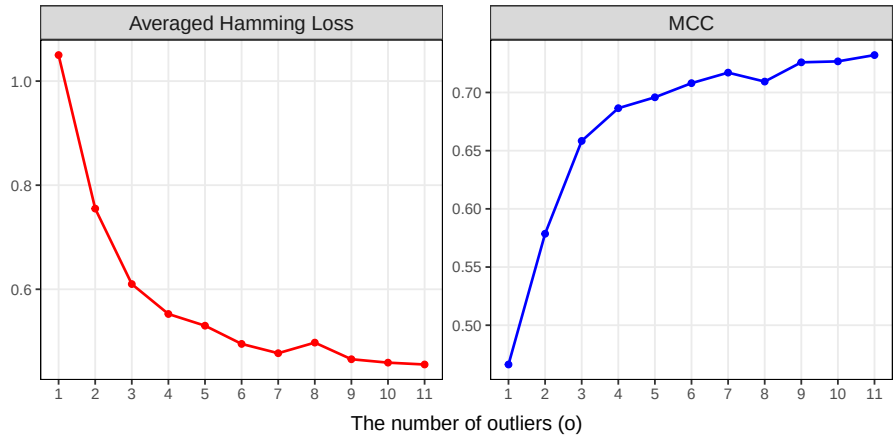

Figure 4: Support recovery performance with increasing o.

increases, the MCC (Matthews Correlation Coefficient) rises and the averaged Hamming loss $\left(\frac{\left|\text{supp}(\theta^*)\triangle\text{supp}(\hat{\theta})\right|}{o}\right)$ decreases. This indicates that a larger o makes the "theta-min" condition easier to satisfy, yielding progressively more accurate outlier identification. And this result aligns with the relationship between o and the signal strength in Assumption 4.

S2 Proof of Proposition 1

Throughout this proof, $C_1 > 0$ is treated as a fixed constant, and Assumption 2 is used in the concrete form stated in Proposition 1. We first prove the restricted isometry of X. For a fixed set $S \subset [p]$ with $|S| = C_1 s$, by Remark 5.40 in Vershynin (2010), we obtain

$$\mathbf{P}\left(\left\|\frac{1}{n}X_{\cdot S}^{\top}X_{\cdot S} - \Sigma_{SS}\right\|_2 > \max(\iota, \iota^2)\right) \leq 2e^{-u}, \tag{S2.1}$$

where $\iota = \left(C \vee \frac{1}{\sqrt{c}}\right)\left(\sqrt{\frac{C_1 s}{n}} + \sqrt{\frac{u}{n}}\right)$, and C, c are constants depending only on $\|\Sigma\|_2$. Let $C_\Sigma := C^2 \vee \frac{1}{c} \vee 1$. By taking $u = 3C_1 s \log p$ and using $n \geq 30M^2 C_1 C_\Sigma \max(s \log p, o \log n)$, we have $\iota \vee \iota^2 = \iota \leq 1/(2M)$, which yields that

$$\begin{aligned}
&\mathbf{P}\left(\sup_{S\subset[p]:\ |S|\leq C_1 s}\left\|\frac{1}{n}X_{\cdot S}^\top X_{\cdot S} - \Sigma_{SS}\right\|_2 > \frac{1}{2M}\right)\\
\leq &\sum_{S\subset[p]:\ |S|=C_1 s}\mathbf{P}\left(\left\|\frac{1}{n}X_{\cdot S}^\top X_{\cdot S} - \Sigma_{SS}\right\|_2 > \max(\iota, \iota^2)\right)\\
\leq &2\binom{p}{C_1 s}\exp\left(-3C_1 s \log p\right)\\
\leq &2\exp\left(-2C_1 s \log p\right).
\end{aligned} \tag{S2.2}$$

By Weyl's inequality, with probability at least $1 - 2\exp(-2C_1 s \log p)$ we have that,

$$\max_{S\subset[p]:\ |S|\leq C_1 s}\ \max_{1\leq k\leq |S|}\left|\Lambda_k\left(\frac{1}{n}X_{\cdot S}^\top X_{\cdot S}\right) - \Lambda_k(\Sigma_{SS})\right| \leq \frac{1}{2M}. \tag{S2.3}$$

Consequently, the restricted isometry property follows immediately from this uniform eigenvalue bound and the definition of eigenvalue.

We then prove the restricted incoherence of X. Similar to (S2.1), for fixed sets $S \subset [p]$ with $|S| = C_1 s$ and $O \subset [n]$ with $|O| = C_1 o$, we have

$$\mathbf{P}\left(\left\|\frac{1}{C_1 o}X_{O,S}^\top X_{O,S} - \Sigma_{SS}\right\|_2 > \max(\tau, \tau^2)\right) \leq 2e^{-v},$$

where $\tau = \left(C \vee \frac{1}{\sqrt{c}}\right) \cdot \left(\sqrt{\frac{s}{o}} + \sqrt{\frac{v}{C_1 o}}\right)$. By setting $v = 3C_1 s \log p + 3C_1 o \log n$ and

applying a union bound, we obtain

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{O\subset[n]:\ |O|\le C_1 o}\ \sup_{S\subset[p]:\ |S|\le C_1 s}\left\|\frac{1}{C_1 o}X_{O,S}^\top X_{O,S}-\Sigma_{SS}\right\|_2>\max(\tau,\tau^2)\right\}\\
\le&\binom{p}{C_1 s}\binom{n}{C_1 o}2e^{-v}\\
\le&2e^{-2C_1 s\log p}.
\end{aligned}$$

Consequently, for any $S\subset[p]$ with $|S|\le C_1 s$ and $O\subset[n]$ with $|O|\le C_1 o$, it holds that

$$\begin{aligned}
\|X_{O,S}\|_2=&\sqrt{\left\|X_{O,S}^\top X_{O,S}\right\|_2}\\
\le&\sqrt{C_1 o\|\Sigma_{SS}\|_2+C_1 o(\tau+\tau^2)}\\
\le&\sqrt{C_1 oM+C_1C_\Sigma\left(\sqrt{so}+\sqrt{\frac{vo}{C_1}}+2s+\frac{2v}{C_1}\right)}\\
\le&\sqrt{C_1 oM+10C_1C_\Sigma\left(s\log p+o\log n\right)}\\
\le&\sqrt{C_1}C_M\cdot\sqrt{s\log p+o\log n}
\end{aligned}\tag{S2.4}$$

with probability at least $1-2\exp(-2C_1 s\log p)$. Here $C_M:=\sqrt{M+10C_\Sigma}>1$ is a constant depending only on M. Therefore, we prove the restricted incoherence property and complete the proof of Proposition 1.

Remark 1 (Incoherence condition) *The restricted incoherence condition* (S2.4) *plays a pivotal role in our theoretical analysis, appearing in key steps such as* (S3.14), (S3.21), (S4.7), *and* (S6.5). *It controls how adversarial contamination distorts the estimation of β^*, and enables us to achieve a sharper estimation accuracy than joint estimation. Existing incoherence conditions in the literature, such as Definition 1.(ii)*

in Dalalyan and Thompson (2019), and equations (2.11)-(2.12) in Minsker et al. (2024), are often tied to the specific penalty forms of Lasso or Slope, and are relatively intricate. In contrast, by leveraging the ℓ_0 structure, we introduce a more concise and intuitive incoherence condition, and prove that it holds with high probability under sub-Gaussian designs.

S3 Proof of Theorem 1

We introduce some definitions that will be used in the following proofs. Define

$$\Phi := \frac{\eta}{n} X^\top X - I_p \in \mathbb{R}^{p\times p}, \qquad \Xi := \frac{1}{n} X^\top \xi \in \mathbb{R}^p.$$

For ease of display, we assume $\sigma = 1$ in the main proof.

Three parts constitute the proof of Theorem 1. We first introduce some useful preliminaries, and then prove the sparsity and error bounds by using mathematical induction. Finally, we verify the validity of our proposed threshold sequence.

S3.1 Preliminary

Define the event

$$\mathcal{E} := \left\{ \|\Xi\|_\infty < 4\sqrt{\frac{M\log p}{n}},\ \text{and}\ \left\|\frac{1}{\sqrt{n}}\xi\right\|_\infty < 3\sqrt{\frac{\log n}{n}} \right\}, \tag{S3.5}$$

and

$$\mathcal{E}_X := \left\{ \begin{array}{c} \frac{1}{2M} \le \Lambda_k\left(\frac{1}{n} X_{\cdot S}^\top X_{\cdot S}\right) \le 2M, \text{ for every } S \subset [p] : \ |S| \le (2B+1)s \text{ and } 1 \le k \le |S|, \\ \sup_{S\subset[p]:\ |S|\le(B+1)s} \ \sup_{O\subset[n]:\ |O|\le(B+1)o} \|X_{O,S}\|_2 \le (\sqrt{B}+1) f\sqrt{n} \end{array} \right\},$$

where

$$B := \left(\frac{\kappa+\delta}{\kappa-\delta}\right)^2 > 1, \quad \delta := \frac{4M^2}{4M^2+1} \in (0,1), \quad \kappa \in (\delta, 1), \quad f := C_M \sqrt{\frac{s\log p + o\log n}{n}}. \tag{S3.6}$$

Throughout the proof of Theorem 1, Assumption 2 is used in the concrete form

$$n \ge C_{\mathrm{Th1}}\{s\log p + o\log n\},$$

where C_{Th1} is a fixed constant depending only on M, η, κ, chosen large enough so that

$$C_{\mathrm{Th1}} \ge \max\left\{30M^2(2B+1)C_\Sigma,\ \frac{32\eta^2\kappa^2 C_M^2}{(\kappa-\delta)^4},\ 4MC_M^2\right\}.$$

Therefore, with $n \ge 30M^2(2B+1)C_\Sigma\{s\log p + o\log n\}$, we confirm that the restricted isometry part of $\mathcal{E}_X$ follows from Proposition 1 with $C_1 = 2B+1$, and the restricted incoherence part follows from Proposition 1 with $C_1 = B+1$. Since $B = \left(\frac{\kappa+\delta}{\kappa-\delta}\right)^2$ is fixed once M and $\kappa \in (\delta, 1)$ are fixed, these choices of C_1 are admissible under Proposition 1, hence we have $\mathbf{P}(\mathcal{E}_X) \ge 1 - 4p^{-2(1+B)s}$.

Similarly, with $n \ge 30M^2 C_\Sigma \max(s\log p, o\log n)$, by Lemma 1 we have $\mathbf{P}(\mathcal{E}) \ge$

$1 - 2p^{-2s} - 2p^{-3} - 2n^{-3}$. And we conclude that

$$\mathbf{P}(\mathcal{E} \cap \mathcal{E}_X) \geq 1 - O(p^{-2} + n^{-3}).$$

The proof of Theorem 1 is then based on the event $\mathcal{E} \cap \mathcal{E}_X$.

In the iteration algorithm, we take the learning rate $\eta \in \left[\frac{2M}{4M^2+1}, \frac{4M}{4M^2+1}\right]$ and the decay rate $\kappa \in (\delta, 1)$. And we use the thresholds

$$\begin{aligned}
\lambda_{\beta,\infty} &= \frac{16\eta\kappa}{\kappa-\delta}\sqrt{\frac{M\log p}{n}} + \frac{48\eta^2\kappa^2 C_M}{(\kappa-\delta)^3}\sqrt{\frac{s\log p + o\log n}{ns}}\sqrt{\frac{o\log n}{n}} \\
&\asymp \sqrt{\frac{\log p}{n}} + \frac{o\log n}{n\sqrt{s}}, \\
\lambda_{\theta,\infty} &= \frac{12\eta\kappa}{\kappa-\delta}\sqrt{\frac{\log n}{n}} + \frac{64\eta^2\kappa^2 C_M\sqrt{M}}{(\kappa-\delta)^3}\sqrt{\frac{s\log p + o\log n}{no}}\sqrt{\frac{s\log p}{n}} \\
&\asymp \sqrt{\frac{\log n}{n}} + \frac{s\log p}{n\sqrt{o}}.
\end{aligned} \tag{S3.7}$$

For the given thresholds $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$ in (S3.7), we can always choose sufficiently large initial thresholds $\lambda_{\beta,0}$ ($> \lambda_{\beta,\infty}$) and $\lambda_{\theta,0}$ ($> \lambda_{\theta,\infty}$) satisfying

$$\sqrt{s}\lambda_{\beta,0} > \|\beta^*\|_2, \quad \sqrt{o}\lambda_{\theta,0} > \|\theta^*\|_2, \quad \frac{\lambda_{\beta,0}}{\lambda_{\theta,0}} = \frac{\lambda_{\beta,\infty}}{\lambda_{\theta,\infty}}. \tag{S3.8}$$

The detailed threshold initialization is provided in Section S3.4, where we utilize the conditions of Theorem 1 and the high-probability restricted incoherence property from Proposition 1 to establish a feasible range for the initial thresholds.

Additionally, by the decomposition, we define

$$\begin{aligned} H_\beta^{t+1} :=& \beta^t + \frac{\eta}{n} X^\top (Y - X\beta^t - \sqrt{n}\theta^t) \\ =& \beta^* + \Phi(\beta^* - \beta^t) + \frac{\eta}{\sqrt{n}} X^\top (\theta^* - \theta^t) + \eta \Xi \in \mathbb{R}^p, \\ H_\theta^{t+1} :=& \theta^t + \frac{\eta}{\sqrt{n}} (Y - X\beta^t - \sqrt{n}\theta^t) \\ =& \theta^* + (\eta - 1)(\theta^* - \theta^t) + \frac{\eta}{\sqrt{n}} X(\beta^* - \beta^t) + \frac{\eta}{\sqrt{n}} \xi \in \mathbb{R}^n. \end{aligned} \tag{S3.9}$$

S3.2 Mathematical induction

Recall $S^* = supp(\beta^*)$ and $O^* = supp(\theta^*)$. We aim to use mathematical induction to prove the following

$$\|\beta^t_{(S^*)^c}\|_0 < Bs, \quad \|\theta^t_{(O^*)^c}\|_0 < Bo, \tag{S3.10}$$

$$\|\beta^t - \beta^*\|_2 \leq (\sqrt{B} + 1) \cdot \sqrt{s}\lambda_{\beta,t}, \quad \|\theta^t - \theta^*\|_2 \leq (\sqrt{B} + 1) \cdot \sqrt{o}\lambda_{\theta,t}. \tag{S3.11}$$

hold for all $t \geq 0$.

First, by (S3.8), we guarantee that both (S3.10) and (S3.11) hold at $t = 0$, with $\beta^0 = \mathbf{0}_p$ and $\theta^0 = \mathbf{0}_n$. By using mathematical induction, assume that (S3.10) and (S3.11) hold at iteration t for some $t \geq 0$, and then we aim to prove that they remain valid at iteration $t + 1$.

Proof of sparsity In the $(t + 1)$-th iteration, we first prove (S3.10) reasoning by the absurd. Assume that $\|\beta^{t+1}_{(S^*)^c}\|_0 \geq Bs$ or $\|\theta^{t+1}_{(O^*)^c}\|_0 \geq Bo$ holds at first. It then follows that there exist subsets $\tilde{S} \subset (S^*)^c$ or $\tilde{O} \subset (O^*)^c$ such that $\|\tilde{S}\|_0 = Bs$ or

$\|\tilde{O}\|_0 = Bo$, satisfying

$$Bs\lambda_{\beta,t+1}^2 < \sum_{i\in\tilde{S}} (H_{\beta,i}^{t+1})^2 \mathbf{1}\{|H_{\beta,i}^{t+1}| \geq \lambda_{\beta,t+1}\} \tag{S3.12}$$

or

$$Bo\lambda_{\theta,t+1}^2 < \sum_{j\in\tilde{O}} (H_{\theta,j}^{t+1})^2 \mathbf{1}\{|H_{\theta,j}^{t+1}| \geq \lambda_{\theta,t+1}\}. \tag{S3.13}$$

Since $\beta_{\tilde{S}}^* = \mathbf{0}_{Bs}$ and $\theta_{\tilde{O}}^* = \mathbf{0}_{Bo}$, under event $\mathcal{E}$, if (S3.12) holds, by decomposition (S3.9) we have

$$\begin{aligned}
\sqrt{Bs}\lambda_{\beta,t+1} &\leq \sqrt{\sum_{i\in\tilde{S}} \langle \Phi_{\cdot i}, \beta^* - \beta^t\rangle^2} + \sqrt{\sum_{i\in\tilde{S}} \frac{\eta^2}{n} \langle X_{\cdot i}, \theta^* - \theta^t\rangle^2} + \eta\sqrt{Bs}\|\Xi\|_\infty \\
&\overset{(i)}{\leq} \delta\|\beta^* - \beta^t\|_2 + \eta(\sqrt{B}+1) f \cdot \|\theta^* - \theta^t\|_2 + \eta\sqrt{Bs}\|\Xi\|_\infty \\
&\overset{(ii)}{<} \delta(\sqrt{B}+1)\cdot\sqrt{s}\lambda_{\beta,t} + \eta(\sqrt{B}+1)^2 f\cdot\sqrt{o}\lambda_{\theta,t} + 4\eta(\sqrt{B}+1)\sqrt{s}\sqrt{\frac{M\log p}{n}} \\
&\overset{(iii)}{\leq} \frac{2\delta}{\kappa-\delta}\cdot\sqrt{s}\lambda_{\beta,t+1} + \frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{o}\lambda_{\theta,t+1} + \frac{8\eta\kappa}{\kappa-\delta}\sqrt{s}\sqrt{\frac{M\log p}{n}},
\end{aligned} \tag{S3.14}$$

where:

- Inequality (i) follows from the fact

$$\begin{aligned}
&\sum_{i\in\tilde{S}}\langle\Phi_{\cdot i},\beta^*-\beta^t\rangle^2 \le \sum_{i\in S'}\langle\Phi_{\cdot i},\beta^*-\beta^t\rangle^2 \qquad (\text{define } S^{'}=\tilde{S}\cup\operatorname{supp}(\beta^*-\beta^t))\\
&= \ (\beta^*-\beta^t)_{S'}^\top\Phi_{S'S'}^\top\Phi_{S'S'}(\beta^*-\beta^t)_{S'} \ \le \ \|\Phi_{S'S'}\|_2^2\|\beta^*-\beta^t\|_2^2\\
&= \left\{\left|\Lambda_{\max}\left(\frac{\eta(X^\top X)_{S'S'}}{n}\right)-1\right|\bigvee\left|1-\Lambda_{\min}\left(\frac{\eta(X^\top X)_{S'S'}}{n}\right)\right|\right\}^2\|\beta^*-\beta^t\|_2^2\\
&\le (1-\eta/(2M))^2\|\beta^*-\beta^t\|_2^2 \quad (\text{By } \mathcal{E}_X \text{ and the range of } \eta)\\
&\le \delta^2\|\beta^*-\beta^t\|_2^2, \quad (\text{Equation (S3.6)})
\end{aligned} \tag{S3.15}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{i\in\tilde{S}}\langle X_{\cdot i},\theta^*-\theta^t\rangle^2 &= \frac{1}{n}(\theta^*-\theta^t)_{O'}^\top X_{O'\tilde{S}}X_{O'\tilde{S}}^\top(\theta^*-\theta^t)_{O'} \qquad (\text{define } O^{'}=\operatorname{supp}(\theta^*-\theta^t))\\
&\le \frac{1}{n}\|X_{O'\tilde{S}}\|_2^2\cdot\|\theta^*-\theta^t\|_2^2\\
&\le (\sqrt{B}+1)^2f^2\cdot\|\theta^*-\theta^t\|_2^2. \quad (\text{By } \mathcal{E}_X)
\end{aligned} \tag{S3.16}$$

- Inequality (ii) follows from the event $\mathcal{E}$ and the assumption (S3.11) at the t-th iteration.

- Inequality (iii) follows the relationship:

$$\sqrt{B}+1=\frac{2\kappa}{\kappa-\delta}, \quad \lambda_{\beta,t+1}=\max(\kappa\lambda_{\beta,t},\ \lambda_{\beta,\infty}), \quad \lambda_{\theta,t+1}=\max(\kappa\lambda_{\theta,t},\ \lambda_{\theta,\infty}). \tag{S3.17}$$

Similarly, if (S3.13) holds, we have

$$\begin{aligned}\sqrt{Bo}\lambda_{\theta,t+1} &\le |\eta-1|\cdot\|\theta^*-\theta^t\|_2+\sqrt{\sum_{j\in\tilde{O}}\frac{\eta^2}{n}\langle X_{j\cdot}^\top,\beta^*-\beta^t\rangle^2}+\eta\sqrt{\sum_{j\in\tilde{O}}\frac{1}{n}\xi_j^2}\\ &\le \delta\|\theta^*-\theta^t\|_2+\eta(\sqrt{B}+1)f\|\beta^*-\beta^t\|_2+3\eta\sqrt{Bo}\sqrt{\frac{\log n}{n}}\\ &< \frac{2\delta}{\kappa-\delta}\cdot\sqrt{o}\lambda_{\theta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{s}\lambda_{\beta,t+1}+\frac{6\eta\kappa}{\kappa-\delta}\sqrt{o}\sqrt{\frac{\log n}{n}},\end{aligned}\tag{S3.18}$$

where:

- The second inequality follows from the fact

$$\begin{aligned}\frac{1}{n}\sum_{j\in\tilde{O}}\langle X_{j\cdot}^\top,\beta^*-\beta^t\rangle^2 &\le \frac{1}{n}(\beta^*-\beta^t)_{S''}^\top X_{\tilde{O}S''}^\top X_{\tilde{O}S''}(\beta^*-\beta^t)_{S''} \qquad (\text{define } S''=\operatorname{supp}(\beta^*-\beta^t))\\ &\le \frac{1}{n}\|X_{\tilde{O}S''}\|_2^2\cdot\|\beta^*-\beta^t\|_2^2\\ &\le (\sqrt{B}+1)^2f^2\cdot\|\beta^*-\beta^t\|_2^2 \quad (\text{By } \mathcal{E}_X),\end{aligned}\tag{S3.19}$$

  and the fact $|\eta-1|<\delta$ in the case $\eta\in\left[\frac{2M}{4M^2+1},\frac{4M}{4M^2+1}\right]$.

- The derivation of the last inequality is analogous to that of (ii) and (iii) in (S3.14).

Combining (S3.14) and (S3.18), to prove $\|\beta^{t+1}_{(S^*)^c}\|_0<Bs$ and $\|\theta^{t+1}_{(O^*)^c}\|_0<Bo$ by contradiction, it suffices to show that:

$$\begin{cases}\frac{2\delta}{\kappa-\delta}\cdot\sqrt{s}\lambda_{\beta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{o}\lambda_{\theta,t+1}+\frac{8\eta\kappa}{\kappa-\delta}\sqrt{s}\sqrt{\frac{M\log p}{n}}\le\sqrt{Bs}\lambda_{\beta,t+1},\\ \frac{2\delta}{\kappa-\delta}\cdot\sqrt{o}\lambda_{\theta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{s}\lambda_{\beta,t+1}+\frac{6\eta\kappa}{\kappa-\delta}\sqrt{o}\sqrt{\frac{\log n}{n}}\le\sqrt{Bo}\lambda_{\theta,t+1}.\end{cases}\tag{S3.20}$$

We first assume that there exist suitable sequence $\{(\lambda_{\beta,t},\ \lambda_{\theta,t})\}_{t\geq 0}$ satisfying this system (S3.20), and we will construct its explicit form in Section S3.3.

Proof of error bounds Once the sparsity results are established in the $(t+1)$-th iteration, we turn to derive upper bounds on estimation errors. We establish that

$$
\begin{aligned}
&\|\beta^{t+1}-\beta^*\|_2 \\
&\leq \Bigg\{ \sum_{i\in S^*} \left(-H_{\beta,i}^{t+1}\mathbf{1}(|H_{\beta,i}^{t+1}|<\lambda_{\beta,t+1}) + \langle \Phi_{\cdot i}, \beta^*-\beta^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^*-\theta^t\rangle + \eta\Xi_i \right)^2 \\
&\qquad + \sum_{i\in S^{t+1}\setminus S^*} \left(\langle \Phi_{\cdot i}, \beta^*-\beta^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^*-\theta^t\rangle + \eta\Xi_i \right)^2 \Bigg\}^{1/2} \\
&\leq \sqrt{\sum_{i\in S^*} \left(H_{\beta,i}^{t+1}\right)^2 \mathbf{1}(|H_{\beta,i}^{t+1}|<\lambda_{\beta,t+1})} + \sqrt{\sum_{i\in S^{t+1}\cup S^*} \langle \Phi_{\cdot i}, \beta^*-\beta^t\rangle^2} \\
&\qquad + \eta\sqrt{\sum_{i\in S^{t+1}\cup S^*} \frac{1}{n}\langle X_{\cdot i}, \theta^*-\theta^t\rangle^2} + \eta\sqrt{\sum_{i\in S^{t+1}\cup S^*} \Xi_i^2} \\
&\overset{(i)}{\leq} \sqrt{s}\lambda_{\beta,t+1} + \delta\|\beta^t-\beta^*\|_2 + \eta(\sqrt{B}+1)f\cdot\|\theta^*-\theta^t\|_2 + 4\eta(\sqrt{B}+1)\sqrt{s}\sqrt{\frac{M\log p}{n}} \\
&\overset{(ii)}{\leq} \sqrt{s}\lambda_{\beta,t+1} + \frac{2\delta}{\kappa-\delta}\cdot\sqrt{s}\lambda_{\beta,t+1} + \frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{o}\lambda_{\theta,t+1} + \frac{8\eta\kappa}{\kappa-\delta}\sqrt{s}\sqrt{\frac{M\log p}{n}},
\end{aligned}
\tag{S3.21}
$$

where recall $S^{t+1}=\mathrm{supp}(\beta^{t+1}),\ S^*=\mathrm{supp}(\beta^*)$. Here, inequality (i) follows from (S3.15) and (S3.16) again, and inequality (ii) follows analogously to (ii) and (iii) in

(S3.14). Similar to (S3.21), we derive that

$$\begin{aligned}\|\theta^{t+1}-\theta^*\|_2 &\le \sqrt{o}\lambda_{\theta,t+1}+\delta\|\theta^t-\theta^*\|_2+\eta(\sqrt{B}+1)f\|\beta^t-\beta^*\|_2+3\eta\sqrt{(B+1)o}\sqrt{\frac{\log n}{n}}\\ &< \sqrt{o}\lambda_{\theta,t+1}+\frac{2\delta}{\kappa-\delta}\cdot\sqrt{o}\lambda_{\theta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{s}\lambda_{\beta,t+1}+\frac{6\eta\kappa}{\kappa-\delta}\sqrt{o}\sqrt{\frac{\log n}{n}}.\end{aligned} \tag{S3.22}$$

Therefore, to ensure that the ℓ_2 error bounds after the $(t+1)$-th iteration satisfies (S3.11), it suffices for the sequence $\{(\lambda_{\beta,t},\lambda_{\theta,t})\}_{t\ge 0}$ to satisfy

$$\begin{cases}\sqrt{s}\lambda_{\beta,t+1}+\frac{2\delta}{\kappa-\delta}\cdot\sqrt{s}\lambda_{\beta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{o}\lambda_{\theta,t+1}+\frac{8\eta\kappa}{\kappa-\delta}\sqrt{s}\sqrt{\frac{M\log p}{n}}\le(\sqrt{B}+1)\sqrt{s}\lambda_{\beta,t+1},\\ \sqrt{o}\lambda_{\theta,t+1}+\frac{2\delta}{\kappa-\delta}\cdot\sqrt{o}\lambda_{\theta,t+1}+\frac{4\eta\kappa}{(\kappa-\delta)^2}\cdot f\sqrt{s}\lambda_{\beta,t+1}+\frac{6\eta\kappa}{\kappa-\delta}\sqrt{o}\sqrt{\frac{\log n}{n}}\le(\sqrt{B}+1)\sqrt{o}\lambda_{\theta,t+1},\end{cases} \tag{S3.23}$$

which is equivalent to the system (S3.20). Consequently, we next focus on constructing a suitable set of solutions for $\{(\lambda_{\beta,t},\lambda_{\theta,t})\}_{t\ge 0}$.

S3.3 Appropriate threshold sequence

Define

$$W_1:=\frac{4\eta\kappa}{(\kappa-\delta)^2},\quad W_2:=\frac{8\eta\kappa}{\kappa-\delta}\sqrt{\frac{Ms\log p}{n}},\quad W_3:=\frac{6\eta\kappa}{\kappa-\delta}\cdot\sqrt{\frac{o\log n}{n}},$$

and note that $\sqrt{B}=\frac{\kappa+\delta}{\kappa-\delta}$. Then the system (S3.20) reduces to the inequalities

$$\begin{cases}W_1\cdot f\sqrt{o}\lambda_{\theta,t+1}+W_2\le\sqrt{s}\lambda_{\beta,t+1},\\ W_1\cdot f\sqrt{s}\lambda_{\beta,t+1}+W_3\le\sqrt{o}\lambda_{\theta,t+1}.\end{cases} \tag{S3.24}$$

The feasible region for $(\sqrt{s}\lambda_{\beta,t+1}, \sqrt{o}\lambda_{\theta,t+1})$ is illustrated as the shaded area in Figure 5. Given the update rule

$$\lambda_{\beta,t+1} = (\kappa \times \lambda_{\beta,t}) \vee \lambda_{\beta,\infty},$$

$$\lambda_{\theta,t+1} = (\kappa \times \lambda_{\theta,t}) \vee \lambda_{\theta,\infty},$$

maintaining the ratio $\frac{\lambda_{\beta,0}}{\lambda_{\theta,0}} = \frac{\lambda_{\beta,\infty}}{\lambda_{\theta,\infty}}$ ensures that the iterate trajectory $\{(\lambda_{\beta,t}, \lambda_{\theta,t})\}_{t\geq 0}$ follows a linear path toward the origin, such as the solid purple or dashed red rays shown in Figure 5. Specifically, in the case $1 - W_1^2 f^2 > 1/2$, i.e.,

$$n > 32\frac{\eta^2\kappa^2 C_M^2}{(\kappa-\delta)^4}(s\log p + o\log n),$$

the target purple endpoint $(\sqrt{s}\lambda_{\beta,\infty} = 2W_2 + 2fW_1W_3,\ \sqrt{o}\lambda_{\theta,\infty} = 2W_3 + 2fW_1W_2)$ lies within the feasible region. Consequently, the purple ray (passing through this endpoint and maintaining $\frac{\lambda_{\beta,t}}{\lambda_{\theta,t}} = \frac{\lambda_{\beta,\infty}}{\lambda_{\theta,\infty}}$) constitutes a valid threshold update trajectory.

Therefore, by using the thresholds in (S3.7), we prove that

$$\|\beta_{S^c}^{t+1}\|_0 < Bs, \quad \|\theta_{O^c}^{t+1}\|_0 < Bo,$$

$$\|\beta^{t+1} - \beta^*\|_2 \leq (\sqrt{B}+1)\cdot\sqrt{s}\lambda_{\beta,t+1}, \quad \|\theta^{t+1} - \theta^*\|_2 \leq (\sqrt{B}+1)\cdot\sqrt{o}\lambda_{\theta,t+1}$$

hold simultaneously in the $(t+1)$-th iteration, which completes the proof of Theorem 1.

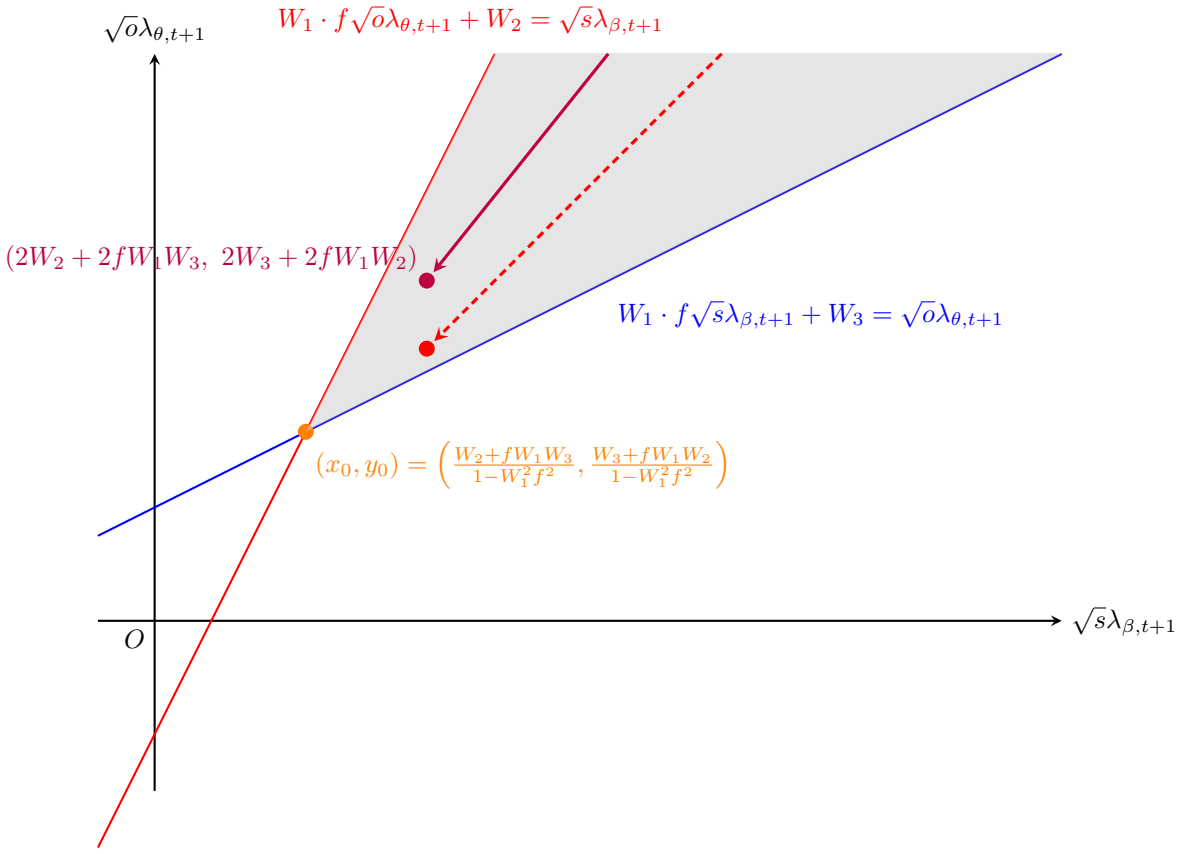

Figure 5: Feasible region (shaded) and two exemplary threshold update trajectories (solid purple and dashed red).

S3.4 Threshold initialization

Here we provide some practical guidance for selecting the initial tuning parameters $\lambda_{\beta,0}$ and $\lambda_{\theta,0}$. With a probability greater than $1 - O(p^{-2} + n^{-3})$, we have the decomposition

$$\begin{aligned}
\sqrt{s}\left\|\frac{1}{n}X^\top Y\right\|_\infty \geq& \left\|\left(\frac{1}{n}X^\top Y\right)_{S^*}\right\|_2 \\
\geq& \left\|\left(\frac{1}{n}X^\top X\right)_{S^*,S^*}\beta^*_{S^*}\right\|_2 - \left\|\left(\frac{1}{\sqrt{n}}X^\top\theta^*\right)_{S^*}\right\|_2 - \|(\Xi)_{S^*}\|_2 \\
\geq& \frac{1}{2M}\|\beta^*\|_2 - f\|\theta^*\|_2 - 4\sqrt{M}\sqrt{\frac{s\log p}{n}},
\end{aligned} \tag{S3.25}$$

and

$$\begin{aligned}
\sqrt{o}\left\|\frac{1}{\sqrt{n}}Y\right\|_\infty \geq& \left\|\left(\frac{1}{\sqrt{n}}Y\right)_{O^*}\right\|_2 \\
\geq& \|\theta^*\|_2 - \left\|\left(\frac{1}{\sqrt{n}}X\beta^*\right)_{O^*}\right\|_2 - \left\|\frac{1}{\sqrt{n}}\xi_{O^*}\right\|_2 \\
\geq& \|\theta^*\|_2 - f\|\beta^*\|_2 - 3\sqrt{\frac{o\log n}{n}}.
\end{aligned} \tag{S3.26}$$

For both inequalities, the final step relies on the restricted incoherence property given in Proposition 1, with $C_1 = 1$. And the sample-size assumption on which Proposition 1 depends is fulfilled under the sample condition presented in Section S3.1.

By combining these inequalities, we obtain

$$\sqrt{s}\left\|\frac{1}{n}X^\top Y\right\|_\infty + 4\sqrt{M}\sqrt{\frac{s\log p}{n}} + f\sqrt{o}\left\|\frac{1}{\sqrt{n}}Y\right\|_\infty + 3f\sqrt{\frac{o\log n}{n}} \geq \left(\frac{1}{2M} - f^2\right)\|\beta^*\|_2 .$$

Then, under the assumption $n \geq 4MC_M^2(s\log p + o\log n)$, it follows that

$$\frac{\|\beta^*\|_2}{\sqrt{s}} \leq 4M\left(4 + \left\|\frac{1}{n}X^\top Y\right\|_\infty + \|Y\|_\infty\right).$$

Similarly, by $4Mf^2 \leq 1$, $M > 1$ and

$$\sqrt{o}\left\|\frac{1}{\sqrt{n}}Y\right\|_\infty + 3\sqrt{\frac{o\log n}{n}} + 2Mf\sqrt{s}\left\|\frac{1}{n}X^\top Y\right\|_\infty + 2Mf4\sqrt{M}\sqrt{\frac{s\log p}{n}} \geq (1-2Mf^2)\left\|\theta^*\right\|_2,$$

we have

$$\frac{\|\theta^*\|_2}{\sqrt{o}} \leq 8\sqrt{M} + 2\left\|\frac{1}{\sqrt{n}}Y\right\|_\infty + \left\|\frac{1}{\sqrt{n}}X^\top Y\right\|_\infty.$$

Consequently, given $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$, with probability at least $1 - O(p^{-2} + n^{-3})$, we can construct valid initial thresholds $\lambda_{\beta,0}$ and $\lambda_{\theta,0}$ based on

$$\lambda_{\beta,0} \geq 4M\left(4 + \left\|\frac{1}{n}X^\top Y\right\|_\infty + \|Y\|_\infty\right) + \lambda_{\beta,\infty},$$
$$\lambda_{\theta,0} \geq 8\sqrt{M} + 2\left\|\frac{1}{\sqrt{n}}Y\right\|_\infty + \left\|\frac{1}{\sqrt{n}}X^\top Y\right\|_\infty + \lambda_{\theta,\infty}.$$

S4 Proof of Theorem 2

Throughout this proof, Assumption 2 is used in the same concrete form as in the proof of Theorem 1, i.e., $n \geq C_{\mathrm{Th1}}(s\log p + o\log n)$. The proof of Theorem 2 proceeds in two steps. First, we introduce some useful preliminaries. Then, we prove that, under Assumption 3, a carefully chosen stopping time yields an error bound sharper than Theorem 1.

S4.1 Preliminary

By Lemma 1, we have

$$\mathbf{P}\left\{\underbrace{\frac{1}{n}\left\|X_{\cdot S^*}^{\top}\xi\right\|_2 \leq \sqrt{\frac{4Ms+6M\log(1/\varrho)}{n}}}_{=:\mathcal{E}'}\right\} \geq 1-O\left(p^{-2s}+\varrho\right),$$

and this proof is based on the event $\mathcal{E}' \cap \mathcal{E} \cap \mathcal{E}_X$.

In the second-stage iteration, we apply the thresholds of the same values as the endpoints in (S3.7):

$$\lambda_\beta = \lambda_{\beta,\infty}, \quad \lambda_\theta = \lambda_{\theta,\infty},$$

therefore, by Theorem 1, with a probability greater than $1-O(p^{-2}+n^{-3})$ we have the following

$$\|\tilde{\beta}^t_{S^{*c}}\|_0 \leq Bs, \quad \|\tilde{\theta}^t_{O^{*c}}\|_0 \leq Bo, \tag{S4.1}$$

$$\|\tilde{\beta}^t-\beta^*\|_2 \leq (\sqrt{B}+1)\cdot\sqrt{s}\lambda_\beta, \quad \|\tilde{\theta}^t-\theta^*\|_2 \leq (\sqrt{B}+1)\cdot\sqrt{o}\lambda_\theta, \tag{S4.2}$$

hold for all $t \geq 0$, regardless of whether Assumption 3 holds (the definition of B follows (S3.6)). Moreover, as Assumption 3 holds, i.e.,

$$\min_{i\in S^*}|\beta_i^*| \geq \left(\frac{4\kappa}{\kappa-\delta}+\frac{\kappa-\delta}{4\kappa}\right)\cdot\lambda_\beta, \tag{S4.3}$$

we can further prove a sharper error bound of $\tilde{\beta}^t$. Before we prove this refined bound,

we rewrite the decomposition as

$$\begin{aligned}\tilde{H}_{\beta}^{t+1} :=& \tilde{\beta}^t + \frac{\eta}{n} X^\top (Y - X\tilde{\beta}^t - \sqrt{n}\tilde{\theta}^t) \\ =& \beta^* + \Phi(\beta^* - \tilde{\beta}^t) + \frac{\eta}{\sqrt{n}} X^\top (\theta^* - \tilde{\theta}^t) + \eta\Xi \in \mathbb{R}^p, \\ \tilde{H}_{\theta}^{t+1} :=& \tilde{\theta}^t + \frac{\eta}{\sqrt{n}} (Y - X\tilde{\beta}^t - \sqrt{n}\tilde{\theta}^t) \\ =& \theta^* + (\eta - 1)(\theta^* - \tilde{\theta}^t) + \frac{\eta}{\sqrt{n}} X(\beta^* - \tilde{\beta}^t) + \frac{\eta}{\sqrt{n}}\xi \in \mathbb{R}^n.\end{aligned} \tag{S4.4}$$

S4.2 Sharper bound under Assumption 3

In the second-stage iteration, by (S4.1) we guarantee the sparsity of solution sequences $\left\{\tilde{\beta}^t\right\}_{t\geq 0}$ and $\left\{\tilde{\theta}^t\right\}_{t\geq 0}$. Then, under the event $\mathcal{E}' \cap \mathcal{E} \cap \mathcal{E}_X$, for every $t \geq 0$ we have

$$\begin{aligned}&\sum_{i\in S^*} \left(\tilde{H}_{\beta,i}^{t+1}\right)^2 \mathbf{1}(|\tilde{H}_{\beta,i}^{t+1}| < \lambda_\beta) \\ \leq& \sum_{i\in S^*} \lambda_\beta^2 \cdot \mathbf{1}\left(|\beta_i^*| - \left|\langle \Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right| - \eta\,|\Xi_i| < \lambda_\beta\right) \\ \leq& \sum_{i\in S^*} \lambda_\beta^2 \cdot \mathbf{1}\left(\left|\langle \Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right| > |\beta_i^*| - \lambda_\beta - \eta\|\Xi\|_\infty \geq \frac{3\kappa+\delta}{\kappa-\delta}\lambda_\beta\right) \\ \leq& \left(\frac{\kappa-\delta}{3\kappa+\delta}\right)^2 \sum_{i\in S^*} \left(\langle \Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right)^2,\end{aligned} \tag{S4.5}$$

where the second inequality follows from the signal condition (S4.3) and the fact $\|\eta\Xi\|_\infty \leq \frac{\kappa-\delta}{4\kappa}\lambda_\beta$ under the event $\mathcal{E}$.

On the set $\tilde{S}^{t+1} \setminus S^*$, by $\|\eta\Xi\|_\infty \le \frac{\kappa-\delta}{4\kappa}\lambda_\beta$ we have

$$
\begin{aligned}
&\sum_{i\in\tilde{S}^{t+1}\setminus S^*} (\eta\Xi_i)^2 \mathbf{1}\{|\tilde{H}^{t+1}_{\beta,i}| \ge \lambda_\beta\} \\
&\le \sum_{i\in\tilde{S}^{t+1}\setminus S^*} (\eta\Xi_i)^2 \mathbf{1}\left\{\left|\langle\Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right| \ge \lambda_\beta - \|\eta\Xi\|_\infty \ge \frac{3\kappa+\delta}{\kappa-\delta}\|\eta\Xi\|_\infty\right\} \\
&\le \left(\frac{\kappa-\delta}{3\kappa+\delta}\right)^2 \sum_{i\in\tilde{S}^{t+1}\setminus S^*} \left|\langle\Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right|^2 .
\end{aligned}
\tag{S4.6}
$$

Therefore,

$$
\begin{aligned}
&\|\tilde{\beta}^{t+1} - \beta^*\|_2 \\
&\le \Bigg\{ \sum_{i\in S^*} \left(-\tilde{H}^{t+1}_{\beta,i}\mathbf{1}(|\tilde{H}^{t+1}_{\beta,i}| < \lambda_\beta) + \langle\Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle + \eta\Xi_i\right)^2 \\
&\qquad + \sum_{i\in\tilde{S}^{t+1}\setminus S^*} \left(\langle\Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle + \eta\Xi_i\mathbf{1}\left(|\tilde{H}^{t+1}_{\beta,i}| \ge \lambda_\beta\right)\right)^2 \Bigg\}^{1/2} \\
&\le \sqrt{\sum_{i\in\tilde{S}^{t+1}\cup S^*} \langle\Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle^2} + \sqrt{\sum_{i\in\tilde{S}^{t+1}\cup S^*} \frac{\eta^2}{n}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle^2} + \eta\sqrt{\sum_{i\in S^*}\Xi_i^2} \\
&\quad + \sqrt{\sum_{i\in S^*}\left(\tilde{H}^{t+1}_{\beta,i}\right)^2 \mathbf{1}(|\tilde{H}^{t+1}_{\beta,i}| < \lambda_\beta) + \sum_{i\in\tilde{S}^{t+1}\setminus S^*} (\eta\Xi_i)^2\mathbf{1}\left(|\tilde{H}^{t+1}_{\beta,i}| \ge \lambda_\beta\right)} \\
&\le \frac{4\kappa\delta}{3\kappa+\delta}\|\tilde{\beta}^t - \beta^*\|_2 + \frac{8\kappa^2\eta}{(3\kappa+\delta)(\kappa-\delta)} \cdot f\|\tilde{\theta}^t - \theta^*\|_2 + \eta\sqrt{\frac{4Ms + 6M\log(1/\varrho)}{n}},
\end{aligned}
\tag{S4.7}
$$

where the last inequality follows from (S4.5), (S4.6), the sparsity results (S4.1) and the event $\mathcal{E}'$. Combining (S4.7) and (S4.2), with a constant C_1 (depending only on

κ, η, and M), we have

$$\|\tilde{\beta}^{t+1} - \beta^*\|_2 \leq \frac{4\kappa\delta}{3\kappa+\delta}\|\tilde{\beta}^t - \beta^*\|_2 + C_1\left(\sqrt{\frac{s+\log(1/\varrho)}{n}} + \frac{s\log p + o\log n}{n}\right),$$

by $4\kappa\delta < 3\kappa + \delta$, we further get that

$$\|\tilde{\beta}^t - \beta^*\|_2 \leq \left(\frac{4\kappa\delta}{3\kappa+\delta}\right)^t \cdot \|\tilde{\beta}^0 - \beta^*\|_2 + \frac{C_1}{1-\frac{4\kappa\delta}{3\kappa+\delta}}\left(\sqrt{\frac{s+\log(1/\varrho)}{n}} + \frac{s\log p + o\log n}{n}\right)$$

holds for every $t \geq 0$. Additionally, from Theorem 1, we have

$$\mathbf{P}\left\{\|\tilde{\beta}^0 - \beta^*\|_2^2 \lesssim \frac{s\log p}{n} + \frac{o^2\log^2 n}{n^2}\right\} \geq 1 - O\left(p^{-2} + n^{-3}\right).$$

Therefore, with a sufficiently large constant C_2, for every $t \geq C_2 \log n$, with a probability greater than $1 - \varrho - O(p^{-2} + n^{-3})$ we get the sharper error bound:

$$\|\tilde{\beta}^t - \beta^*\|_2 \lesssim \sqrt{\frac{s+\log(1/\varrho)}{n}} + \frac{s\log p + o\log n}{n}.$$

Combining this bound and (S4.1), we complete the proof of Theorem 2.

S5 Proof of Corollary 1

Sharper error bound: Under the beta-min Assumption 3 and taking $\varrho = p^{-2}$, from Theorem 2 we learn that

$$\|\tilde{\beta}^t - \beta^*\|_2 \leq C\left(\sqrt{\frac{s+\log p}{n}} + \frac{s\log p + o\log n}{n}\right), \tag{S5.1}$$

holds for every $t \geq C_2 \log n$, with a probability greater than $1 - O(p^{-2} + n^{-3})$.

On true support: On the true support $S^* = \mathrm{supp}(\beta^*)$, we have

$$\begin{aligned}
\left|\mathrm{supp}(\beta^*) - \mathrm{supp}(\tilde{\beta}^t)\right| &= \sum_{i\in S^*} \mathbf{1}\{|\tilde{H}^t_{\beta,i}| < \lambda_\beta\} \\
&\overset{(i)}{\leq} \sum_{i\in S^*} \mathbf{1}\left(\left|\langle \Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right| > \frac{3\kappa+\delta}{\kappa-\delta}\lambda_\beta\right) \\
&\lesssim \sum_{i\in S^*} \frac{\left|\langle \Phi_{\cdot i}, \beta^* - \tilde{\beta}^t\rangle + \frac{\eta}{\sqrt{n}}\langle X_{\cdot i}, \theta^* - \tilde{\theta}^t\rangle\right|^2}{\lambda_\beta^2} \\
&\lesssim \frac{\|\beta^* - \tilde{\beta}^t\|_2^2 + f^2\|\theta^* - \tilde{\theta}^t\|_2^2}{\lambda_\beta^2} \\
&\overset{(ii)}{\lesssim} \frac{\frac{s+\log p}{n} + \left(\frac{s\log p}{n}\right)^2 + \left(\frac{o\log n}{n}\right)^2}{\frac{s\log p}{n} + \left(\frac{o\log n}{n}\right)^2} \cdot s \\
&\overset{(iii)}{=} \left(\frac{1}{\log p} + \frac{1}{s} + \frac{s\log p}{n} + o(1)\right)\cdot s \;=\; o(s),
\end{aligned} \tag{S5.2}$$

where inequality (i) comes from (S4.5), inequality (ii) comes from the sharper bound (S5.1) and (S4.2), and inequality (iii) comes from the condition $n \succ s\log p \succ \frac{o^2\log^2 n}{n}$.

On non-support: By $\sum_{i\notin S^*}\mathbf{1}(\tilde{\beta}_i^t\neq 0)\leq Bs$, we conclude that

$$\begin{aligned}
&\left|\mathrm{supp}(\tilde{\beta}^t)-\mathrm{supp}(\beta^*)\right| = \sum_{i\notin S^*}\mathbf{1}\{|\tilde{H}^t_{\beta,i}|\geq\lambda_\beta\}\\
&\overset{(i)}{\leq}\sum_{i\notin S^*}\mathbf{1}\left\{\left|\langle\Phi_{\cdot i},\beta^*-\tilde{\beta}^t\rangle+\frac{\eta}{\sqrt{n}}\langle X_{\cdot i},\theta^*-\tilde{\theta}^t\rangle\right|\geq\lambda_\beta-\|\eta\Xi\|_\infty\geq\frac{3\kappa+\delta}{4\kappa}\lambda_\beta\right\}\\
&=o(s),
\end{aligned}$$

where inequality follows from (S4.6), and the last equation comes from a similar process in (S5.2). Therefore, by taking the stopping time $t\geq C_2\log n$ we have

$$\mathbf{P}\left\{\left|\mathrm{supp}(\beta^*)\,\triangle\,\mathrm{supp}(\tilde{\beta}^t)\right|=o(s)\right\}\geq 1-O(p^{-2}+n^{-3}),$$

which completes the proof of Corollary 1.

S6 Proof of Theorem 3

Throughout this proof, Assumption 2 is used in the concrete form

$$n\geq C_{\mathrm{Th3}}\{s\log p+o\log n\},$$

where C_{Th3} is a fixed constant depending only on M,η,κ, chosen large enough to dominate the concrete sample-size constants used in the proof of Theorem 1, Lemma 2, and Lemma 3, satisfying:

$$C_{\mathrm{Th3}}\geq\max\left\{C_{\mathrm{Th1}},\ 30M^2C_\Sigma,\ 6MC_M^2,\ \frac{16C_M^2}{(\kappa-\delta)^4}\right\}.$$

The proof of Theorem 3 proceeds in three steps: First, we introduce some useful preliminaries. Second, by mathematical induction, we prove that, under beta-min and theta-min conditions, i.e., Assumptions 3 and 4, the output of Algorithm 2 converges to the oracle estimation $\beta^\dagger$. Finally, we derive the variable selection consistency.

S6.1 Preliminary

Assume $supp(\theta^*) \neq \emptyset$. For ease of exposition, we will denote $(O^*)^c$ as $-O^*$ throughout the following proof. Recall the formula of the oracle estimation:

$$\begin{aligned}
\beta^\dagger_{S^*} &:= \beta^*_{S^*} + (X^\top_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^\top_{-O^*,S^*} \xi_{-O^*} \in \mathbb{R}^{|S^*|}, \\
\beta^\dagger_{-S^*} &:= \mathbf{0} \in \mathbb{R}^{p-|S^*|}, \\
\theta^\dagger_{O^*} &:= \theta^*_{O^*} + \frac{1}{\sqrt{n}} \Big\{ \xi_{O^*} - X_{O^*,S^*} (X^\top_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^\top_{-O^*,S^*} \xi_{-O^*} \Big\} \in \mathbb{R}^{|O^*|}, \\
\theta^\dagger_{-O^*} &:= \mathbf{0} \in \mathbb{R}^{n-|O^*|}.
\end{aligned} \tag{S6.1}$$

Therefore, we can rewrite the decomposition in the second-stage algorithm as:

$$\begin{aligned}
\tilde{H}^{t+1}_\beta =& \tilde{\beta}^t + \frac{\eta}{n} X^\top (Y - X\tilde{\beta}^t - \sqrt{n}\tilde{\theta}^t) \\
=& \beta^\dagger + \Phi(\beta^\dagger - \tilde{\beta}^t) + \frac{\eta}{\sqrt{n}} X^\top (\theta^\dagger - \tilde{\theta}^t) + \eta \Xi^\dagger, \\
\tilde{H}^{t+1}_\theta =& \tilde{\theta}^t + \frac{\eta}{\sqrt{n}} (Y - X\tilde{\beta}^t - \sqrt{n}\tilde{\theta}^t) \\
=& \theta^\dagger + (\eta - 1)(\theta^\dagger - \tilde{\theta}^t) + \frac{\eta}{\sqrt{n}} X (\beta^\dagger - \tilde{\beta}^t) + \frac{\eta}{\sqrt{n}} \xi^\dagger,
\end{aligned} \tag{S6.2}$$

where

$$\begin{aligned}
\Xi^{\dagger}_{S^*} &:= \mathbf{0} \in \mathbb{R}^{|S^*|}, \\
\Xi^{\dagger}_{-S^*} &:= \frac{1}{n} X^{\top}_{-O^*,-S^*} \Big\{ I_{n-o} - X_{-O^*,S^*} (X^{\top}_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^{\top}_{-O^*,S^*} \Big\} \xi_{-O^*} \in \mathbb{R}^{p-|S^*|}, \\
\xi^{\dagger}_{O^*} &:= \mathbf{0} \in \mathbb{R}^{|O^*|},
\end{aligned}$$

and

$$\xi^{\dagger}_{-O^*} := \Big\{ I_{n-o} - X_{-O^*,S^*} (X^{\top}_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^{\top}_{-O^*,S^*} \Big\} \xi_{-O^*} \in \mathbb{R}^{n-|O^*|}.$$

We specify Assumptions 3 and 4 as

$$\min_{i \in S^*} |\beta^*_i| \geq \frac{4\kappa}{\kappa - \delta} \lambda_\beta + 5\sqrt{\frac{M \log p}{n}}, \quad \min_{k \in O^*} |\theta^*_k| \geq \frac{4\kappa}{\kappa - \delta} \lambda_\theta + 4\sqrt{\frac{\log n}{n}}, \tag{S6.3}$$

where $\lambda_\beta = \lambda_{\beta,\infty}$ and $\lambda_\theta = \lambda_{\theta,\infty}$ follows from (S3.7). Define the event

$$\mathcal{E}_{oracle} := \left\{ \begin{array}{cc} \|\Xi^{\dagger}\|_\infty < 4\sqrt{\frac{M \log p}{n}}, & \left\| \frac{\xi^{\dagger}}{\sqrt{n}} \right\|_\infty < 3\sqrt{\frac{\log n}{n}}; \\ \left\|\beta^{\dagger} - \beta^*\right\|_\infty < 5\sqrt{\frac{M \log p}{n}}, & \left\|\theta^{\dagger} - \theta^*\right\|_\infty < 4\sqrt{\frac{\log n}{n}}. \end{array} \right\}$$

And the following proof is based on the event $\mathcal{E}_{oracle} \cap \mathcal{E}_X$, which holds with a probability greater than $1 - O(p^{-2} + n^{-3})$ by Lemma 2 and Proposition 1.

S6.2 Convergence to oracle estimation

Similar to (S4.5), under the event $\mathcal{E}_{oracle}$ and the beta-min condition in (S6.3), we have

$$
\begin{aligned}
&\sum_{i\in S^*}\left(\tilde{H}^{t+1}_{\beta,i}\right)^2\mathbf{1}(|\tilde{H}^{t+1}_{\beta,i}|<\lambda_\beta)\\
\le&\sum_{i\in S^*}\lambda_\beta^2\cdot\mathbf{1}\left(|\beta_i^\dagger|-\left|\langle\Phi_{\cdot i},\beta^\dagger-\tilde{\beta}^t\rangle+\frac{\eta}{\sqrt{n}}\langle X_{\cdot i},\theta^\dagger-\tilde{\theta}^t\rangle\right|<\lambda_\beta\right)\\
\le&\sum_{i\in S^*}\lambda_\beta^2\cdot\mathbf{1}\left(\left|\langle\Phi_{\cdot i},\beta^\dagger-\tilde{\beta}^t\rangle+\frac{1}{\sqrt{n}}\langle X_{\cdot i},\theta^\dagger-\tilde{\theta}^t\rangle\right|>|\beta_i^*|-\|\beta^\dagger-\beta^*\|_\infty-\lambda_\beta\ge\frac{3\kappa+\delta}{\kappa-\delta}\lambda_\beta\right)\\
\le&\left(\frac{\kappa-\delta}{3\kappa+\delta}\right)^2\sum_{i\in S^*}\left(\langle\Phi_{\cdot i},\beta^\dagger-\tilde{\beta}^t\rangle+\frac{\eta}{\sqrt{n}}\langle X_{\cdot i},\theta^\dagger-\tilde{\theta}^t\rangle\right)^2,
\end{aligned}
\tag{S6.4}
$$

leading that

$$
\begin{aligned}
&\|\tilde{\beta}^{t+1}-\beta^{\dagger}\|_2 \\
\leq&\left\{\sum_{i\in S^*}\left(-\tilde{H}_{\beta,i}^{t+1}\mathbf{1}(|\tilde{H}_{\beta,i}^{t+1}|<\lambda_\beta)+\langle\Phi_{\cdot i},\beta^{\dagger}-\tilde{\beta}^{t}\rangle+\frac{\eta}{\sqrt{n}}\langle X_{\cdot i},\theta^{\dagger}-\tilde{\theta}^{t}\rangle\right)^2\right. \\
&\left.+\sum_{i\in\tilde{S}^{t+1}\setminus S^*}\left(\langle\Phi_{\cdot i},\beta^{\dagger}-\tilde{\beta}^{t}\rangle+\frac{\eta}{\sqrt{n}}\langle X_{\cdot i},\theta^{\dagger}-\tilde{\theta}^{t}\rangle+\eta\Xi_i^{\dagger}\mathbf{1}\left(|\tilde{H}_{\beta,i}^{t+1}|\geq\lambda_\beta\right)\right)^2\right\}^{1/2} \\
\leq&\sqrt{\sum_{i\in\tilde{S}^{t+1}\cup S^*}\langle\Phi_{\cdot i},\beta^{\dagger}-\tilde{\beta}^{t}\rangle^2}+\sqrt{\sum_{i\in\tilde{S}^{t+1}\cup S^*}\frac{\eta^2}{n}\langle X_{\cdot i},\theta^{\dagger}-\tilde{\theta}^{t}\rangle^2} \\
&+\left\{\sum_{i\in S^*}\left(\tilde{H}_{\beta,i}^{t+1}\right)^2\mathbf{1}(|\tilde{H}_{\beta,i}^{t+1}|<\lambda_\beta)\right. \\
&\left.+\sum_{i\in\tilde{S}^{t+1}\setminus S^*}\left(\eta\Xi_i^{\dagger}\right)^2\mathbf{1}\left(\left|\langle\Phi_{\cdot i},\beta^{\dagger}-\tilde{\beta}^{t}\rangle+\frac{\eta\langle X_{\cdot i},\theta^{\dagger}-\tilde{\theta}^{t}\rangle}{\sqrt{n}}\right|\geq\lambda_\beta-\|\eta\Xi^{\dagger}\|_\infty\geq\frac{3\kappa+\delta}{\kappa-\delta}|\eta\Xi_i^{\dagger}|\right)\right\}^{1/2} \\
\leq&\frac{4\kappa}{3\kappa+\delta}\left(\delta\|\tilde{\beta}^{t}-\beta^{\dagger}\|_2+\frac{2\kappa\eta}{\kappa-\delta}\cdot f\|\tilde{\theta}^{t}-\theta^{\dagger}\|_2\right),
\end{aligned}
\tag{S6.5}
$$

where the second inequality follows from the event $\mathcal{E}_{oracle}$. The last inequality follows from (S6.4), the techniques (S3.15) and (S3.16), and the sparsity result (S4.1) derived in Theorem 2, which holds consistently for every $t \geq 0$ with a probability greater than $1 - O\left(p^{-2} + n^{-3}\right)$.

Similarly, the event $\mathcal{E}_{oracle}$ and the theta-min condition in (S6.3) yield that

$$\begin{aligned}
&\sum_{i\in O^*}\left(\tilde{H}^{t+1}_{\theta,i}\right)^2\mathbf{1}(|\tilde{H}^{t+1}_{\theta,i}|<\lambda_\theta)\\
\le&\sum_{i\in O^*}\lambda_\theta^2\cdot\mathbf{1}\left(\left|(\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)\right|>|\theta^*_i|-\|\theta^\dagger-\theta^*\|_\infty-\lambda_\theta\ge\frac{3\kappa+\delta}{\kappa-\delta}\lambda_\theta\right)\\
\le&\left(\frac{\kappa-\delta}{3\kappa+\delta}\right)^2\sum_{i\in O^*}\left((\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)\right)^2,
\end{aligned}\tag{S6.6}$$

which leads to

$$\begin{aligned}
&\|\tilde{\theta}^{t+1}-\theta^\dagger\|_2\\
\le&\left\{\sum_{i\in O^*}\left(-\tilde{H}^{t+1}_{\theta,i}\mathbf{1}(|\tilde{H}^{t+1}_{\theta,i}|<\lambda_\theta)+(\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)\right)^2\right.\\
&\left.\quad+\sum_{i\in\tilde{O}^{t+1}\setminus O^*}\left((\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)+\frac{\eta}{\sqrt{n}}\xi^\dagger_i\mathbf{1}\left(|\tilde{H}^{t+1}_{\theta,i}|\ge\lambda_\theta\right)\right)^2\right\}^{1/2}\\
\le&\sqrt{\sum_{i\in\tilde{O}^{t+1}\cup O^*}\left((\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)\right)^2}\\
&+\left\{\sum_{i\in O^*}\left(\tilde{H}^{t+1}_{\theta,i}\right)^2\mathbf{1}(|\tilde{H}^{t+1}_{\theta,i}|<\lambda_\theta)\right.\\
&\left.\quad+\sum_{i\in\tilde{O}^{t+1}\setminus O^*}\left(\frac{\eta\xi^\dagger_i}{\sqrt{n}}\right)^2\mathbf{1}\left(\left|(\eta-1)(\theta^\dagger_i-\tilde{\theta}^t_i)+\frac{\eta}{\sqrt{n}}X_{i\cdot}(\beta^\dagger-\tilde{\beta}^t)\right|\ge\lambda_\theta-\left|\frac{\eta\xi^\dagger_i}{\sqrt{n}}\right|\ge\frac{3\kappa+\delta}{\kappa-\delta}\left|\frac{\eta\xi^\dagger_i}{\sqrt{n}}\right|\right)\right\}^{1/2}\\
\le&\frac{4\kappa}{3\kappa+\delta}\left(\delta\|\tilde{\theta}^t-\theta^\dagger\|_2+\frac{2\kappa\eta}{\kappa-\delta}\cdot f\|\tilde{\beta}^t-\beta^\dagger\|_2\right).
\end{aligned}\tag{S6.7}$$

Combining (S6.5) and (S6.7), we conclude that

$$\begin{aligned}\|\tilde{\beta}^t-\beta^\dagger\|_2 &\le \left(\frac{4\kappa\delta}{3\kappa+\delta}+\frac{8\kappa^2\eta\cdot f}{(3\kappa+\delta)(\kappa-\delta)}\right)^t\cdot\frac{\|\tilde{\beta}^0-\beta^\dagger\|_2+\|\tilde{\theta}^0-\theta^\dagger\|_2}{2}\\ &\quad+\left(\frac{4\kappa\delta}{3\kappa+\delta}-\frac{8\kappa^2\eta\cdot f}{(3\kappa+\delta)(\kappa-\delta)}\right)^t\cdot\frac{\|\tilde{\beta}^0-\beta^\dagger\|_2-\|\tilde{\theta}^0-\theta^\dagger\|_2}{2},\\ &\le \left(\frac{4\kappa\delta}{3\kappa+\delta}+\frac{8\kappa^2\eta\cdot f}{(3\kappa+\delta)(\kappa-\delta)}\right)^t\cdot\left(\|\tilde{\beta}^0-\beta^\dagger\|_2+\|\tilde{\theta}^0-\theta^\dagger\|_2\right)\end{aligned}\tag{S6.8}$$

simultaneously hold for every $t\ge 0$. Define $C_0:=\|\tilde{\beta}^0-\beta^\dagger\|_2+\|\tilde{\theta}^0-\theta^\dagger\|_2<\infty$ and $r:=\frac{4\kappa\delta}{3\kappa+\delta}+\frac{8\kappa^2\eta\cdot f}{(3\kappa+\delta)(\kappa-\delta)}$, by the definition of f in (S3.6) and the sample size assumption $n\ge\frac{16C_M^2}{(\kappa-\delta)^4}\cdot(s\log p+o\log n)$, we conclude that $r\le\frac{2\kappa+4\kappa\delta-2\delta}{3\kappa+\delta}<1$. Therefore, we prove that, with a probability greater than $1-O(p^{-2}+n^{-3})$,

$$\|\tilde{\beta}^t-\beta^\dagger\|_2\le C_0\times r^t$$

holds for every $t\ge 0$ in the second-stage algorithm.

S6.3 Oracle estimation rate

According to (S6.1), with a probability greater than $1-O(p^{-2})$, we have

$$\begin{aligned}\|\beta^\dagger-\beta^*\|_2&=\left\|(X_{-O^*,S^*}^\top X_{-O^*,S^*})^{-1}X_{-O^*,S^*}^\top\xi_{-O^*}\right\|_2\\ &\le\frac{3M}{n}\cdot\left\|X_{-O^*,S^*}^\top\xi_{-O^*}\right\|_2,\end{aligned}\tag{S6.9}$$

where the last inequality follows from Lemma 3, which implies

$$\left\|(X_{-O^*,S^*}^\top X_{-O^*,S^*})^{-1}\right\|_2=\frac{1}{\Lambda_{\min}\left(X_{-O^*,S^*}^\top X_{-O^*,S^*}\right)}\le\frac{3M}{n}.$$

For a given design X, from Theorem 2.1 of Hsu et al. (2012), we learn that

$$\mathbf{P}_{\xi|X}\Bigg\{ \left\|X^\top_{-O^*,S^*}\xi_{-O^*}\right\|_2^2 \geq tr\left(X^\top_{-O^*,S^*}X_{-O^*,S^*}\right) + 2\left\|X^\top_{-O^*,S^*}X_{-O^*,S^*}\right\|_F\sqrt{u} + 2\Lambda_{\max}\left(X^\top_{-O^*,S^*}X_{-O^*,S^*}\right)u \;\Big|\; X\Bigg\} \leq e^{-u}. \tag{S6.10}$$

Additionally, if $X \in \mathcal{E}_X$, we further conclude that

$$\begin{aligned} tr\left(X^\top_{-O^*,S^*}X_{-O^*,S^*}\right) \leq & tr\left(X^\top_{\cdot,S^*}X_{\cdot,S^*}\right) \leq 2Mns, \\ \left\|X^\top_{-O^*,S^*}X_{-O^*,S^*}\right\|_F^2 \leq & s\left\|X^\top_{-O^*,S^*}X_{-O^*,S^*}\right\|_2^2 \leq s\left\|X^\top_{\cdot,S^*}X_{\cdot,S^*}\right\|_2^2 \leq 4M^2n^2s, \\ \Lambda_{\max}\left(X^\top_{-O^*,S^*}X_{-O^*,S^*}\right) \leq & \left\|X^\top_{\cdot,S^*}X_{\cdot,S^*}\right\|_2 \leq 2Mn. \end{aligned}$$

Therefore, by taking $u = \log(1/\varrho)$ into (S6.10), we have

$$\begin{aligned} & \mathbf{P}_{\xi,X}\left(\left\|X^\top_{-O^*,S^*}\xi_{-O^*}\right\|_2^2 \geq 2Mn(2s+3\log(1/\varrho))\right) \\ \leq & \mathbf{P}_{\xi,X}\left(\mathcal{E}_X^c\right) + \mathbf{E}_X\left\{\mathbf{1}(\mathcal{E}_X)\cdot\mathbf{P}_{\xi|X}\left(\left\|X^\top_{-O^*,S^*}\xi_{-O^*}\right\|_2^2 \geq 2Mn(2s+3\log(1/\varrho))\;\big|X\right)\right\} \\ \leq & 4p^{-2s} + \varrho. \end{aligned} \tag{S6.11}$$

Combining (S6.8), (S6.9) and (S6.11), with taking the stopping time $t \geq \frac{\log(C_0n)}{\log(1/r)}$, we guarantee that

$$\sup_{i\in[p]}|\beta_i^\dagger - \tilde{\beta}_i^t| \leq \|\tilde{\beta}^t - \beta^\dagger\|_2 \leq \frac{1}{n}, \tag{S6.12}$$

and

$$\|\tilde{\beta}^t - \beta^*\|_2 \leq \|\tilde{\beta}^t - \beta^\dagger\|_2 + \|\beta^\dagger - \beta^*\|_2 \;\leq\; \frac{1}{n} + 9M^{3/2}\sqrt{\frac{s+\log(1/\varrho)}{n}} \;\asymp\; \sqrt{\frac{s+\log(1/\varrho)}{n-o}},$$

with a probability at least $1 - \varrho - O(p^{-2} + n^{-3})$.

S6.4 Variable selection consistency

By (S6.12), we conclude that:

- For every $i \in S^*$, under the signal condition (S6.3) we have

$$|\tilde{\beta}_i^t| \geq |\beta_i^*| - |\beta_i^\dagger - \beta_i^*| - |\beta_i^\dagger - \tilde{\beta}_i^t| \geq \frac{4\kappa}{\kappa - \delta}\lambda_\beta - \frac{1}{n} > 0,$$

  which proves that the whole support set S^* is recovered.

- For every $i \notin S^*$, we have

$$|\tilde{\beta}_i^t| \leq |\beta_i^\dagger| + |\beta_i^\dagger - \tilde{\beta}_i^t| \leq \frac{1}{n} < \lambda_\beta,$$

  which proves that we cannot discover any entry on $(S^*)^c$ by using the hard-thresholding parameter λ_β.

Therefore, with a probability greater than $1 - O(p^{-2} + n^{-3})$, we have

$$\mathrm{supp}(\tilde{\beta}^t) = \mathrm{supp}(\beta^\dagger) = \mathrm{supp}(\beta^*), \quad \text{for every } t \geq C \log n,$$

which achieves the variable selection consistency of $\tilde{\beta}$. Meanwhile, by utilizing the technique similar to (S6.8), we also establish the convergence of the estimator $\{\tilde{\theta}^t\}_{t\geq 0}$ to the oracle $\theta^\dagger$, therefore demonstrating the selection consistency of $\tilde{\theta}$.

S7 Proof of Corollary 2

For every vector $\gamma \in \mathbb{R}^s$ with $\|\gamma\|_2 < \infty$, according to Theorem 3, we have

$$\sqrt{n}\gamma^\top \tilde{\beta}_{S^*} = \sqrt{n}\gamma^\top \beta^\dagger_{S^*} + o_p(1).$$

Therefore, by Slutsky's Lemma, we only need to prove the asymptotic normality of $\beta^\dagger_{S^*}$. By (S6.1),

$$\begin{aligned}
&\sqrt{n}\gamma^\top(\beta^\dagger_{S^*} - \beta^*_{S^*}) \\
=&\sqrt{n}\gamma^\top \left(\frac{X^\top_{-O^*,S^*}X_{-O^*,S^*}}{n-o}\right)^{-1} \frac{X^\top_{-O^*,S^*}\xi_{-O^*}}{n-o} \\
=&\frac{n}{n-o}\cdot\frac{1}{\sqrt{n}}\gamma^\top\Sigma^{-1}_{S^*,S^*}X^\top_{-O^*,S^*}\xi_{-O^*} \\
&+\sqrt{n}\gamma^\top\Sigma^{-1}_{S^*,S^*}\left\{\Sigma_{S^*,S^*} - \frac{X^\top_{-O^*,S^*}X_{-O^*,S^*}}{n-o}\right\}\left(\frac{X^\top_{-O^*,S^*}X_{-O^*,S^*}}{n-o}\right)^{-1}\frac{X^\top_{-O^*,S^*}\xi_{-O^*}}{n-o} \\
=&\frac{n}{n-o}\cdot\frac{1}{\sqrt{n}}\gamma^\top\Sigma^{-1}_{S^*,S^*}X^\top_{-O^*,S^*}\xi_{-O^*} + O_p\left(\frac{ns\log p}{(n-o)^{3/2}}\right).
\end{aligned}$$

where the last equality follows from (S6.11) and a similar technique used in (S2.2). Define $\epsilon_i := \frac{1}{\sqrt{n}}\gamma^\top\Sigma^{-1}_{S^*,S^*}X^\top_{i,S^*}\xi_i$. Clearly, $\{\epsilon_i\}_{i\in -O^*}$ are independent and zero-mean, and

$$\sum_{i\in -O^*}\mathrm{Var}(\epsilon_i) = \sum_{i\in -O^*}\frac{1}{n}c_\xi\sigma^2\gamma^\top\Sigma^{-1}_{S^*,S^*}\gamma \geq \frac{c_\xi\sigma^2\|\gamma\|_2^2}{2M}, \tag{S7.1}$$

where $c_\xi = \frac{\mathrm{var}(\xi_i)}{\sigma^2}$ is a posotive constant. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sum_{i\in -O^*} \mathbf{E}|\epsilon_i|^3 &= n^{-3/2} \sum_{i\in -O^*} \mathbf{E}\left(|\gamma^\top \Sigma^{-1}_{S^*,S^*} X^\top_{i,S^*}|^3\right) \cdot \mathbf{E}|\xi_i|^3 \\ &\lesssim \frac{\|\gamma\|_2^3 \cdot \|\Sigma^{-1}_{S^*,S^*}\|_2^3 \cdot \mathbf{E}(\|X^\top_{i,S^*}\|_2^3) \cdot \sigma^3}{n^{1/2}} \\ &\lesssim \frac{\sigma^3 s^{3/2}}{n^{1/2}} \|\gamma\|_2^3 \ ,\end{aligned} \tag{S7.2}$$

where the last inequality comes from the property of sub-Gaussian random variable. Since

$$\frac{\sum_{i\in -O^*} \mathbf{E}|\epsilon_i|^3}{\left(\sum_{i\in -O^*} \mathrm{Var}(\epsilon_i)\right)^{3/2}} \lesssim s^{3/2} n^{-1/2},$$

by Lyapunov's central limit theorem and Slutsky's Lemma, with $n \succ s^3,\ n \gtrsim o \log n$, we conclude that

$$\frac{1}{\sqrt{n}} \gamma^\top \Sigma^{-1}_{S^*,S^*} X^\top_{-O^*,S^*} \xi_{-O^*} \xrightarrow{D} \mathcal{N}\left(0\ ,\ c_\xi \sigma^2 \gamma^\top \Sigma^{-1}_{S^*,S^*} \gamma\right).$$

Furthermore, combining with $\sqrt{n} \succ s \log p$, we get

$$\sqrt{n}\gamma^\top (\tilde{\beta}_{S^*} - \beta^*_{S^*}) \xrightarrow{D} \mathcal{N}\left(0\ ,\ c_\xi \sigma^2 \gamma^\top \Sigma^{-1}_{S^*,S^*} \gamma\right)$$

for every $\gamma \in \mathbb{R}^s$ with $\|\gamma\|_2 < \infty$. Therefore, we complete the proof of Corollary 2.

S8 Proof of minimax lower bounds

The proof of lower bounds proceeds in three steps. We first follow the proof technique in Chen et al. (2018) and get an estimation lower bound of the ϵ-contamination model. Then, we employ a Bayesian method to connect our empirical contamination model with the ϵ-contamination model, thereby establishing Theorem 4. We finally apply a similar technique to obtain the lower bound for variable selection, and complete the proof of Theorem 5.

For ease of display, we denote by $\pi_\Sigma(X_{i,\cdot})$ the distribution of the i-th observation $X_{i,\cdot} \in \mathbb{R}^p$ for each uncontaminated index i. We also assume the noise ξ_i follows from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ independently for each $i \in [n]$.

S8.1 Estimation lower bound related to ϵ'-contamination model

We assume s is an even number, and construct two coefficient vectors as

$$\begin{aligned} \beta^{(1)} &:= (\underbrace{\lambda, \cdots, \lambda}_{s}, \underbrace{0, \cdots, 0}_{p-s})^\top \in \mathbb{R}^p, \\ \beta^{(2)} &:= (\underbrace{0, \cdots, 0}_{s/2}, \underbrace{\lambda, \cdots, \lambda}_{s}, \underbrace{0, \cdots, 0}_{p-3s/2})^\top \in \mathbb{R}^p, \end{aligned} \tag{S8.1}$$

where $\lambda = \sqrt{\frac{1}{8C_{2s}}} \cdot \frac{\sigma}{\sqrt{s}} \cdot \frac{o}{n}$. Let $\mathbf{P}_1$ denote the joint distribution of $(X_{i\cdot}, Y_i)$ given $\beta^{(1)}$, where assume that $X_{i,\cdot} \in \mathbb{R}^{1\times p}$ has marginal density function $\pi_\Sigma(\cdot)$ and, conditional on $X_{i\cdot}$ and $\beta^{(1)}$, Y_i follows a normal distribution $\mathcal{N}(X_{i,\cdot}\beta^{(1)}, \sigma^2)$. Similarly, we define

$\mathbf{P}_2$ as the joint distribution of $(X_{i\cdot}, Y_i)$ given $\beta^{(2)}$. It is straightforward to learn that

$$\begin{aligned} 0 < \{TV(\mathbf{P}_1, \mathbf{P}_2)\}^2 \; \leq & KL(\mathbf{P}_1 \| \mathbf{P}_2) \\ = & \frac{1}{2\sigma^2} \left(\beta^{(1)} - \beta^{(2)}\right)^\top \Sigma \left(\beta^{(1)} - \beta^{(2)}\right) \\ \overset{(i)}{\leq} & \frac{C_{2s}}{2\sigma^2} \cdot s\lambda^2 \\ = & \left(\frac{o}{4n}\right)^2 \; < \; \left(\frac{o/(4n)}{1 - o/(4n)}\right)^2, \end{aligned} \tag{S8.2}$$

where inequality (i) follows from $\sup_{S \subset [p]: |S| \leq 2s} \Lambda_1(\Sigma_{S,S}) \leq C_{2s}$ and the construction in (S8.1). We then define $\epsilon' := \frac{TV(\mathbf{P}_1, \mathbf{P}_2)}{1 + TV(\mathbf{P}_1, \mathbf{P}_2)}$, and it is directly to learn that $0 < \epsilon' < o/(4n)$.

Define density functions

$$p_1 = \frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\left(\mathbf{P}_1 + \mathbf{P}_2\right)}, \quad p_2 = \frac{\mathrm{d}\mathbf{P}_2}{\mathrm{d}\left(\mathbf{P}_1 + \mathbf{P}_2\right)}.$$

Define $\mathbf{Q}_1$ and $\mathbf{Q}_2$ (both are the joint distributions of $(X_{i\cdot}, Y_i)$) by their density functions

$$\begin{aligned} \frac{\mathrm{d}\mathbf{Q}_1}{\mathrm{d}\left(\mathbf{P}_1 + \mathbf{P}_2\right)} &= \frac{(p_2 - p_1) \cdot \mathbf{1}(p_2 \geq p_1)}{\mathrm{TV}\left(\mathbf{P}_1, \mathbf{P}_2\right)}, \\ \frac{\mathrm{d}\mathbf{Q}_2}{\mathrm{d}\left(\mathbf{P}_1 + \mathbf{P}_2\right)} &= \frac{(p_1 - p_2) \cdot \mathbf{1}(p_1 > p_2)}{\mathrm{TV}\left(\mathbf{P}_1, \mathbf{P}_2\right)}. \end{aligned} \tag{S8.3}$$

Following the proof of Theorem 5.1 in (Chen et al., 2018), it is easy to check that

both $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are well-defined probability measures. It can also be checked that

$$\begin{aligned}\frac{\mathrm{d}\left((1-\epsilon')\mathbf{P}_1+\epsilon'\mathbf{Q}_1\right)}{\mathrm{d}\left(\mathbf{P}_1+\mathbf{P}_2\right)} &= (1-\epsilon')\,p_1+\epsilon'\frac{(p_2-p_1)\cdot\mathbf{1}(p_2\geq p_1)}{\epsilon'/(1-\epsilon')}\\ &= (1-\epsilon')\,p_2+\epsilon'\frac{(p_1-p_2)\cdot\mathbf{1}(p_1>p_2)}{\epsilon'/(1-\epsilon')}\\ &= \frac{\mathrm{d}\left((1-\epsilon')\mathbf{P}_2+\epsilon'\mathbf{Q}_2\right)}{\mathrm{d}\left(\mathbf{P}_1+\mathbf{P}_2\right)}.\end{aligned} \tag{S8.4}$$

We construct the prior of β as

$$\pi_\beta(\beta=\beta^{(1)})=\pi_\beta(\beta=\beta^{(2)})=1/2, \tag{S8.5}$$

where recall the construction of $\beta^{(1)}$ and $\beta^{(2)}$ in (S8.1). Therefore, for any $q\in[1,2]$, we derive that

$$\begin{aligned}&\inf_{\hat\beta}\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)}\left(\|\hat\beta-\beta\|_q^q\right)\\ \geq&\frac{s\lambda^q}{2^q}\cdot\inf_{\hat\beta}\frac{1}{2}\left\{[(1-\epsilon')\mathbf{P}_1+\epsilon'\mathbf{Q}_1]\left(\|\hat\beta-\beta^{(1)}\|_q\geq\frac{s^{1/q}\lambda}{2}\right)\right.\\ &\qquad\left.+[(1-\epsilon')\mathbf{P}_2+\epsilon'\mathbf{Q}_2]\left(\|\hat\beta-\beta^{(2)}\|_q\geq\frac{s^{1/q}\lambda}{2}\right)\right\}\\ \overset{(i)}{\geq}&\frac{s\lambda^q}{8}\cdot\inf_{\hat\beta}\left\{[(1-\epsilon')\mathbf{P}_1+\epsilon'\mathbf{Q}_1]\left(\psi^*(\hat\beta)=2\right)+[(1-\epsilon')\mathbf{P}_2+\epsilon'\mathbf{Q}_2]\left(\psi^*(\hat\beta)=1\right)\right\}\\ =&\frac{1}{8(8C_{2s})^{q/2}}\cdot\frac{\sigma^q s^{1-q/2}o^q}{n^q}\quad\left(=\frac{s\lambda^q}{8}\right),\end{aligned} \tag{S8.6}$$

where $\mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)}$ denotes an expectation based on the joint distribution of $(X_{i,\cdot},Y_i)_{i\in[n]}$, with each $(X_{i,\cdot},Y_i)\sim(1-\epsilon')\mathbf{P}_\ell+\epsilon'\mathbf{Q}_\ell$ independently for $\beta=\beta^{(\ell)}$, and we use $\mathbf{Q}(\beta^{(\ell)})=\mathbf{Q}_\ell$ to emphasize that in our construction the contamination distribu-

tion is partially determined by β. Additionally, in inequality (i) we define the selector $\psi^*(\hat{\beta}) = \arg\min_{\ell\in\{1,2\}} \|\hat{\beta} - \beta^{(\ell)}\|_q$, and this inequality holds because that $\psi^*(\hat{\beta}) = 2 \Rightarrow \|\hat{\beta} - \beta^{(1)}\|_q \geq \|\hat{\beta} - \beta^{(2)}\|_q \Rightarrow s^{1/q}\lambda = \|\beta^{(1)} - \beta^{(2)}\|_q \leq 2\|\hat{\beta} - \beta^{(1)}\|_q$. And the last equality follows from (S8.4).

S8.2 Estimation lower bound in Theorem 4

Section S8.1 gives a minimax estimation lower bound under the Huber contamination model $(X_{i,\cdot}, Y_i) \sim (1-\epsilon')\mathbf{P} + \epsilon'\mathbf{Q}$. However, this lower bound (S8.6) cannot be applied directly in our frequentist-setting model. Given β, n, k, σ and $\pi_\Sigma(\cdot)$, we define the distribution class of $(X_i, Y_i)_{i\in[n]}$ as

$$\begin{aligned}\mathcal{P}_{\beta,k} = \big\{&(n-k) \text{ observations are drawn from } \pi_\Sigma(X_{i,\cdot}) \times \mathcal{N}(Y_i, X_{i,\cdot}\beta, \sigma^2),\\ &k \text{ observations are drawn from arbitrary } \mathbf{Q} \,\big\}.\end{aligned} \tag{S8.7}$$

Define $\mathcal{M}(\beta, o) := \bigcup_{0\leq k\leq o} \mathcal{P}_{\beta,k}$, and in this subsection, we focus on deriving a lower bound as

$$\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{\mathbf{R}} \left(\|\hat{\beta} - \beta\|_q^q\right) \gtrsim \sigma^q s \left(\frac{\log(ep/s)}{n}\right)^{q/2} + \sigma^q s^{1-q/2}\frac{o^q}{n^q}, \tag{S8.8}$$

where $\mathbf{R}$ is a joint distribution of $(X_{i\cdot}, Y_i)_{i\in[n]}$.

It is straightforward to apply the uncontaminated result (Raskutti et al., 2011;

Ndaoud, 2019) to get

$$\begin{aligned}&\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{\mathbf{R}}\left(\|\hat{\beta}-\beta\|_q^q\right)\\ \geq &\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{\mathbf{R}\in\mathcal{P}_{\beta,k=0}} \mathbf{E}_{\mathbf{R}}\left(\|\hat{\beta}-\beta\|_q^q\right) \gtrsim \sigma^q s\left(\frac{\log(ep/s)}{n}\right)^{q/2},\end{aligned} \tag{S8.9}$$

for any $q\in[1,2]$, therefore we only need to derive the term $\sigma^q s^{1-q/2}\frac{o^q}{n^q}$.

We first get a lower bound as

$$\begin{aligned}&\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\|\hat{\beta}(Y,X)-\beta\right\|_q^q\\ =&\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{k\leq o} \sup_{\mathbf{R}\in\mathcal{P}_{\beta,k}} \mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\|\hat{\beta}(Y,X)-\beta\right\|_q^q\\ \overset{(i)}{\geq}&\inf_{\hat{\beta}} \mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k^o}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\sum_{j\in[p]}\left|\hat{\beta}_j(Y,X)-\beta_j\right|^q\\ \overset{(ii)}{\geq}&\sum_{j\in[p]}\inf_{\hat{\beta}_j(Y,X)} \mathbf{E}_{\beta,k}\mathbf{E}_{X,Y|\beta,k}\left|\hat{\beta}_j(Y,X)-\beta_j\right|^q\\ \overset{(iii)}{=}&\sum_{j\in[p]}\inf_{\hat{\beta}_j(Y,X)} \mathbf{E}_{Y,X}\mathbf{E}_{\beta,k|Y,X}\left|\hat{\beta}_j(Y,X)-\beta_j\right|^q\\ =&\sum_{j\in[p]}\mathbf{E}_{Y,X}\ \mathbf{E}_{\beta,k|Y,X}\left|\hat{M}_j(Y,X)-\beta_j\right|^q.\end{aligned} \tag{S8.10}$$

Some remarks on (S8.10):

1. In inequality (i), we first construct a distribution $\pi_k := Binomial(n,\epsilon')$ and denote π_k^o the conditional distribution of π_k given $0\leq k\leq o$. Once $\beta(=\beta^{(\ell)})$ and k are generated, we use $\mathbf{Q}(\beta)$ (derived in (S8.3)) to represent the contaminated distribution, and denote by $\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))$ the subset of $\mathcal{P}_{\beta,k}$, given

the contamination $\mathbf{Q} = \mathbf{Q}(\beta)$. It is then clear that $|\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))| = \binom{n}{k}$, and we write $\mathbf{R} \sim U\big(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\big)$ to denote the uniform distribution over the collection $\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))$. Equivalently, each subset of size k is chosen as the outlier set with probability $1/\binom{n}{k}$.

2. In inequality (ii), we write $\mathbf{E}_{\beta,k}$ as an abbreviation of $\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k^o}$. And we write the abbreviation $\mathbf{E}_{X,Y|\beta,k}$ since the joint distribution of (X,Y) is fully specified given β and k:

$$\mathbb{P}\left(X,Y|\beta=\beta^{(\ell)},k\right)=\frac{1}{\binom{n}{k}}\sum_{O\subset[n]:|O|=k}\left\{\prod_{i\in O}Q(X_i.,Y_i)\cdot\prod_{i\notin O}\pi_\Sigma(X_i.)\mathcal{N}\left(Y_i,X_i.\beta^{(\ell)},\sigma^2\right)\right\},$$

as we discussed in inequality (i).

3. In equality (iii), we denote by $\mathbf{E}_{\beta,k|Y,X}$ the conditional expectation based on the probability measure

$$\pi^o_{\beta,k|Y,X}(\beta^{(\ell)},k):=\frac{\mathbb{P}\left(\beta=\beta^{(\ell)},k,X,Y\right)}{\sum_{\ell'\in\{1,2\},0\le k'\le o}\mathbb{P}\left(\beta=\beta^{(\ell')},k=k',X,Y\right)},$$

where

$$\begin{aligned}&\mathbb{P}\left(\beta=\beta^{(\ell)},k,X,Y\right)\\:=&\frac{1}{2}\times\frac{\binom{n}{k}(\epsilon')^k(1-\epsilon')^{n-k}}{\sum_{0\le k'\le o}\binom{n}{k'}(\epsilon')^{k'}(1-\epsilon')^{n-k'}}\\&\times\frac{1}{\binom{n}{k}}\sum_{O\subset[n]:|O|=k}\left\{\prod_{i\in O}Q(X_i.,Y_i)\cdot\prod_{i\notin O}\pi_\Sigma(X_i.)\mathcal{N}\left(Y_i,X_i.\beta^{(\ell)},\sigma^2\right)\right\}.\end{aligned}$$

4. In the last equality, by Theorem 1.1 on page 228 in (Lehmann and Casella, 2006), there exists a Bayes estimator $\hat{M}_j(Y,X)$ achieving the infimum. Additionally, this estimator is better than the zero estimator, leading

$$\begin{aligned}|\hat{M}_j(Y,X)| \leq& \left|\mathbf{E}_{\beta,k|Y,X}(\hat{M}_j(Y,X)-\beta_j)\right| + \left|\mathbf{E}_{\beta,k|Y,X}\beta_j\right| \\ \overset{\text{(Jensen)}}{\leq}& \left(\mathbf{E}_{\beta,k|Y,X}\left|\hat{M}_j(Y,X)-\beta_j\right|^q\right)^{1/q} + \left(\mathbf{E}_{\beta,k|Y,X}|\beta_j|^q\right)^{1/q} \\ \leq& 2\left(\mathbf{E}_{\beta,k|Y,X}|\beta_j|^q\right)^{1/q}.\end{aligned} \tag{S8.11}$$

Therefore, we bridge the ϵ'-contamination model (S8.6) and our frequentist-setting model for any $q \in [1,2]$:

$$\begin{aligned}&\inf_{\hat{\beta}} \mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)}\left(\|\hat{\beta}-\beta\|_q^q\right) \\ =&\inf_{\hat{\beta}} \mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\sum_{j\in[p]}\left|\hat{\beta}_j(Y,X)-\beta_j\right|^q \\ \leq&\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\{\mathbf{1}(k\leq o)\cdot\sum_{j\in[p]}\left|\hat{M}_j(Y,X)-\beta_j\right|^q\right\} \\ &+\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\{\mathbf{1}(k>o)\cdot\sum_{j\in[p]}\left|\hat{M}_j(Y,X)-\beta_j\right|^q\right\} \\ \overset{\text{(Jensen)}}{\leq}&\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k^o}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\{\sum_{j\in[p]}\left|\hat{M}_j(Y,X)-\beta_j\right|^q\right\} \\ &+2^{q-1}\sum_{j\in[p]}\mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{k\sim\pi_k}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\beta,k}(\mathbf{Q}(\beta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\{\mathbf{1}(k>o)\cdot\left(|\hat{M}_j(Y,X)|^q+|\beta_j|^q\right)\right\} \\ \leq&\inf_{\hat{\beta}}\sup_{\beta:\|\beta\|_0\leq s}\sup_{\mathbf{R}\in\mathcal{M}(\beta,o)}\mathbf{E}_{(X,Y)\sim\mathbf{R}}\left\|\hat{\beta}(Y,X)-\beta\right\|_q^q + 15s\lambda^q\cdot\pi_k(k>o),\end{aligned} \tag{S8.12}$$

where the last inequality follows from (S8.10), (S8.11), and also the fact

$$\mathbf{E}_{\beta,k|Y,X}\left(|\beta_j|^q\right)\begin{cases}\leq \lambda^q & \text{if } j \leq 3s/2,\\ = 0 & \text{if } 3s/2 < j \leq p,\end{cases}$$

leading by the construction of π_β ((S8.1) and (S8.5)).

By Bernstein's inequality (Appendix D.4 in Mohri et al. (2018)), we have

$$\pi_k(k > o) \leq \exp\left(-\frac{\frac{1}{n}(o-n\epsilon')^2}{2\epsilon' + \frac{2}{3}\frac{o-n\epsilon'}{n}}\right) \leq \exp(-9o/16),$$

where the last inequality follows from $\epsilon' \leq o/(4n)$. Therefore, combining (S8.6) and (S8.12), for any $q \in [1,2]$, we have the lower bound

$$\begin{aligned}&\inf_{\hat{\beta}} \sup_{\beta:\|\beta\|_0\leq s} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \left\|\hat{\beta}(Y,X)-\beta\right\|_q^q\\ \geq& s\lambda^q\left(\frac{1}{8} - 15\exp(-9o/16)\right)\\ \geq& \frac{(8C_{2s})^{-q/2}}{80}\cdot\sigma^q s^{1-q/2}\frac{o^q}{n^q},\end{aligned} \tag{S8.13}$$

where the last inequality follows from the assumption $o \geq 9$. Hence, we complete the proof of Theorem 4.

S8.3 Selection lower bound in Theorem 5

The proof of Theorem 5 is quite similar to that of Theorem 4. Define

$$\mathcal{B}(s,a) := \left\{ \beta \in \mathbb{R}^p : \ \|\beta\|_0 \le s, \min_{i:\beta_i \neq 0} |\beta_i| \ge a \right\}.$$

From Wainwright (2007) and Fano's inequality (Tsybakov, 2009), we have

$$\inf_{\hat{S}} \sup_{\beta \in \mathcal{B}\left(s,\ c_1\sigma\sqrt{(\log(ep/s))/n}\right)} \sup_{\mathbf{R} \in \mathcal{P}_{\beta,k=0}} \mathbf{E}_{\mathbf{R}} \left\{ \left| \hat{S}(X,Y) \triangle \mathrm{supp}(\beta) \right| \right\} \ge c_2 s,$$

where $\inf_{\hat{S}}$ denotes the infimum over all support set estimation $\hat{S}$ based on (X,Y), and $c_1, c_2 > 0$ are two absolute constants. Therefore, we only need to prove the selection lower bound based on the parameter subspace $\mathcal{B}(s,\lambda)$ with $\lambda = \sqrt{\frac{1}{8C_{2s}}} \cdot \frac{\sigma}{\sqrt{s}} \cdot \frac{o}{n}$. We introduce the decoder $\eta = \eta(\beta) \in \{0,1\}^p$, with each $\eta_i = \{\eta(\beta)\}_i = \mathbf{1}(\beta_i \neq 0)$. Denote the selector $\hat{\eta}(X,Y) \in \{0,1\}^p$ as the estimator of $\eta(\beta)$. By using the same technique in Section S8.1, we have

$$\begin{aligned}
&\inf_{\hat{\eta}} \mathbf{E}_{\beta \sim \pi_\beta} \mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)} \left(\|\hat{\eta}(X,Y) \ - \eta(\beta)\|_2^2 \right) \\
=& \inf_{\hat{S}} \mathbf{E}_{\beta \sim \pi_\beta} \mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)} \left(\left| \hat{S}(X,Y) \triangle \mathrm{supp}(\beta) \right| \right) \\
\ge& \frac{s}{2} \cdot \inf_{\hat{\beta}} \ \frac{1}{2} \cdot \Big\{ [(1-\epsilon')\mathbf{P}_1 + \epsilon'\mathbf{Q}_1] \left(\left| \hat{S}(X,Y) \triangle \mathrm{supp}(\beta^{(1)}) \right| \ge \frac{s}{2} \right) \\
&\qquad + [(1-\epsilon')\mathbf{P}_2 + \epsilon'\mathbf{Q}_2] \left(\left| \hat{S}(X,Y) \triangle \mathrm{supp}(\beta^{(2)}) \right| \ge \frac{s}{2} \right) \Big\} \\
\ge& \frac{s}{4}.
\end{aligned} \tag{S8.14}$$

We next analyze the selection lower bound in our setting. Here we rewrite the prior π_β as a prior on η, i.e., $\pi_\eta(\eta = \eta^{(\ell)}) = 1/2, \quad \ell = 1, 2$, where

$$\begin{aligned} \eta^{(1)} &:= (\underbrace{1, \cdots, 1}_{s}, \underbrace{0, \cdots, 0}_{p-s})^\top \in \mathbb{R}^p, \\ \eta^{(2)} &:= (\underbrace{0, \cdots, 0}_{s/2}, \underbrace{1, \cdots, 1}_{s}, \underbrace{0, \cdots, 0}_{p-3s/2})^\top \in \mathbb{R}^p. \end{aligned} \tag{S8.15}$$

Thus $\beta^{(\ell)} = \lambda\eta^{(\ell)}$. And we get two results similar to (S8.10) and (S8.12):

$$\begin{aligned} &\inf_{\hat{\eta}} \sup_{\beta\in\mathcal{B}(s,\lambda)} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \|\hat{\eta}(Y,X) - \eta(\beta)\|_2^2 \\ \geq &\inf_{\hat{\eta}} \ \mathbf{E}_{\eta\sim\pi_\eta}\mathbf{E}_{k\sim\pi_k^o}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\lambda\eta,k}(\mathbf{Q}(\lambda\eta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}} \sum_{j\in[p]} \left(\hat{\eta}_j(Y,X) - \eta_j\right)^2 \\ \geq &\sum_{j\in[p]} \mathbf{E}_{Y,X} \ \mathbf{E}_{\eta,k|Y,X} \left(\hat{T}_j(Y,X) - \eta_j\right)^2, \end{aligned} \tag{S8.16}$$

and

$$\begin{aligned} &\inf_{\hat{\eta}} \mathbf{E}_{\beta\sim\pi_\beta}\mathbf{E}_{Y,X|\epsilon',\beta,\mathbf{Q}(\beta)} \left(\|\hat{\eta} - \eta\|_2^2\right) \\ \leq &\mathbf{E}_{\eta\sim\pi_\eta}\mathbf{E}_{k\sim\pi_k^o}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\lambda\eta,k}(\mathbf{Q}(\lambda\eta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}} \left\{\sum_{j\in[p]} \left(\hat{T}_j(Y,X) - \eta_j\right)^2\right\} \\ &+ 2s\cdot\mathbf{E}_{\eta\sim\pi_\eta}\mathbf{E}_{k\sim\pi_k}\mathbf{E}_{\mathbf{R}\sim U\left(\mathcal{P}_{\lambda\eta,k}(\mathbf{Q}(\lambda\eta))\right)}\mathbf{E}_{(X,Y)\sim\mathbf{R}} \left\{\mathbf{1}(k > o)\right\} \\ \leq &\inf_{\hat{\eta}} \sup_{\beta\in\mathcal{B}(s,\lambda)} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \|\hat{\eta}(Y,X) - \eta(\beta)\|_2^2 \ + \ 2s\cdot\pi_k(k > o), \end{aligned} \tag{S8.17}$$

where $\hat{T}_j(Y,X) := \mathbf{E}_{\eta,k|Y,X}(\eta_j) \in [0,1]$. Therefore,

$$\begin{aligned}
&\inf_{\hat{S}} \sup_{\beta\in\mathcal{B}(s,\lambda)} \sup_{\mathbf{R}\in\mathcal{M}(\beta,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \left|\hat{S}(X,Y) \triangle \mathrm{supp}(\beta)\right| \\
\geq &\frac{s}{4} - 2s\cdot\exp(-9o/16) \\
\geq &\frac{s}{5},
\end{aligned} \tag{S8.18}$$

where the last inequality follows from $o \geq 8$, and thus we complete the proof of Theorem 5.

S9 Technical Lemmas

Lemma 1 (Support error control and element-wise error control) *Assume $\xi_i \sim SG(0,\sigma^2)$ independently for $i \in [n]$. Recall $S^* = \mathrm{supp}(\beta^*)$. Then, under Assumptions 1 and 2, we have*

$$\mathbf{P}\left(\frac{1}{n}\left\|X_{\cdot S^*}^\top \xi\right\|_2 \leq \sigma\sqrt{\frac{4Ms + 6M\log(1/\varrho)}{n}}\right) \geq 1 - 2p^{-2s} - \varrho. \tag{S9.1}$$

Additionally, we have

$$\mathbf{P}\left(\sup_{i\in[p]}\left|\frac{1}{n}X_{\cdot i}^\top \xi\right| \leq 4\sigma\sqrt{\frac{M\log p}{n}}\right) > 1 - 2p^{-2s} - 2p^{-3}, \tag{S9.2}$$

and

$$\mathbf{P}\left(\sup_{k\in[n]}\left|\frac{1}{\sqrt{n}}\xi_k\right| \leq 3\sigma\sqrt{\frac{\log n}{n}}\right) > 1 - 2n^{-3}. \tag{S9.3}$$

Proof 1 (Proof of Lemma 1) *In this lemma, Assumption 2 is used through Proposition 1 with $C_1 = 1$, namely in the concrete form*

$$n \geq 30M^2 C_\Sigma \max(s \log p, o \log n).$$

Define

$$\mathcal{E}_{L1} = \left\{ \sup_{S \subset [p]:\ |S| \leq s} \max_{1 \leq k \leq |S|} \Lambda_k \left(\frac{1}{n} X_{\cdot S}^\top X_{\cdot S} \right) \leq 2M \right\} \tag{S9.4}$$

Then by Proposition 1, $\mathbf{P}(\mathcal{E}_{L1}) \geq 1 - 2\exp(-2s \log p)$. The next two parts constitute the main proof.

For Equation (S9.1) *By using Theorem 2.1 of Hsu et al. (2012) and Assumption 1, for a given X, we have*

$$\mathbf{P}\left\{ \frac{\left\| X_{\cdot S^*}^\top \xi \right\|_2^2}{n\sigma^2} \geq tr\left(\frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right) + 2\left\| \frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right\|_F \sqrt{t} + 2\Lambda_{\max}\left(\frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right) t \ \middle| X \right\} \leq e^{-t}.$$

Combining with $|S^*| = s$, under event $\mathcal{E}_{L1}$,

$$\begin{aligned} tr\left(\frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right) &= \sum_{k=1}^{s} \Lambda_k \left(\frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right) \leq 2Ms, \\ \left\| \frac{X_{\cdot S^*}^\top X_{\cdot S^*}}{n} \right\|_F^2 &= \sum_{k=1}^{s} \Lambda_k^2 \left(\frac{X_{\cdot S}^\top X_{\cdot S}}{n} \right) \leq 4M^2 s. \end{aligned}$$

By taking $t = \log(1/\varrho)$ *(where* $\varrho \in (0,1)$*), we obtain*

$$\begin{aligned}
&\mathbf{P}\left(\frac{1}{n\sigma^2}\left\|X_{\cdot S^*}^\top \xi\right\|_2^2 \geq 4Ms + 6M\log(1/\varrho)\right) \\
\leq &\mathbf{P}(\mathcal{E}_{L1}^c) + \mathbf{E}_X\left\{\mathbf{1}(\mathcal{E}_{L1}) \cdot \mathbf{P}\left(\frac{1}{n\sigma^2}\left\|X_{\cdot S^*}^\top \xi\right\|_2^2 \geq 4Ms + 6M\log(1/\varrho) \;\middle|\; X\right)\right\} \\
\leq &2p^{-2s} + \varrho,
\end{aligned}$$

which completes the proof of equation (S9.1)*.*

For Equation (S9.2) **and** (S9.3) *For a given* X*, it is straightforward that* $\Xi_i = \frac{1}{n} X_{\cdot i}^\top \xi$ *is sub-Gaussian random variable with sub-Gaussian parameter* $\sigma\|X_{\cdot i}\|_2/n$ *(Wainwright, 2019), leading that*

$$\begin{aligned}
&\mathbf{P}\left(\sup_{i\in[p]}\left|\frac{1}{n}X_{\cdot i}^\top \xi\right| > 4\sigma\sqrt{\frac{M\log p}{n}}\right) \\
\leq &\mathbf{P}(\mathcal{E}_{L1}^c) + \mathbf{E}_X\left\{\mathbf{1}(\mathcal{E}_{L1}) \cdot \mathbf{P}\left(\sup_{i\in[p]}\left|\frac{1}{n}X_{\cdot i}^\top \xi\right| > 4\sigma\sqrt{\frac{M\log p}{n}} \;\middle|\; X\right)\right\} \\
\leq &2p^{-2s} + p \cdot 2e^{-4\log p} \;=\; 2p^{-2s} + 2p^{-3},
\end{aligned}$$

where the second inequality applies the union bound and the fact $\max_{i\in[p]}\|X_{\cdot i}\|_2^2 \leq 2Mn$ *under the event* $\mathcal{E}_{L1}$*. A similar union bound yields that*

$$\mathbf{P}\left(\sup_{k\in[n]}\left|\frac{1}{\sqrt{n}}\xi_k\right| > 3\sigma\sqrt{\frac{\log n}{n}}\right) \leq 2\exp\left(-3\log n\right).$$

Therefore, we complete the proof of Lemma 1.

Next, we control the error related to the oracle estimator. Recall that

$$\begin{aligned}
\beta^{\dagger}_{S^*} &= \beta^{*}_{S^*} + (X^{\top}_{-O^*,S^*}X_{-O^*,S^*})^{-1}X^{\top}_{-O^*,S^*}\xi_{-O^*}, \quad \beta^{\dagger}_{-S^*} = \mathbf{0}, \\
\theta^{\dagger}_{O^*} &= \theta^{*}_{O^*} + \frac{1}{\sqrt{n}}\Big\{\xi_{O^*} - X_{O^*,S^*}(X^{\top}_{-O^*,S^*}X_{-O^*,S^*})^{-1}X^{\top}_{-O^*,S^*}\xi_{-O^*}\Big\}, \quad \theta^{\dagger}_{-O^*} = \mathbf{0}, \\
\Xi^{\dagger}_{S^*} &= \mathbf{0}, \quad \xi^{\dagger}_{O^*} = \mathbf{0}, \\
\Xi^{\dagger}_{-S^*} &= \frac{1}{n}X^{\top}_{-O^*,-S^*}\Big\{I_{n-o} - P_{-O^*,S^*}\Big\}\xi_{-O^*}, \\
\xi^{\dagger}_{-O^*} &= \Big\{I_{n-o} - P_{-O^*,S^*}\Big\}\xi_{-O^*},
\end{aligned}$$

where we define

$$P_{-O^*,S^*} := X_{-O^*,S^*}(X^{\top}_{-O^*,S^*}X_{-O^*,S^*})^{-1}X^{\top}_{-O^*,S^*} \in \mathbb{R}^{(n-o)\times(n-o)}.$$

Lemma 2 (Oracle error control) *Assume that $\xi_i \sim SG(0,\sigma^2)$ independently for $i \in [n]$. Under Assumptions 1 and 2, we have*

$$\begin{aligned}
\mathbf{P}\left(\sup_{i\in(S^*)^c}|\Xi^{\dagger}_i| < 4\sigma\sqrt{\frac{M\log p}{n}}\right) &> 1 - 2p^{-3} - 2p^{-2s}, \\
\mathbf{P}\left(\sup_{i\in(O^*)^c}\left|\frac{\xi^{\dagger}_i}{\sqrt{n}}\right| < 3\sigma\sqrt{\frac{\log n}{n}}\right) &> 1 - 2n^{-3}.
\end{aligned} \tag{S9.5}$$

Additionally, under signal condition (S6.3)*, there are element-wise estimation bounds*

on the oracle estimators $\beta^\dagger$ and $\theta^\dagger$:

$$\begin{aligned}\mathbf{P}\left(\sup_{i\in S^*}\left|\beta_i^\dagger-\beta_i^*\right|<5\sigma\sqrt{\frac{M\log p}{n}}\right)&\geq 1-4p^{-2s}-2p^{-3},\\ \mathbf{P}\left(\sup_{i\in O^*}\left|\theta_i^\dagger-\theta_i^*\right|<4\sigma\sqrt{\frac{\log n}{n}}\right)&\geq 1-4p^{-2s}-2n^{-3}.\end{aligned}\tag{S9.6}$$

Proof 2 (Proof of Lemma 2) *In this lemma, Assumption 2 is used in the concrete form*

$$n\geq C_{\text{oracle}}\{s\log p+o\log n\},\qquad C_{\text{oracle}}\geq\max\left\{30M^2C_\Sigma,\ 6MC_M^2\right\}.$$

This condition covers the application of Proposition 1 with $C_1=1$, and the minimum-eigenvalue bound in Lemma 3. Then, two parts constitute the proof.

For Equation (S9.5) For every $i\in(O^*)^c$, it is straightforward that $\xi_i^\dagger=e_i^\top(I_{n-o}-P_{-O^*,S^*})\xi_{-O^*}$ satisfies

$$\mathbf{E}_{\xi,X}\left(e^{\lambda\xi_i^\dagger}\right)\leq\mathbf{E}_X\exp\left(\frac{\lambda^2\sigma^2}{2}\left\|e_i^\top\left(I_{n-o}-P_{-O^*,S^*}\right)\right\|_2^2\right)\leq\exp\left(\frac{\lambda^2\sigma^2}{2}\right)\tag{S9.7}$$

for all $\lambda\in\mathbb{R}$, where $e_i\in\mathbb{R}^{(n-o)\times 1}$ is a vector with 1 at the i-th position and 0 elsewhere. Therefore, $\xi_i^\dagger$ is a sub-Gaussian variable with sub-Gaussian parameter σ (Wainwright, 2019), which leads that

$$\mathbf{P}_{\xi,X}\left(\sup_{i\in(O^*)^c}\left|\xi_i^\dagger\right|\geq 3\sigma\sqrt{\log n}\right)\leq\sum_{i\in(O^*)^c}2\exp\left(-\frac{9\log n}{2}\right)\ \leq\ 2\exp\left(-3\log n\right).$$

Similarly, with a fixed design $X \in \mathbb{R}^{n\times p}$, for every $i \in (S^*)^c$,

$$\mathbf{E}_{\xi|X}\left(e^{\lambda \Xi_i^\dagger}\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2n^2}\left\|(I_{n-o} - P_{-O^*,S^*})X_{-O^*,i}\right\|_2^2\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2n^2}\left\|X_{\cdot,i}\right\|_2^2\right),$$

by the event $\mathcal{E}_{L1}$ (defined in (S9.4)), we conclude that

$$\begin{aligned}\mathbf{P}\left(\sup_{i\in(S^*)^c}|\Xi_i^\dagger| \geq 4\sigma\sqrt{\frac{M\log p}{n}}\right) &\leq \mathbf{P}\left(\sup_{i\in(S^*)^c}|\Xi_i^\dagger| \geq 4\sigma\sqrt{\frac{M\log p}{n}}, X \in \mathcal{E}_{L1}\right) + \mathbf{P}(\mathcal{E}_{L1}^c)\\ &\leq 2p^{-3} + 2p^{-2s},\end{aligned}$$

therefore we complete the proof of (S9.5).

For Equation (S9.6) Similar to (S9.7), with a fixed design $X \in \mathbb{R}^{n\times p}$, for every $i \in S^*$ and $\lambda \in \mathbb{R}$ we have

$$\begin{aligned}\mathbf{E}_{\xi|X}\left(e^{\lambda(\beta_i^\dagger - \beta_i^*)}\right) &\overset{(i)}{\leq} \exp\left(\frac{\lambda^2\sigma^2}{2}\left\|e_i^\top(X_{-O^*,S^*}^\top X_{-O^*,S^*})^{-1}X_{-O^*,S^*}^\top\right\|_2^2\right)\\ &\leq \exp\left(\frac{\lambda^2\sigma^2}{2}\left\|(X_{-O^*,S^*}^\top X_{-O^*,S^*})^{-1}X_{-O^*,S^*}^\top\right\|_2^2\right)\\ &\overset{(ii)}{=} \exp\left(\frac{\lambda^2\sigma^2}{2}\left\|(X_{-O^*,S^*}^\top X_{-O^*,S^*})^{-1}\right\|_2\right)\\ &= \exp\left(\frac{\lambda^2\sigma^2}{2}\frac{1}{\Lambda_{\min}(X_{-O^*,S^*}^\top X_{-O^*,S^*})}\right),\end{aligned}$$

where inequality (i) follows from the relationship (S6.1), inequality (ii) follows from $\|A\|_2^2 = \|AA^\top\|_2$.

Define the event

$$\mathcal{E}_{L2} := \left\{ \|X_{O^*,S^*}\|_2 \le C_M \sqrt{s \log p + o \log n}, \quad \Lambda_{\min}\left(X^\top_{-O^*,S^*} X_{-O^*,S^*}\right) \ge \frac{n}{3M} \right\}.$$

Then, by Proposition 1 and Lemma 3, we have $\mathbf{P}(\mathcal{E}_{L2}) \ge 1 - 4p^{-2s}$. Therefore, by the probability union bound, we have

$$\begin{aligned}
\mathbf{P}\left(\sup_{i \in S^*} \left| \beta_i^\dagger - \beta_i^* \right| \ge 5\sigma \sqrt{\frac{M \log p}{n}} \right) &\le \mathbf{P}\left(\sup_{i \in S^*} \left| \beta_i^\dagger - \beta_i^* \right| \ge 5\sigma \sqrt{\frac{M \log p}{n}} \ \middle| \mathcal{E}_{L2} \right) \mathbf{P}(\mathcal{E}_{L2}) + \mathbf{P}(\mathcal{E}_{L2}^c) \\
&\le 2p^{-3} + 4p^{-2s}.
\end{aligned}$$

Additionally, with a fixed design $X \in \mathbb{R}^{n \times p}$, for every $i \in O^*$ and $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbf{E}_{\xi|X}\left(e^{\lambda(\theta_i^\dagger - \theta_i^*)} \right) &= \left\{ \mathbf{E}_\xi \ e^{\lambda \xi_i / \sqrt{n}} \right\} \cdot \left\{ \mathbf{E}_{\xi|X} \ \exp\left(-\frac{\lambda}{\sqrt{n}} e_i^\top X_{O^*,S^*} (X^\top_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^\top_{-O^*,S^*} \xi_{-O^*} \right) \right\} \\
&\le \exp\left(\frac{\lambda^2 \sigma^2}{2n} \right) \cdot \exp\left(\frac{\lambda^2 \sigma^2}{2n} \left\| e_i^\top X_{O^*,S^*} (X^\top_{-O^*,S^*} X_{-O^*,S^*})^{-1} X^\top_{-O^*,S^*} \right\|_2^2 \right) \\
&\le \exp\left\{ \frac{\lambda^2 \sigma^2}{2n} \left(1 + \|X_{O^*,S^*}\|_2^2 \left\| (X^\top_{-O^*,S^*} X_{-O^*,S^*})^{-1} \right\|_2 \right) \right\} \\
&= \exp\left\{ \frac{\lambda^2 \sigma^2}{2n} \left(1 + \frac{\|X_{O^*,S^*}\|_2^2}{\Lambda_{\min}\left(X^\top_{-O^*,S^*} X_{-O^*,S^*} \right)} \right) \right\}
\end{aligned}$$

Therefore, under the assumption $n \ge 3MC_M^2(s \log p + o \log n)$, we have

$$\begin{aligned}
\mathbf{P}\left(\sup_{i \in O^*} \left| \theta_i^\dagger - \theta_i^* \right| \ge 4\sigma \sqrt{\frac{\log n}{n}} \right) &\le \mathbf{P}\left(\sup_{i \in O^*} \left| \theta_i^\dagger - \theta_i^* \right| \ge 4\sigma \sqrt{\frac{\log n}{n}} \ , \ X \in \mathcal{E}_{L2} \right) + \mathbf{P}\left(\mathcal{E}_{L2}^c \right) \\
&\le 2n^{-3} + 4p^{-2s}.
\end{aligned}$$

Therefore, we complete the proof of (S9.6), and hence the proof of Lemma 2 is completed.

Lemma 3 (Restricted minimum eigenvalue control) *Recall* $S^* = \mathrm{supp}(\beta^*)$ *and* $O^* = \mathrm{supp}(\theta^*)$. *Under Assumptions 1 and 2,*

$$\mathbf{P}\left(\Lambda_{\min}\left(\frac{X_{-O^*,S^*}^\top X_{-O^*,S^*}}{n}\right) \geq \frac{1}{3M}\right) \geq 1 - 4p^{-2s}.$$

Proof 3 (Proof of Lemma 3) *In this proof, Assumption 2 is used in the concrete form*

$$n \geq \max\left\{30M^2C_\Sigma,\, 6MC_M^2\right\} \cdot (s\log p + o\log n).$$

This condition covers both the application of Proposition 1 with $C_1 = 1$ *and the bound*

$$C_M^2 \frac{s\log p + o\log n}{n} \leq \frac{1}{6M}. \tag{S9.8}$$

By Weyl's inequality and the decomposition $X_{\cdot,S^*}^\top X_{\cdot,S^*} = X_{O^*,S^*}^\top X_{O^*,S^*} + X_{-O^*,S^*}^\top X_{-O^*,S^*}$, *with a probability greater than* $1 - 4p^{-2s}$, *we have*

$$\begin{aligned}
\Lambda_{\min}\left(\frac{X_{-O^*,S^*}^\top X_{-O^*,S^*}}{n}\right) &\geq \Lambda_{\min}\left(\frac{X_{\cdot,S^*}^\top X_{\cdot,S^*}}{n}\right) - \Lambda_{\max}\left(\frac{X_{O^*,S^*}^\top X_{O^*,S^*}}{n}\right) \\
&\geq \frac{1}{2M} - \frac{\|X_{O^*,S^*}\|_2^2}{n} \\
&\geq \frac{1}{2M} - C_M^2 \frac{s\log p + o\log n}{n} \\
&\geq \frac{1}{3M},
\end{aligned}$$

where the second inequality follows from the restricted isometry in Proposition 1, the third inequality follows from the restricted incoherence in Proposition 1, and the last inequality follows from the condition (S9.8)*. This completes the proof of Lemma 3.*

S10 Additional Details about the Extension to GLM

In the Discussion section, we extend the two-stage AC-IHT algorithm to the setting of generalized linear models (GLMs). Compared to the linear model case, only the gradient descent step requires modification. Therefore, we begin by analyzing the decomposition of H_β^{t+1} and H_θ^{t+1}.

S10.1 Preliminary

Let $X^{(i)} \in \mathbb{R}^{p\times 1}$ be the transformation of the ith row of X. We have

$$\begin{aligned}
H_\beta^{t+1} :=& \beta^t - \frac{\eta}{n}\sum_{i=1}^n X^{(i)}\left(b'(\zeta_i^t) - Y_i\right) \\
=& \beta^* - (\beta^* - \beta^t) - \frac{\eta}{n}\sum_{i=1}^n X^{(i)}\Big\{\left(b'(\zeta_i^t) - b'(\zeta_i^*)\right) - (Y_i - b'(\zeta_i^*))\Big\} \\
=& \beta^* + \left(\frac{\eta}{n}\sum_{i=1}^n b''(\tilde{\zeta}_i^t)X^{(i)}(X^{(i)})^\top - \mathbf{I}_p\right)\left(\beta^* - \beta^t\right) \\
&+ \frac{\eta}{\sqrt{n}}\left(\sum_{i=1}^n b''(\tilde{\zeta}_i^t)X^{(i)}e_i^\top\right)(\theta^* - \theta^t) + \frac{\eta}{n}\sum_{i=1}^n X^{(i)}\left(Y_i - b'(\zeta_i^*)\right) \\
=& \beta^* + \tilde{\Phi}^t\left(\beta^* - \beta^t\right) + \frac{\eta}{\sqrt{n}}(\tilde{X}^t)^\top\left(\theta^* - \theta^t\right) + \eta\tilde{\Xi},
\end{aligned} \tag{S10.1}$$

$$
\begin{aligned}
H_\theta^{t+1} :=& \theta^t - \frac{\eta}{\sqrt{n}} \sum_{i=1}^n e_i \left(b'(\zeta_i^t) - Y_i \right) \\
=& \theta^* - (\theta^* - \theta^t) - \frac{\eta}{\sqrt{n}} \sum_{i=1}^n e_i \Big\{ \left(b'(\zeta_i^t) - b'(\zeta_i^*) \right) - \left(Y_i - b'(\zeta_i^*) \right) \Big\} \\
=& \theta^* + \frac{\eta}{\sqrt{n}} \left(\sum_{i=1}^n b''(\tilde{\zeta}_i^t) e_i (X^{(i)})^\top \right) \left(\beta^* - \beta^t \right) + \left(\eta \sum_{i=1}^n b''(\tilde{\zeta}_i^t) e_i e_i^\top - \mathbf{I}_n \right) (\theta^* - \theta^t) \\
&+ \frac{\eta}{\sqrt{n}} \sum_{i=1}^n e_i \left(Y_i - b'(\zeta_i^*) \right) \\
=& \theta^* + \frac{\eta}{\sqrt{n}} \tilde{X}^t \left(\beta^* - \beta^t \right) + \tilde{D}^t \left(\theta^* - \theta^t \right) + \frac{\eta}{\sqrt{n}} \tilde{\xi},
\end{aligned}
\tag{S10.2}
$$

where $\zeta_i^t := (X^{(i)})^\top \beta^t + \sqrt{n} e_i^\top \theta^t$, $\tilde{\zeta}_i^t = \lambda^t \zeta_i^t + (1 - \lambda^t) \zeta_i^*$ for some $\lambda^t \in (0, 1)$, and $\tilde{\Phi}^t := \frac{\eta}{n} \sum_{i=1}^n b''(\tilde{\zeta}_i^t) X^{(i)} (X^{(i)})^\top - I_p \in \mathbb{R}^{p \times p}$, $\tilde{X}^t := \sum_{i=1}^n b''(\tilde{\zeta}_i^t) e_i (X^{(i)})^\top \in \mathbb{R}^{n \times p}$, $\tilde{\Xi} := \frac{1}{n} \sum_{i=1}^n X^{(i)} \left(Y_i - b'(\zeta_i^*) \right) \in \mathbb{R}^p$, $e_i \in \mathbb{R}^{n \times 1}$ be the i-th standard basis vector, $\tilde{D}^t := \eta \mathrm{diag}(b''(\tilde{\zeta}_1^t), \ldots, b''(\tilde{\zeta}_n^t)) - I_n = \eta \sum_{i=1}^n b''(\tilde{\zeta}_i^t) e_i e_i^\top - I_n \in \mathbb{R}^{n \times n}$ and $\tilde{\xi} := \sum_{i=1}^n e_i \left(Y_i - b'(\zeta_i^*) \right)$. The decomposition of H_β^{t+1} and H_θ^{t+1} reveals that the structure of H_β^{t+1} is similar to (S3.9) in the proofs of Theorems 1 and 2.

S10.2 Useful lemmas

The following lemmas establish that $\tilde{X}^t$, $\tilde{\Phi}^t$, $\tilde{\Xi}$, and $\tilde{\xi}$ retain the same properties as those in the proofs of Theorems 1 and 2.

Lemma 4 (Proposition 1, GLM version) *Under Assumptions 1, 2, and 5, for*

any fixed constant $C_1 > 0$, with Assumption 2 imposed in the concrete form

$$n \geq 30M^2 C_1 C_\Sigma \max(s \log p, o \log n),$$

the following properties hold with probability greater than $1 - 4\exp(-2C_1 s \log p)$:

1. ***(Restricted isometry)*** *For every index set $S \subset [p]$ with $|S| \leq C_1 s$, the sample covariance matrix satisfies:*

$$\frac{L}{2M}\|u\|_2^2 \;\leq\; u^\top \left(\frac{1}{n}\sum_{i=1}^n b''(\tilde{\zeta}_i^t) X^{(i)} (X^{(i)})^\top \right)_{S,S} u \;\leq\; 2MU\|u\|_2^2, \quad \textit{for each } u \in \mathbb{R}^{|S|}. \tag{S10.3}$$

2. ***(Restricted incoherence)*** *There exists a constant $C_M > 0$ depending only on M such that:*

$$\sup_{S \subset [p]:\ |S| \leq C_1 s} \; \sup_{O \subset [n]:\ |O| \leq C_1 o} \left\| \tilde{X}^t_{O,S} \right\|_2 \leq U\sqrt{C_1} C_M \sqrt{s \log p + o \log n}. \tag{S10.4}$$

Proof 4 (Proof of Lemma 4) *Under Assumption 5, for every index set $S \subset [p]$ with $|S| \leq C_1 s$, we have*

$$\frac{L}{n}\sum_{i=1}^n u^\top \left(X^{(i)} (X^{(i)})^\top \right)_{S,S} u \leq \frac{1}{n}\sum_{i=1}^n b''(\tilde{\zeta}_i^t)\; u^\top \left(X^{(i)} (X^{(i)})^\top \right)_{S,S} u \leq \frac{U}{n}\sum_{i=1}^n u^\top \left(X^{(i)} (X^{(i)})^\top \right)_{S,S} u.$$

Therefore, following (S2.3)*, with a probability greater than $1 - 4p^{-2C_1 s}$ we get*

$$\left\| \frac{1}{n}\sum_{i=1}^n b''(\tilde{\zeta}_i^t) \left(X^{(i)} (X^{(i)})^\top \right)_{S,S} \right\|_2 \leq U \left(\|\Sigma\|_2 + \frac{1}{2M} \right) \leq 2MU.$$

A similar technique shows that

$$\Lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} b''(\tilde{\zeta}_i^t)\left(X^{(i)}(X^{(i)})^\top\right)_{S,S}\right) \geq L\left(\Lambda_{\min}(\Sigma) - \frac{1}{2M}\right) \geq \frac{L}{2M},$$

therefore we complete the proof of (S10.3).

By the definition of $\tilde{X}^t$*, for every index set* $S \subset [p]$ *with* $|S| \leq C_1 s$ *and set* $O \subset [n]$ *with* $|O| \leq C_1 o$*, we have*

$$\begin{aligned}\left\|\tilde{X}^t_{O,S}\right\|_2^2 = \left\|(\tilde{X}^t_{O,S})^\top \tilde{X}^t_{O,S}\right\|_2 &= \left\|\sum_{i\in O}(b''(\tilde{\zeta}_i^t))^2 X_S^{(i)}(X_S^{(i)})^\top\right\|_2 \\ &\overset{(i)}{\leq} U^2\left\|X_{O,S}^\top X_{O,S}\right\|_2 \\ &\leq U^2 C_1 C_M^2 \cdot (s\log p + o\log n),\end{aligned}$$

where inequality (i) follows from Assumption 5, and the last inequality follows from the restricted incoherence result in Proposition 1, which holds with a probability greater than $1 - 4p^{-2C_1 s}$*. Therefore, we complete the proof of Lemma 4.*

Lemma 5 *Assume that GLM (5.1) holds with* $\zeta_i^* = (X^{(i)})^\top \beta^* + \sqrt{n} e_i^\top \theta^*$*, for every* $i \in [n]$*. Then based on Assumption 5,* $\tilde{\xi}_i := Y_i - b'(\zeta_i^*)$ *is sub-Gaussian with zero mean and sub-Gaussian parameter* $\sigma = \sqrt{aU}$*, that is,*

$$\mathbf{P}\left(|Y_i - b'(\zeta_i^*)| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2aU}\right), \text{ for all } t \geq 0. \tag{S10.5}$$

Proof 5 (Proof of Lemma 5) *Note that for GLM (5.1), we have* $\mathbf{E}_{Y_i|X^{(i)}}(Y_i) =$

$b'(\zeta_i^*)$. *From Theorem 5.10 of Lehmann and Casella (2006), we also have*

$$\mathbf{E}_{Y_i|X^{(i)}}\left(\exp(\lambda Y_i)\right) = \exp\left(\frac{b(\zeta_i^* + \lambda a) - b(\zeta_i^*)}{a}\right),\ \forall \lambda \in \mathbb{R}, \tag{S10.6}$$

which leads to

$$\begin{aligned}\mathbf{E}_{Y_i|X^{(i)}}\left(e^{\lambda(Y_i - b'(\zeta_i^*))}\right) &= \exp\left(\frac{b(\zeta_i^* + \lambda a) - b(\zeta_i^*) - \lambda a \cdot b'(\zeta_i^*)}{a}\right) \\ &\overset{(i)}{=} \exp\left(\frac{\lambda^2 a b''(\tilde{\zeta}_i)}{2}\right) \\ &\leq \exp\left(\frac{\lambda^2 a U}{2}\right),\ \forall \lambda \in \mathbb{R},\end{aligned} \tag{S10.7}$$

where in equality (i), $\tilde{\zeta}_i$ *is between* ζ_i *and* $\zeta_i + \lambda a$ *based on Taylor's Theorem, and the last inequality follows from Assumption 5. Hence, we prove that* $Y_i - b'(\zeta_i^*)$ *is sub-Gaussian with zero mean and sub-Gaussian parameter* $\sigma = \sqrt{aU}$*, and it satisfies the concentration inequality (see, e.g., Wainwright (2019)):*

$$\mathbf{P}\left(|Y_i - b'(\zeta_i^*)| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2aU}\right),\ \forall t \geq 0, \tag{S10.8}$$

which completes the proof of Lemma 5.

S10.3 Proof of Theorem 6

We are now ready to present the proof of Theorem 6. The proof is divided into three steps: **Step 1**: establishing the convergence property of the first-stage algorithm (Theorem 1, GLM version); **Step 2**: establishing the signal adaptive

result of the second-stage algorithm (Theorem 2, GLM version). Define $M' := \max\left(M,\ \frac{M}{L},\ MU\right)$.

Step 1 (First stage Algorithm 1, GLM version) Define the event

$$\mathcal{E}_{GLM} := \left\{ \left\| \tilde{\Xi} \right\|_\infty < 4\sqrt{\frac{aUM'\log p}{n}},\ \text{and}\ \left\| \frac{1}{\sqrt{n}} \tilde{\xi} \right\|_\infty < 3\sqrt{\frac{aU\log n}{n}} \right\},$$

and

$$\mathcal{E}_{X,GLM} := \left\{ \begin{array}{c} \frac{1}{2M'} \le \Lambda_k\left(\frac{1}{n}\sum_{i=1}^n b''(\tilde{\zeta}_i^t) X_S^{(i)} (X_S^{(i)})^\top\right) \le 2M',\ S\subset[p]:\ |S| \le (2B'+1)s \text{ and } 1\le k\le |S|, \\ \sup_{S\subset[p]:\ |S|\le (B'+1)s}\ \sup_{O\subset[n]:\ |O|\le (B'+1)o} \left\| \tilde{X}_{O,S}^t \right\|_2 \le (\sqrt{B'}+1) f' \sqrt{n} \end{array} \right\},$$

where

$$B' := \left(\frac{\kappa'+\delta'}{\kappa'-\delta'}\right)^2 > 1, \quad \delta' := \frac{4M'^2}{4M'^2+1} \in (0,1), \quad f' := C_M U \sqrt{\frac{s\log p + o\log n}{n}},$$

and the decay rate $\kappa' \in (\delta', 1)$. The event $\mathcal{E}_{X,GLM}$ follows from Lemma 4 with $C_1 = 2B'+1$ for the restricted isometry part and $C_1 = B'+1$ for the restricted incoherence part. These choices are fixed once M, L, U and κ' are fixed. By Lemmas 4 and 5, we have $\mathbf{P}(\mathcal{E}_{GLM} \cap \mathcal{E}_{X,GLM}) \ge 1 - O\left(p^{-2} + n^{-3}\right)$. Under the event $\mathcal{E}_{GLM} \cap \mathcal{E}_{X,GLM}$, we choose some proper parameters

$$\begin{aligned} \lambda_{\beta,\infty} &\asymp \sqrt{aU}\left(\sqrt{\frac{\log p}{n}} + \frac{o\log n}{n\sqrt{s}}\right), \\ \lambda_{\theta,\infty} &\asymp \sqrt{aU}\left(\sqrt{\frac{\log n}{n}} + \frac{s\log p}{n\sqrt{o}}\right). \end{aligned} \tag{S10.9}$$

Then, by choosing the learning rate $\eta' \in \left[\frac{2M'}{4M'^2+1}, \frac{4M'}{4M'^2+1}\right]$ and applying a proof

process similar to that in Section S3.2, we ensure that, with probability at least $1 - O(p^{-2} + n^{-3})$, both estimators are ℓ_0-sparse, i.e., $\|\widehat{\beta}^{GLM}\|_0 \lesssim s,\ \|\widehat{\theta}^{GLM}\|_0 \lesssim o$, and satisfies

$$\begin{aligned} \|\widehat{\beta}^{GLM} - \beta^*\|_2^2 \lesssim & aU \left(\frac{s \log p}{n} + \frac{o^2 \log^2 n}{n^2} \right), \\ \|\widehat{\theta}^{GLM} - \theta^*\|_2^2 \lesssim & aU \left(\frac{o \log n}{n} + \frac{s^2 \log^2 p}{n^2} \right). \end{aligned} \tag{S10.10}$$

Step 2 (Second-stage Algorithm 2, GLM version) Following Lemmas 1 and 5, we have

$$\mathbf{P} \left\{ \underbrace{\frac{1}{n} \left\| X_{\cdot S^*}^\top \tilde{\xi} \right\|_2 \leq \sqrt{aUM} \sqrt{\frac{4s + 6\log(1/\varrho)}{n}}}_{=:\mathcal{E}'_{GLM}} \right\} \geq 1 - O\left(p^{-2s} + \varrho\right).$$

We fix the threshold parameters λ_β, λ_θ as in (S10.9). Then, under the signal condition

$$\min_{i \in S^*} |\beta_i^*| \geq \left(\frac{4\kappa'}{\kappa' - \delta'} + \frac{\kappa' - \delta'}{4\kappa'} \right) \cdot \lambda_\beta \asymp \sqrt{aU} \left(\sqrt{\frac{\log p}{n}} + \frac{o \log n}{n\sqrt{s}} \right), \tag{S10.11}$$

by applying a proof technique similar to that in Section S4.2, we get the sharper bound

$$\|\tilde{\beta}^{GLM} - \beta^*\|_2 \lesssim \sqrt{\frac{s + \log(1/\varrho)}{n}} + \frac{s \log p + o \log n}{n},$$

with a probability greater than $1 - \varrho - O(p^{-2} + n^{-3})$. Therefore, we complete the proof of Theorem 6, demonstrating the signal adaptivity of our procedure under the GLM setting.

S11 Proof of Theorem 7

We first establish the estimation error bound for the regime where $\delta \in (0, 1)$. Before proceeding to the main analysis, we provide some necessary probabilistic inequalities.

Concentration of $|\Xi'_j|$. Define

$$\Xi'_j := \frac{1}{n}\sum_{i=1}^{n} X_{ij}\xi_i = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\psi_\tau(\epsilon_i), \text{ for every } j \in [p].$$

For every $k \geq 2$, by the definition of $\psi_\tau(\cdot)$, we have the moment inequality

$$\begin{aligned}
\mathbf{E}(|\xi_i|^k) =& \mathbf{E}\left\{\tau^k \mathbf{1}(|\epsilon_i| > \tau)\right\} + \mathbf{E}\left\{|\epsilon_i|^k \mathbf{1}(|\epsilon_i| \leq \tau)\right\} \\
\leq& \mathbf{E}\left\{\tau^k \frac{|\epsilon_i|^{1+\delta}}{\tau^{1+\delta}} \cdot \mathbf{1}(|\epsilon_i| > \tau)\right\} + \mathbf{E}\left\{|\epsilon_i|^k \frac{\tau^{k-1-\delta}}{|\epsilon_i|^{k-1-\delta}} \cdot \mathbf{1}(|\epsilon_i| \leq \tau)\right\} \\
\leq& \tau^{k-1-\delta}\mathbf{E}\left\{|\epsilon_i|^{1+\delta}\mathbf{1}(|\epsilon_i| > \tau) + |\epsilon_i|^{1+\delta}\mathbf{1}(|\epsilon_i| \leq \tau)\right\} \\
\leq& v_\delta \tau^{k-1-\delta}.
\end{aligned}$$

By the sub-Gaussian property, we obtain

$$\mathbf{E}(|X_{ij}|^k) = \int_{t\in\mathbb{R}^+} \mathbf{P}(|X_{ij}|^k > t)\mathrm{d}t \leq (2\Sigma_{jj})^{k/2} k\Gamma(k/2).$$

Then, for every λ satisfying $|\lambda| \leq \frac{1}{2\tau\sqrt{2\Sigma_{jj}}}$, we get a Bernstein-type bound:

$$\begin{aligned}
\mathbf{E}(e^{\lambda X_{ij}\xi_j}) \leq & 1 + \frac{1}{2}\lambda^2 \mathbf{E} X_{ij}^2 \xi_j^2 + \sum_{k\geq 3} \frac{\lambda^k}{k!} \mathbf{E}|X_{ij}\xi_j|^k \\
\leq & 1 + 2\lambda^2 \Sigma_{jj} v_\delta \tau^{1-\delta} + \sum_{k\geq 3} \frac{\lambda^k v_\delta \tau^{k-1-\delta} \cdot (\sqrt{2\Sigma_{jj}})^k \; k!}{k!} \\
\leq & 1 + 4\lambda^2 \Sigma_{jj} v_\delta \tau^{1-\delta} \\
\leq & \exp(4\lambda^2 \Sigma_{jj} v_\delta \tau^{1-\delta}), \text{ for each } (i,j) \in [n] \times [p].
\end{aligned}$$

Therefore, by the probability union bound, we get the concentration inequality

$$\mathbf{P}\left(\max_{j\in[p]} \left[|\Xi_j'| - \frac{v_\delta \sqrt{2\Sigma_{jj}}}{\tau^\delta} - \frac{6\sqrt{2\Sigma_{jj}} \cdot \tau \log p}{n} \right] \geq 0 \right) \leq 2p^{-2}. \tag{S11.12}$$

Concentration of $\|\theta^*\|_0$. We next bound the term $\|\theta^*\|_0 = \sum_{i\in[n]} \mathbf{1}(\epsilon_i - \psi_\tau(\epsilon_i) \neq 0) = \sum_{i\in[n]} \mathbf{1}(|\epsilon_i| > \tau)$. Define $q := \mathbf{P}(|\epsilon_i| > \tau) \leq \frac{v_\delta}{\tau^{1+\delta}}$. By Bernstein's inequality (Appendix D.4 in Mohri et al. (2018)), we have

$$\begin{aligned}
\mathbf{P}\left(\|\theta^*\|_0 \geq \frac{2nv_\delta}{\tau^{1+\delta}} \right) \leq & \mathbf{P}\left\{ \frac{1}{n} \sum_{i\in[n]} \mathbf{1}(|\epsilon_i| > \tau) \geq q + \frac{v_\delta}{\tau^{1+\delta}} \right\} \\
\leq & \exp\left(-\frac{n(v_\delta/\tau^{1+\delta})^2}{2q(1-q) + 2/3 \cdot v_\delta/\tau^{1+\delta}} \right) \leq \exp\left(-\frac{nv_\delta}{3\tau^{1+\delta}} \right).
\end{aligned} \tag{S11.13}$$

Take $\tau = \left(\frac{nv_\delta}{6\log p} \right)^{\frac{1}{1+\delta}}$. We now replace the event $\mathcal{E}$ (S3.5) (in the proof of Theorem

1) by the newly defined $\mathcal{E}_{heavy}$:

$$\mathcal{E}_{heavy} := \left\{ \begin{array}{c} \|\Xi'\|_\infty \leq \frac{v_\delta \sqrt{2\max_j \Sigma_{jj}}}{\tau^\delta} + \frac{6\sqrt{2\max_j \Sigma_{jj}} \cdot \tau \log p}{n}, \\ o \leq \frac{2nv_\delta}{\tau^{1+\delta}}, \text{ and } \|\xi\|_\infty \leq \tau \end{array} \right\}.$$

By (S11.12) and (S11.13), we have $\mathbf{P}(\mathcal{E}_{heavy}) \geq 1 - 3p^{-2}$.

Near-optimal error bound. We first construct a suitable $\lambda_{\theta,0}$. According to (S3.8), it suffices to ensure that

$$\sqrt{o}\lambda_{\theta,0} > \|\theta^*\|_2 \vee (\sqrt{o}\lambda_{\theta,\infty}). \tag{S11.14}$$

Following the proof of Theorem 1, we aim to choose a proper $\lambda_{\theta,\infty}$ such that

$$\sqrt{o}\lambda_{\theta,\infty} \gtrsim \sqrt{o}\left\|\frac{1}{\sqrt{n}}\xi\right\|_\infty + \sqrt{\frac{s\log p + o\log n}{n}} \cdot \sqrt{s}\|\Xi'\|_\infty.$$

Additionally,

$$\begin{aligned} &\sqrt{o}\left\|\frac{1}{\sqrt{n}}\xi\right\|_\infty + \sqrt{\frac{s\log p + o\log n}{n}} \cdot \sqrt{s}\|\Xi'\|_\infty \\ \lesssim_\tau &\sqrt{\frac{o}{n}} + \sqrt{\frac{s\log p + o\log n}{n}} \cdot \sqrt{s}\left(\frac{v_\delta}{\tau^\delta} + \frac{\tau\log p}{n}\right) \\ \asymp_\tau &\sqrt{\frac{o}{n}}\left(1 + \sqrt{\frac{s\log p}{n}}\sqrt{\frac{s\log p + o\log n}{n}}\right) \\ \asymp_\tau &\sqrt{\frac{o}{n}}, \end{aligned}$$

where the second equality uses $o \asymp \log p$, and the last equality follows from the sample size condition $n \gtrsim (s + \log n) \log p$. Therefore, it is enough to choose

$$\lambda_{\theta,0} \gtrsim \frac{1}{\sqrt{n}} \left(\|\epsilon\|_\infty \vee \tau \right) \asymp \frac{1}{\sqrt{n}} \left(\|\epsilon\|_\infty \bigvee \left(\frac{n v_\delta}{\log p} \right)^{\frac{1}{1+\delta}} \right),$$

which guarantees (S11.14). Moreover, by the union bound and the Markov inequality, for every $r > 0$,

$$\mathbf{P}(\|\epsilon\|_\infty > r) \leq n \mathbf{E} \left\{ \mathbf{1} \left(|\epsilon|^{1+\delta} > r^{1+\delta} \right) \right\} \leq \frac{n v_\delta}{r^{1+\delta}}.$$

Hence, by taking $\lambda_{\theta,0} \gtrsim n^{-1/2} (n v_\delta / \varrho)^{\frac{1}{1+\delta}}$, we ensure that, with probability at least $1 - \varrho - O(p^{-2})$, the initial threshold $\lambda_{\theta,0}$ satisfies (S11.14).

Following the proof of Theorems 1 and 2, with a probability greater than $1 - \varrho - O(p^{-2})$, the output of our AC-IHT algorithm satisfies

$$\begin{aligned}
\|\tilde{\beta} - \beta^*\|_2 \lesssim & \sqrt{s} \|\Xi'\|_\infty + \sqrt{\frac{s \log p + o \log n}{n}} \cdot \sqrt{o} \left\| \frac{1}{\sqrt{n}} \xi \right\|_\infty \\
\lesssim & \frac{\sqrt{s} v_\delta}{\tau^\delta} + \frac{\sqrt{s} \log p}{n} \tau + \sqrt{\frac{v_\delta s \log p}{n} \tau^{1-\delta}} + \frac{v_\delta \sqrt{\log n}}{\tau^\delta} \\
\lesssim & \frac{v_\delta \sqrt{s + \log n}}{\tau^\delta} + \frac{\sqrt{s + \log n} \log p}{n} \tau \\
\asymp & v_\delta^{\frac{1}{1+\delta}} \sqrt{s + \log n} \left(\frac{\log p}{n} \right)^{\frac{\delta}{1+\delta}},
\end{aligned}$$

where the first two inequalities follow from Assumption 1, which imposes constant upper and lower bounds on the spectrum of Σ.

In the case $\delta \geq 1$, by the moments inequality, we have $v_1^{1/2} := (\mathbf{E}|\epsilon_i|^2)^{1/2} \leq$

$(\mathbf{E}|\epsilon_i|^{1+\delta})^{1/(1+\delta)} \leq v_\delta^{1/(1+\delta)}$. Then we take $\delta' = 1$ and $v_{\delta'} = v_1$. Following the proof sketch provided above, with a probability at least $1 - \varrho - O(p^{-2})$, we can choose $\lambda_{\theta,0} \gtrsim \sqrt{v_1/\varrho}$ and get the upper bound

$$\|\tilde{\beta} - \beta^*\|_2 \lesssim v_{\delta'}^{\frac{1}{1+\delta'}} \sqrt{s + \log n} \left(\frac{\log p}{n}\right)^{\frac{\delta'}{1+\delta'}} = \sqrt{\frac{v_1(s + \log n)\log p}{n}}.$$

Therefore, we complete the proof of Theorem 7, demonstrating the applicability of our algorithm to the high-dimensional heavy-tailed regression.

References

Chen, M., C. Gao, and Z. Ren (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics 46*(5), 1932 – 1960.

Dalalyan, A. and P. Thompson (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's m-estimator. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Hsu, D., S. Kakade, and T. Zhang (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability 17*(none), 1 – 6.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Minsker, S., M. Ndaoud, and L. Wang (2024). Robust and tuning-free sparse linear regression via square-root slope. *SIAM Journal on Mathematics of Data Science 6*(2), 428 – 453.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of Machine Learning, second edition*.

The MIT Press.

Ndaoud, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. In A. Garivier and S. Kale (Eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, Volume 98 of *Proceedings of Machine Learning Research*, pp. 647 – 668. PMLR.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory 57*(10), 6976 – 6994.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Ser. Stat. New York, NY: Springer.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wainwright, M. (2007). Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *2007 IEEE International Symposium on Information Theory*, pp. 961 – 965.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.