Statistica Sinica Preprint No: SS-2025-0224	
Title	Adversarial Contamination Meets Hard Thresholding: an Iterative Algorithm with Signal Adaptivity and Minimax Optimality
Manuscript ID	SS-2025-0224
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0224
Complete List of Authors	Shixiang Liu and Hanming Yang
Corresponding Authors	Hanming Yang
E-mails	yanghanming@ruc.edu.cn
Notice: Accepted author version.	

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

Shixiang Liu, Hanming Yang*

Renmin University of China

Abstract: Pervasive data contamination—stemming from measurement errors, outliers, or adversarial corruption—has motivated the development of robust statistical methods. In this context, we propose a two-stage Adversarial Contamination-resistant Iterative Hard Thresholding (AC-IHT) algorithm for high-dimensional regression with contamination. Our nonconvex algorithm achieves minimax near-optimal (up to logarithmic terms) estimation by iteratively updating the coefficient vector and the contamination vector with different thresholding scales. We further demonstrate that our AC-IHT estimator is signal-adaptive: under proper signal conditions, it adaptively attains a sharper estimation rate and more accurate support recovery. Moreover, it enjoys the strong oracle property, laying a theoretical foundation for asymptotic inference. Numerical experiments confirm its superior finite-sample performance. Finally, we discuss theoretical extensions of the proposed procedure to generalized linear models and to heavy-tailed noise settings.

Key words and phrases: Adversarial contamination, Iterative hard thresholding, Non-convex optimization, Signal adaptivity, Strong oracle property.

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

1. Introduction

Adversarial contamination has become a significant concern in both statistical theory and its applications. Models that explicitly address contamination yield enhanced robustness, improved generalization, and a more realistic representation of real-world data. From a statistical perspective, this paper considers a basic high-dimensional linear model with adversarial contamination, given by

$$Y = X\beta^* + \sqrt{n}\theta^* + \xi, \tag{1.1}$$

where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix, and $\beta^* \in \mathbb{R}^p$ is the coefficient vector. The term $\sqrt{n}\theta^* \in \mathbb{R}^n$ represents an adversarial contamination that models potentially malicious or outlying observations in the response vector. The $\sqrt{n}$ factor is introduced for technical convenience, ensuring that the columns of the augmented design matrix $(X \mid \sqrt{n}I_n) \in \mathbb{R}^{n\times(p+n)}$ are of comparable magnitude. The noise term ξ is an independent centered σ-subGaussian random vector, i.e., $\mathbf{E}(e^{\lambda^\top \xi}) \le \exp(\sigma^2\|\lambda\|_2^2/2)$ for all $\lambda \in \mathbb{R}^n$, and ξ is independent of X. We assume that β^* is s-sparse and θ^* is $\tilde{o}$-sparse, i.e., $\|\beta^*\|_0 = s < p$ and $\|\theta^*\|_0 = \tilde{o} < n$.

Such a formulation provides a unified framework that can potentially benefit various domains, including robust principal component analysis with missing data (Chen et al., 2021), figure classification with covariate-shifted samples (Heng and Soh, 2025), and econometric models with heterogeneous treatment effects (Goldsmith-

Pinkham et al., 2024). To address model (1.1) in the high-dimensional regime $p \gtrsim n$, this paper introduces a two-stage **A**dversarial **C**ontamination-resistant **I**terative **H**ard **T**hresholding (AC-IHT) algorithm and analyzes how the signal strengths of β^* and θ^* affect the statistical inference of β^*.

1.1 Literature review

Adversarial contamination To achieve robust statistical inference, one often treats the contamination as additional covariates, thereby reducing its influence (Sardy et al., 2001; Gannaz, 2007). In high-dimensional settings, Nguyen and Tran (2013) applied simultaneous ℓ_1 penalties to both coefficient β and contamination θ, deriving a joint error bound. Alternative penalty forms have also been explored: She and Owen (2011) introduced nonconvex penalties for outlier detection, Lee et al. (2012) studied ℓ_2 regularization on contamination within regression and classification frameworks, and Kong et al. (2018) proposed an adaptive ℓ_1 penalty with weights determined from an initial robust fit. From an algorithmic viewpoint, Bhatia et al. (2015, 2017); Suggala et al. (2019) analyzed a series of iterative hard thresholding methods under sparse contamination, establishing their convergence properties.

Recent work on adversarial contamination has shifted toward the non-asymptotic minimax estimation rates. Following the general minimax theory for ϵ-Huber contamination model proposed by Chen et al. (2016, 2018), Gao (2020) introduced a multivariate-depth estimator that is minimax-optimal (but not efficiently computable). To bridge this gap, efficient ℓ_1-regularized methods achieving near-optimal

rates were developed (Dalalyan and Thompson, 2019; Chinot, 2020; Sasai and Fujisawa, 2020). Subsequent contributions produced estimators that adapt to the noise level σ, sparsity s, and contamination o (Finocchio et al., 2021; Minsker et al., 2024). There are also advanced extensions considering non-sparse settings (Pensia et al., 2025; Hammouda et al., 2024), low-rank matrix regression under contamination (Thompson, 2020; Shen et al., 2025) and models where both covariates and responses are contaminated (Sasai and Fujisawa, 2025). For a comprehensive overview, we refer readers to Loh (2025).

Iterative hard thresholding and signal adaptivity Iterative optimization has emerged as a central analysis paradigm in statistics (Jain et al., 2014; Huang et al., 2018; She et al., 2021) that goes beyond the traditional Empirical Risk Minimization (ERM) framework. This is because the iteration trajectory itself acts as a form of regularization, such as early stopping (Fan et al., 2023) and iterative thresholding (Blumensath and Davies, 2008, 2009; Liu and Foygel Barber, 2019), and provides a finer balance between computational efficiency and statistical precision. Furthermore, iterative schemes provide superior flexibility by allowing for the dynamic tuning of hyperparameters, such as learning rates (She et al., 2023; Shen et al., 2025), truncation thresholds (Ndaoud, 2020), compression coefficients (She et al., 2023), and so on. As a widely adopted iterative scheme, Iterative Hard Thresholding (IHT, Blumensath and Davies (2008)) is particularly noted for applying no additional shrinkage to the selected signals. Notably, Ndaoud (2020) demonstrated that IHT exhibits a remarkable **signal adaptivity** property: if the signal scale satisfies

$\min_{i:\beta_i^*\neq 0}|\beta_i| \geq C\sigma\sqrt{n^{-1}\log(p/s)}$, the IHT estimator adaptively achieves an ℓ_2-estimation rate of order $\sigma\sqrt{s/n}$. This refined rate not only improves upon the standard high-dimensional minimax rate $\sigma\sqrt{n^{-1}s\log(p/s)}$ (Raskutti et al., 2011), but also aligns with the minimax phase transition phenomenon (Ndaoud, 2019).

1.2 Motivation and inspiration

In the high-dimensional contamination setting, most existing works established minimax (near) optimal estimations (Dalalyan and Thompson, 2019; Finocchio et al., 2021; She et al., 2022; Minsker et al., 2024; Shen et al., 2025), without studying how the signal strengths of β^* and θ^* affect the estimation accuracy. Moreover, support recovery and asymptotic distributions remain unexplored: Minsker and Shen (2025) proved support recovery for the Lasso estimator under the incoherence condition, but did not address its asymptotic behavior. Therefore, we address the following questions:

> *In the high-dimensional contamination model* (1.1)*, do the signal strengths of β^* and θ^* influence the recovery of β^*? Does the IHT estimator possess signal-adaptivity to β^* estimation, and retain the strong oracle property? Are these results supported by minimax-optimal guarantees?*

The following works inspire our study. From the upper-bound perspective, Dalalyan and Thompson (2019); Minsker et al. (2024) showed that consistent estimation of β^* is possible only if the columns of $X \in \mathbb{R}^{n\times p}$ are "nearly uncorrelated" with those of

the identity matrix I_n. This insight motivates our incoherence-type requirement in Proposition 1. From the lower-bound perspective, Chen et al. (2018); Chinot (2020) described how to construct least-favorable distributions specific to contamination settings; in the uncontaminated sparse model, Butucea et al. (2018) derived a minimax lower bound via a Bayesian-risk approach, which crucially guides our lower bound construction.

1.3 Main contributions

We affirmatively address the questions posed in Section 1.2. The main contributions of this paper are threefold:

- **Signal adaptivity and minimax near-optimality** We propose a two-stage AC-IHT algorithm for sparse linear regression under adversarial contamination. We establish that our estimator $\tilde{\beta}$ exhibits signal adaptivity, and explicitly characterize how the signal strengths of β^* and θ^* affect the statistical inference of β^*, as summarized in Table 1. Our results are supported by minimax near-optimal guarantees.

- **Strong oracle property** Under proper signal strength conditions, we prove that the AC-IHT algorithm converges to the oracle estimator, enabling exact support recovery of β^*. Furthermore, we establish asymptotic normality for our estimator $\tilde{\beta}$, providing a foundation for valid inference.

- **Additional theoretical extensions** We extend the AC-IHT algorithm

to generalized linear models, accommodating a diverse range of response distributions. Furthermore, we establish a theoretical connection between adversarial contamination and heavy-tailed noise, demonstrating that AC-IHT achieves minimax near optimality in both settings.

Table 1: The signal adaptivity of the AC-IHT estimator $\tilde{\beta}$ in signal estimation and support recovery.

Theoretical Result / Signal Condition	**Convergence Rate** $\|\tilde{\beta} - \beta^*\|_2$	**Support Recovery**
None	$\sigma\left(\sqrt{\frac{s\log p}{n}} + \frac{o\log n}{n}\right)$ Theorem 2	$\left\|\mathrm{supp}(\beta^*) \triangle \mathrm{supp}(\tilde{\beta})\right\| = O(s)$ Theorem 2
With Signal Condition : $\min_{i:\beta_i^*\neq 0}\|\beta_i^*\| \geq \beta^\ddagger$	$\sigma\left(\sqrt{\frac{s+\log(1/\varrho)}{n}} + \frac{s\log p + o\log n}{n}\right)$ Theorem 2	$\left\|\mathrm{supp}(\beta^*) \triangle \mathrm{supp}(\tilde{\beta})\right\| \prec s$ Corollary 1
With Signal Conditions: $\min_{i:\beta_i^*\neq 0}\|\beta_i^*\| \geq \beta^\ddagger$, $\min_{k:\theta_k^*\neq 0}\|\theta_k^*\| \geq \theta^\ddagger$	$\sigma\sqrt{\frac{s+\log(1/\varrho)}{n-o}}$ Theorem 3	$\mathrm{supp}(\tilde{\beta}) = \mathrm{supp}(\beta^*)$ Theorem 3

Note: Here $\beta^\ddagger := C_\beta\sigma\left\{\frac{\log p}{n} + \frac{o^2\log^2 n}{n^2 s}\right\}^{1/2}$, $\theta^\ddagger := C_\theta\sigma\left\{\frac{\log n}{n} + \frac{s^2\log^2 p}{n^2 o}\right\}^{1/2}$, where $o = \|\theta^*\|_0 \vee 1$, and $C_\beta, C_\theta > 0$ are some absolute constants. All results hold with a probability greater than $1 - \varrho - O(p^{-2} + n^{-3})$.

However, solely estimating β^* is far more challenging than jointly estimating β^* and θ^*. To overcome this, we innovatively employ separate thresholding levels for $\tilde{\beta}$ and $\tilde{\theta}$ in each iteration. This separation enables a refined analysis of how the nuisance component θ^* impacts the estimation of β^*, and yields both delicate sparsity patterns and sharp ℓ_2-error bounds for our estimator $\tilde{\beta}$.

1.4 Organization of the paper

The present paper is organized as follows. Section 2 establishes the procedure of the two-stage AC-IHT algorithm. Section 3 presents the theoretical guarantees of our algorithm. Section 4 presents numerical experiments that illustrate our theoretical findings. Section 5 discusses extensions of AC-IHT to generalized linear models, its connection to heavy-tailed regression, and related future directions. Detailed proofs and additional simulations are available in the supplementary material.

Notation For sequences a_n and b_n, we write $a_n = O(b_n)$ (or $a_n \lesssim b_n$) if $a_n \leq Cb_n$ for some constant $C > 0$ and all large n, and $a_n \prec b_n$ if $a_n/b_n \to 0$ as $n \to \infty$. We write that $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Let $[m] = \{1, 2, \ldots, m\}$, and $\mathbf{1}(\cdot)$ be the indicator function. Define $x \vee y = \max\{x, y\}$. Let $S^* = \{i : \beta_i^* \neq 0\} \subseteq [p]$ and $O^* = \{k : \theta_k^* \neq 0\} \subseteq [n]$ denote the support sets of β^* and θ^*, respectively. For sets A and B with sizes $|A|$ and $|B|$, let $\beta_A = (\beta_j)_{j\in A} \in \mathbb{R}^{|A|}$, $X_{\cdot,A} = (X_j)_{j\in A} \in \mathbb{R}^{n\times|A|}$, and $X_{A,B} \in \mathbb{R}^{|A|\times|B|}$ be the submatrix of $X \in \mathbb{R}^{n\times p}$ with rows and columns in A and B. The symmetric difference of A and B is defined as $A \triangle B = (A \setminus B) \cup (B \setminus A)$. For a vector β, denote $\|\beta\|_2$ as its Euclidean norm, $\|\beta\|_0$ as the number of its nonzero entries, and $\mathrm{supp}(\beta)$ as its support set. For matrix X, denote $\|X\|_2$ as its operator norm.

2. Two-stage AC-IHT Algorithm

This section presents the two-stage AC-IHT algorithm, with its first and second stages detailed in Sections 2.1 and 2.2, respectively. The first-stage procedure provides an initial estimator with a near-optimal estimation rate. The second-stage algorithm refines this estimate to obtain a final estimator with the desirable theoretical properties summarized in Table 1.

2.1 The first stage: dynamic thresholding iteration

We start by defining the squared ℓ_2 loss for model (1.1) as

$$L(\beta, \theta) = \frac{1}{2n}\big\|Y - X\beta - \sqrt{n}\,\theta\big\|_2^2.$$

Based on $L(\beta, \theta)$, we propose the first-stage AC-IHT Algorithm 1. Each iteration in Algorithm 1 can be summarized in three steps: **gradient update**, **threshold parameters update**, and **hard thresholding operation**.

Gradient update: We derive the partial derivatives of $L(\beta, \theta)$ with respect to β and θ as

$$\frac{\partial L}{\partial \beta}(\beta^t, \theta^t) = -\frac{1}{n}X^\top\big(Y - X\beta^t - \sqrt{n}\,\theta^t\big),$$
$$\frac{\partial L}{\partial \theta}(\beta^t, \theta^t) = -\frac{1}{\sqrt{n}}\big(Y - X\beta^t - \sqrt{n}\,\theta^t\big),$$

and employ the gradient descent approach to update the parameters:

$$\textbf{Step 1.1:} \quad \begin{aligned} H_\beta^{t+1} &\leftarrow \beta^t - \eta \frac{\partial L}{\partial \beta}(\beta^t, \theta^t), \\ H_\theta^{t+1} &\leftarrow \theta^t - \eta \frac{\partial L}{\partial \theta}(\beta^t, \theta^t), \end{aligned} \tag{2.1}$$

where $\eta > 0$ denotes the learning rate, the explicit choice of which will be specified in Theorem 1.

Threshold parameters update: For $\lambda > 0$, define the hard thresholding operator $\mathcal{T}_\lambda^m : \mathbb{R}^m \to \mathbb{R}^m$, such that

$$\left(\mathcal{T}_\lambda^m(z)\right)_j = z_j \times \mathbf{1}\left(|z_j| \geq \lambda\right), \quad \text{for every } z \in \mathbb{R}^m \text{ and } j \in [m]. \tag{2.2}$$

In Algorithm 1, the thresholding parameters in operator $\mathcal{T}_\lambda^m$ are dynamically updated at each iteration. Specifically, the thresholds $\lambda_{\beta,0}$ and $\lambda_{\theta,0}$ are initialized to sufficiently large values. These thresholds are then iteratively decreased and used at each iteration for hard-thresholding operations, until they reach their respective universal statistical levels $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$ (see Theorem 1 for their specific forms). The following scheme guarantees that the sequences $\{\lambda_{\beta,t}\}_{t\geq 0}$ and $\{\lambda_{\theta,t}\}_{t\geq 0}$ are monotonically non-increasing:

$$\textbf{Step 1.2:} \quad \begin{aligned} \lambda_{\beta,t+1} &\leftarrow (\kappa \times \lambda_{\beta,t}) \vee \lambda_{\beta,\infty}, \\ \lambda_{\theta,t+1} &\leftarrow (\kappa \times \lambda_{\theta,t}) \vee \lambda_{\theta,\infty}. \end{aligned} \tag{2.3}$$

Here, $\kappa \in (0, 1)$ controls the decay rate, typically chosen as 0.9 in practice. This

Algorithm 1: The first stage of the AC-IHT algorithm

Input: $\beta^0 = \mathbf{0}_p,\ \theta^0 = \mathbf{0}_n,\ Y,\ X,\ \lambda_{\beta,0},\ \lambda_{\theta,0},\ \lambda_{\beta,\infty},\ \lambda_{\theta,\infty},\ \kappa,\ t = 0$

Output: $\widehat{\beta},\ \widehat{\theta}$

while $t < \max\left\{\log_{1/\kappa}(\lambda_{\beta,0}/\lambda_{\beta,\infty}), \log_{1/\kappa}(\lambda_{\theta,0}/\lambda_{\theta,\infty})\right\}$ **do**

  Step 1.1: Update H_β^{t+1} and H_θ^{t+1} using (2.1);

  Step 1.2: Update $\lambda_{\beta,t+1}$ and $\lambda_{\theta,t+1}$ using (2.3);

  Step 1.3: Update β^{t+1} and θ^{t+1} using (2.4);

  $t \leftarrow t + 1$;

end

$\widehat{\beta} \leftarrow \beta^t, \quad \widehat{\theta} \leftarrow \theta^t$

process not only provides an explicit stopping time for the procedure, but also enables sparsity control in each iteration.

Hard thresholding operation: In this step, we apply the hard-thresholding operator (2.2), together with the updated thresholds (2.3), to the raw updates in (2.1), ensuring a dynamic regularization at each iteration:

$$\textbf{Step 1.3:} \qquad \begin{aligned} \beta^{t+1} &\leftarrow \mathcal{T}^p_{\lambda_{\beta,t+1}}(H_\beta^{t+1}), \\ \theta^{t+1} &\leftarrow \mathcal{T}^n_{\lambda_{\theta,t+1}}(H_\theta^{t+1}). \end{aligned} \tag{2.4}$$

Steps 1.2 and 1.3 are inspired by the strategy of Ndaoud (2020), using gradually decreasing hard-threshold levels to limit variable inclusion in each iteration. It also aligns with the motivation of the LARS algorithm (Efron et al., 2004). This approach ensures computational efficiency and sparsity in the outputs. A key novelty of our Algorithm 1 is the separate handling of β^t and θ^t updates, which allows

us to derive a delicate sparse pattern and a minimax near-optimal ℓ_2-error bound for $\widehat{\beta}$, as shown in Theorem 1. In contrast, choosing common threshold levels $\lambda_{\beta,\infty} \asymp \lambda_{\theta,\infty} \asymp \sigma\{\log(p+n)/n\}^{1/2}$ may not deliver comparably precise estimates.

2.2 The second stage: fixed thresholding iteration

While the first-stage Algorithm 1 delivers a minimax near-optimal initial estimate $\widehat{\beta}$ (shown in Theorem 1), it tends to omit some true support variables and falls short in estimation accuracy in practice, even when the support signals of β^* are relatively strong. Therefore, a refinement step is required: Starting from $\widehat{\beta}$ and $\widehat{\theta}$, we execute successive iterations, as detailed in Algorithm 2. The second-stage Algorithm 2 differs from Algorithm 1 in two ways: First, it initializes at $\tilde{\beta}^0 = \widehat{\beta}$ and $\tilde{\theta}^0 = \widehat{\theta}$, the estimators produced by the first-stage Algorithm 1. Second, instead of updating the threshold parameters in each iteration, in the second stage, we use two **fixed** values λ_β and λ_θ (see Theorem 2 for their specific forms). Each iteration in the second stage consists of two steps: a **gradient update** (identical to **Step 1.1** (2.1)), and a **hard thresholding operation** using the fixed thresholds:

$$\textbf{Step 2.2:} \quad \begin{aligned} \tilde{\beta}^{t+1} &\leftarrow \mathcal{T}^p_{\lambda_\beta}(H^{t+1}_\beta), \\ \tilde{\theta}^{t+1} &\leftarrow \mathcal{T}^n_{\lambda_\theta}(H^{t+1}_\theta). \end{aligned} \tag{2.5}$$

The complete second-stage algorithm is presented in Algorithm 2, and our two-stage AC-IHT algorithm combines Algorithm 1 (initial estimation) and Algorithm 2 (refinement): Algorithm 1 iteratively updates the estimator while decreasing the

Algorithm 2: The second stage of the AC-IHT algorithm

Input: $\tilde{\beta}^0 = \widehat{\beta},\ \tilde{\theta}^0 = \widehat{\theta},\ Y,\ X,\ \lambda_\beta,\ \lambda_\theta,\ t = 0$

Output: $\tilde{\beta},\ \tilde{\theta}$

while $t \le C \log n$ **do**

  Step 2.1: Update H_β^{t+1} and H_θ^{t+1} using (2.1);

  Step 2.2: Update $\tilde{\beta}^{t+1}$ and $\tilde{\theta}^{t+1}$ using (2.5);

  $t \leftarrow t + 1$;

end

$\tilde{\beta} \leftarrow \tilde{\beta}^t, \quad \tilde{\theta} \leftarrow \tilde{\theta}^t$

thresholding levels $\lambda_{\beta,t}$ and $\lambda_{\theta,t}$ until they reach their limiting values $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$. Algorithm 2 then continues the iteration with fixed thresholding levels, yielding a refined final output. In this sense, the two algorithms together form a two-step "debiased" procedure, which does not require data splitting.

Remark 1 (Practical Role of Algorithm 1)**.** The second-stage Algorithm 2 performs a refined bias correction on initial estimates $\widehat{\beta}$. Moreover, any initial estimator satisfying the guarantees of Theorem 1, such as the Lasso estimator (Dalalyan and Thompson, 2019) or the square-root Slope estimator (Minsker et al., 2024), can be used as input to Algorithm 2. This demonstrates the generality of the proposed IHT framework. However, Algorithm 1 still has its practical advantages: It provides a computationally efficient initial estimation, while the refinement of estimation accuracy is handled by Algorithm 2. Moreover, as shown in Theorem 2, the limiting thresholds $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$ used in Algorithm 1 can be directly used in Algorithm 2, avoiding additional tuning and reducing computational cost.

3. Theoretical Guarantees

In this section, we study the statistical properties of the two-stage AC-IHT algorithm. Define $o := \|\theta^*\|_0 \vee 1$, and assume $\|\beta^*\|_0 = s \geq 1$ without loss of generality. Recall that the noise ξ in (1.1) is assumed to be σ-subGaussian. We further impose two key assumptions.

Assumption 1 (Sub-Gaussian design)**.** The design matrix $X \in \mathbb{R}^{n\times p}$ is row-wise independent and sub-Gaussian: each row $X_{i,\cdot} \overset{d}{=} Z_i\Sigma^{1/2}$, where $Z_i = (Z_{i1}, \cdots, Z_{ip}) \in \mathbb{R}^{1\times p}$ and each Z_{ij} is i.i.d. centered 1-sub-Gaussian random variable such that $\mathbf{E}(Z_i^\top Z_i) = I_p$. The population covariance $\Sigma \in \mathbb{R}^{p\times p}$ satisfies

$$M^{-1} \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq M,$$

where $M > 1$ is a universal constant.

Assumption 2 (Sample size)**.** The sample size n satisfies $\max(s\log p,\ o\log n) \lesssim n$.

Assumption 1 controls the correlation among covariates and is commonly used in the literature on iterative algorithms (Fan et al., 2023; Han et al., 2026). These assumptions yield the following proposition, which underpins the theoretical guarantees of our iterative procedure.

Proposition 1 (Restricted isometry and incoherence)**.** *For any fixed constant $C_1 > 0$, assume Assumption 1 holds and Assumption 2 holds in the specific form $n \geq$*

$30M^2C_1C_\Sigma \max(s\log p, o\log n)$, *where* $C_\Sigma > 0$ *is a constant depending only on* $\|\Sigma\|_2$. *Then 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* $|S| \le C_1 s$*, the sample covariance matrix satisfies:*

$$\frac{1}{2M}\|u\|_2^2 \;\le\; u^\top \left(\frac{X^\top X}{n}\right)_{S,S} u \;\le\; 2M\|u\|_2^2, \text{ for every } u \in \mathbb{R}^{|S|}. \tag{3.1}$$

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

$$\sup_{S\subset[p]:\ |S|\le C_1 s}\ \sup_{O\subset[n]:\ |O|\le C_1 o} \|X_{O,S}\|_2 \le \sqrt{C_1}C_M\sqrt{s\log p + o\log n}. \tag{3.2}$$

This proposition characterizes both the restricted isometry property of the design matrix X and the restricted correlation between columns of X and the identity matrix I_n (since $X_{O,S}^\top = X_{\cdot,S}^\top I_{\cdot,O}$). The latter controls how adversarial contamination can distort the estimation of β^*, and such an incoherence-type condition is standard in outlier analyses (Dalalyan and Thompson, 2019; Minsker et al., 2024). If (3.2) fails, for example, in the case $n = p$ and $X = \sqrt{n}\, I_n$, then model (1.1) reduces to $Y = \sqrt{n}(\beta^* + \theta^*) + \xi$, making it impossible to estimate β^* alone consistently. Further analysis of this property is provided in Section S2 of the Supplementary Material.

Remark 2 (Sub-Gaussian limitation)**.** Proposition 1 relies on a sub-Gaussian design and a spectrally bounded covariance matrix Σ, and thus cannot be directly generalized to heavy-tailed designs addressed in some robust statistics literature (Sun et al., 2020; Pensia et al., 2025). We speculate that such extensions may require some technical tools like truncation (Section 4 in Sun et al. (2020)), and leave this for future work.

3.1 Property of the first stage estimation

We begin with the statistical guarantee of the first-stage Algorithm 1.

Theorem 1 (Initial Estimation)**.** *Assume that Assumptions 1 and 2 hold. For tuning parameters in Algorithm 1, suppose that the learning rate* $\eta \in \left[\frac{2M}{4M^2+1}, \frac{4M}{4M^2+1}\right]$*, the decay rate* $\kappa \in \left(\frac{4M^2}{4M^2+1}, 1\right)$*, and the initial thresholds satisfy* $\sqrt{s}\lambda_{\beta,0} > \|\beta^*\|_2$ *and* $\sqrt{o}\lambda_{\theta,0} > \|\theta^*\|_2$*. Let*

$$\lambda_{\beta,\infty} = C_{\beta,1}\sigma\sqrt{\frac{M\log p}{n}} + C_{\beta,2}\sigma\sqrt{\frac{s\log p + o\log n}{ns}}\sqrt{\frac{o\log n}{n}},$$
$$\lambda_{\theta,\infty} = \frac{3C_{\beta,1}\sigma}{4}\sqrt{\frac{\log n}{n}} + \frac{4C_{\beta,2}\sigma\sqrt{M}}{3}\sqrt{\frac{s\log p + o\log n}{no}}\sqrt{\frac{s\log p}{n}},$$

where $C_{\beta,1}$*,* $C_{\beta,2}$ *are two constants depending on* M, κ*, and* η*. Let* $(\widehat{\beta}, \widehat{\theta})$ *denote the output of Algorithm 1. Then with probability at least* $1 - O(p^{-2} + n^{-3})$*, both estimators are* ℓ_0*-sparse, i.e.,* $\|\widehat{\beta}\|_0 \lesssim s$*,* $\|\widehat{\theta}\|_0 \lesssim o$*, and satisfy*

$$\begin{aligned}\|\widehat{\beta} - \beta^*\|_2^2 \lesssim& \sigma^2\left(\frac{s\log p}{n} + \frac{o^2\log^2 n}{n^2}\right),\\ \|\widehat{\theta} - \theta^*\|_2^2 \lesssim& \sigma^2\left(\frac{o\log n}{n} + \frac{s^2\log^2 p}{n^2}\right).\end{aligned} \tag{3.3}$$

The ℓ_2 error of $\widehat{\beta}$ consists of two terms: (i) the estimation rate of an s-sparse vector $(\sigma^2 s \log p)/n$, and (ii) the contamination proportion $(\sigma^2 o^2 \log^2 n)/n^2$. Furthermore, by interpolating the ℓ_0 sparsity and ℓ_2 rate, we conclude that

$$\|\widehat{\beta} - \beta^*\|_q^q \lesssim \sigma^q \left\{ s \left(\frac{\log p}{n} \right)^{q/2} + s^{1-q/2} \left(\frac{o \log n}{n} \right)^q \right\}, \text{ for every } q \in [1, 2],$$

demonstrating that $\widehat{\beta}$ is minimax near-optimal in terms of the ℓ_q error for all $q \in [1, 2]$ (see Theorem 4 for the lower bound). Additionally, please refer to Section S3.1 of the Supplementary Material for the specific sample size requirements of this theorem.

Remark 3 (Two phases of estimation accuracy)**.** The bound for $\widehat{\beta}$ in (3.3) can be rewritten as

$$\|\widehat{\beta} - \beta^*\|_2^2 \lesssim \sigma^2 \max\left(\frac{s \log p}{n}, \frac{o^2 \log^2 n}{n^2} \right) = \begin{cases} \sigma^2 \frac{s \log p}{n}, & \text{if } o \leq \frac{\sqrt{ns \log p}}{\log n}, \\ \sigma^2 \frac{o^2 \log^2 n}{n^2}, & \text{if } \frac{\sqrt{ns \log p}}{\log n} < o \lesssim \frac{n}{\log n}. \end{cases}$$

Thus, there are two distinct regimes: When the number of contaminated samples o is relatively small, the estimation error attains the uncontaminated near-optimal rate $(\sigma^2 s \log p)/n$ (Raskutti et al., 2011). As o increases, the error of $\widehat{\beta}$ becomes dominated by the squared contamination proportion (up to a logarithmic term).

Remark 4 (Symmetric rate and joint bound)**.** Within the bounds of (3.3), the terms $o \log n$ and $s \log p$ play symmetric roles: swapping them in the bound of $\widehat{\beta}$ yields the corresponding bound of $\widehat{\theta}$. This symmetry reflects a duality between β^* and

θ^*: If one views the model (1.1) as a sparse linear regression, then $\sqrt{n}\theta^*$ constitutes an ℓ_0-sparse contamination of the observations; conversely, if one views (1.1) as a (sub-Gaussian) location model, the term $X\beta^*$ acts as contamination, representing the image of an ℓ_0-sparse vector β^* into the observation space via X. Summing the individual bounds in (3.3) gives the joint error control:

$$\|\widehat{\beta} - \beta^*\|_2^2 + \|\widehat{\theta} - \theta^*\|_2^2 \lesssim \sigma^2 \left(\frac{s \log p + o \log n}{n} \right), \tag{3.4}$$

which is minimax near-optimal (see Theorem 6 in She et al. (2022) for the joint lower bound). A comparison of (3.3) and (3.4) shows that focusing solely on β^* delivers a more refined guarantee.

3.2 Property of the two-stage estimation

This subsection establishes signal adaptivity and the strong oracle property of the final estimator $\tilde{\beta}$, which is obtained from the second-stage Algorithm 2. We first specify the signal condition imposed on β^*.

Assumption 3 (Signal condition for β^*)**.** There exists a constant $C_\beta > 0$ such that

$$\min_{i \in S^*} |\beta_i^*| \geq C_\beta \sigma \left(\sqrt{\frac{\log p}{n}} + \frac{o \log n}{n\sqrt{s}} \right), \tag{3.5}$$

This signal condition guarantees that all nonzero components of β^* are well separated from zero, thereby facilitating sharper estimation rates.

Theorem 2 (Signal-adaptive estimation)**.** *Assume that Assumptions 1 and 2 hold. We set the tuning parameters in Algorithm 2 as* $\eta \in \left[\frac{2M}{4M^2+1}, \frac{4M}{4M^2+1}\right]$, $\lambda_\beta = \lambda_{\beta,\infty}$, *and* $\lambda_\theta = \lambda_{\theta,\infty}$, *where* $\lambda_{\beta,\infty}, \lambda_{\theta,\infty}$ *are specified in Theorem 1. Let* $(\tilde{\beta}, \tilde{\theta})$ *denote the output of Algorithm 2. Then, for any given* $\varrho \in (0,1)$, *with probability at least* $1 - \varrho - O(p^{-2} + n^{-3})$, *we have* $\|\tilde{\beta}\|_0 \lesssim s$, $\|\tilde{\theta}\|_0 \lesssim o$, *and the estimation error is signal-adaptive:*

$$\|\tilde{\beta} - \beta^*\|_2^2 \lesssim \begin{cases} \sigma^2 \left(\dfrac{s + \log(1/\varrho)}{n} + \dfrac{(s \log p + o \log n)^2}{n^2} \right), & \text{if Assumption 3 holds,} \\ \sigma^2 \left(\dfrac{s \log p}{n} + \dfrac{o^2 \log^2 n}{n^2} \right), & \text{otherwise.} \end{cases} \tag{3.6}$$

Theorem 2 establishes the signal adaptivity of the second-stage AC-IHT algorithm: If the signal condition in Assumption 3 fails, $\tilde{\beta}$ achieves a near-optimal rate; if it holds, $\tilde{\beta}$ adaptively achieves a sharper estimation rate than the minimax near-optimal rate. This signal adaptivity is a key advantage of hard-thresholding-based methods and is generally difficult to achieve by convex procedures such as the Lasso (Bellec, 2018). Furthermore, if Assumption 3 holds and $n \gtrsim s \log^2 p$, from (3.6) we can get a sharper bound:

$$\|\tilde{\beta} - \beta^*\|_2^2 \lesssim \sigma^2 \left(\frac{s + \log(1/\varrho)}{n} + \frac{o^2 \log^2 n}{n^2} \right),$$

which matches the estimation rate (up to a $\log n$ factor) as if the support of β^* were known (Hammouda et al., 2024).

Signal-adaptive estimation has been well-studied in uncontaminated models (Fan et al., 2018; Ndaoud, 2019, 2020; Fan et al., 2023). However, to our knowledge, no existing work addresses this property under adversarial contamination, and Theorem 2 fills this gap. Moreover, by utilizing this property, we can derive a more refined variable selection guarantee than that in Theorem 2.

Corollary 1 (Selection error)**.** *Under the conditions of Theorem 2 and Assumption 3, if* $\frac{o^2 \log^2 n}{n} \prec s \log p \prec n$*, then, as* $s \succ 1$*, with probability at least* $1 - O(p^{-2} + n^{-3})$ *we have* $\left| supp(\beta^*) \triangle supp(\tilde{\beta}) \right| \prec s$.

Utilizing the interplay between estimation and selection, Corollary 1 effectively controls the variable selection error of the estimator $\tilde{\beta}$ (obtained from two-stage AC-IHT). Moreover, Theorem 5 establishes that the required signal strength Assumption 3 is nearly necessary, demonstrating that our procedure is minimax near-optimal (up to logarithmic terms) in the variable selection task.

We further analyze the oracle property of $\tilde{\beta}$. An additional signal strength condition on θ^* is presented as follows.

Assumption 4 (Signal condition for θ^*)**.** There exists a sufficiently large (absolute) constant $C_\theta > 0$ such that

$$\min_{k \in O^*} |\theta_k^*| \geq C_\theta \sigma \left(\sqrt{\frac{\log n}{n}} + \frac{s \log p}{n\sqrt{o}} \right). \tag{3.7}$$

This assumption is instrumental and sufficient for identifying contaminated samples, which allows our procedure to further remove the bias induced by corruption. Similar requirements appear in the recent literature (see Theorem 3 in Hammouda et al. (2024)). Moreover, it implies that a larger number of contaminated samples o relaxes the required signal strength for outlier identification. This phenomenon is empirically validated through numerical experiments in Supplementary Material S1.5. Define the oracle estimator $\beta^\dagger$ as

$$\beta^\dagger := \mathrm{Proj}_\beta \left\{ \underset{\substack{\beta_{(S^*)^c}=\mathbf{0} \\ \theta_{(O^*)^c}=\mathbf{0}}}{\arg\min} \ \frac{1}{2n} \left\| Y - X\beta - \sqrt{n}\theta \right\|_2^2 \right\}, \tag{3.8}$$

where Proj_β denotes the projection onto the β component of the joint vector $(\beta^\top, \theta^\top)^\top$.

Theorem 3 (Oracle estimation and selection consistency)**.** *Assume that all conditions in Theorem 2 hold, and Assumptions 3 and 4 hold. Let $\{\tilde{\beta}^t\}_{t\geq 0}$ be the sequence of iterates from Algorithm 2. Then with probability at least $1 - O(p^{-2} + n^{-3})$, there exist two absolute constants $r \in (0,1)$ and $C > 0$ such that*

$$\|\tilde{\beta}^t - \beta^\dagger\|_2 \leq C \times r^t, \quad \textit{for every } t \geq 0. \tag{3.9}$$

Additionally, by terminating Algorithm 2 after $t \geq C' \log n$ iterations, for any given $\varrho \in (0,1)$, with probability at least $1 - \varrho - O(p^{-2} + n^{-3})$ the output $\tilde{\beta}$ reaches the

oracle rate and achieves selection consistency:

$$\|\tilde{\beta} - \beta^*\|_2^2 \lesssim \sigma^2 \left(\frac{s + \log(1/\varrho)}{n - o} \right), \quad supp(\tilde{\beta}) = supp(\beta^*), \quad supp(\tilde{\theta}) = supp(\theta^*).$$

Under suitable signal conditions, $\tilde{\beta}$ (obtained from two-stage AC-IHT) converges geometrically to the oracle estimator $\beta^\dagger$, therefore reaching the ℓ_2 estimation rate $\sigma\{s/(n-o)\}^{1/2}$ as if the true support of β^* were known and the contaminated data were excluded. Leveraging this fact, one can further establish the asymptotic property of each linear functional of $\tilde{\beta}$.

Corollary 2 (Asymptotic normality)**.** *Assume that all conditions in Theorem 3 hold, and assume* $n \succ \max\left(s^2 \log^2 p,\ s^3\right)$. *Define* $c_\xi := Var(\xi_1)/\sigma^2$. *Then for every* $\gamma \in \mathbb{R}^s$ *with* $0 < \|\gamma\|_2 < \infty$, *as* $n, p \to \infty$, *we have*

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

Corollary 2 demonstrates that $\tilde{\beta}$ possesses asymptotic normality, thereby enabling statistical inference. This property distinguishes our method from existing methods tailored for adversarial contamination (Dalalyan and Thompson, 2019; Minsker et al., 2024).

Remark 5 (Revisit signal adaptivity)**.** Assumptions 3 and 4 provide the technical prerequisites for the AC-IHT procedure to converge to the oracle estimator, ensuring

both selection consistency and asymptotic normality. When these signal conditions are not satisfied, our estimator may no longer enjoy these strong guarantees; nevertheless, it at least maintains minimax near-optimality and ℓ_0 sparsity, as established in Theorem 2. This underscores that our procedure does not require prior knowledge of the true signals for practical execution, and thus is inherently signal-adaptive: it delivers the (nearly) best possible estimation accuracy for the given data, with stronger signals adaptively yielding sharper results.

For clarity, we summarize the conditions on sample size, signal strength, and corruption required by the above theoretical results in Table 2.

Table 2: Summary of sample size, signal strength, and corruption assumptions for main theoretical results.

<table>
<tr><th>Assumption / Theoretical Result</th><th>Sample Size</th><th>Strength of β^* and θ^*</th><th>Corruption Number</th></tr>
<tr><td>Theorem 1
(Initial Estimation)</td><td rowspan="3">$\max(s\log p, o\log n) \lesssim n$
(Assumption 2)</td><td>None</td><td rowspan="2">$o \lesssim \frac{n}{\log n}$</td></tr>
<tr><td>Theorem 2
(Signal-adaptive estimation)</td><td rowspan="2">$\min_{i:\beta_i^*\neq 0}|\beta_i^*| \geq \beta^\ddagger$
(Assumption 3)</td></tr>
<tr><td>Corollary 1
(Selection error)</td><td>$o \prec \frac{\sqrt{ns\log p}}{\log n}$</td></tr>
<tr><td>Theorem 3
(Oracle estimation and selection consistency)</td><td rowspan="2">$\max\left(s^2\log^2 p, s^3, o\log n\right) \prec n$</td><td>$\min_{i:\beta_i^*\neq 0}|\beta_i^*| \geq \beta^\ddagger$
$\min_{k:\theta_k^*\neq 0}|\theta_k^*| \geq \theta^\ddagger$
(Assumption 3 and 4)</td><td rowspan="2">$o \lesssim \frac{n}{\log n}$</td></tr>
<tr><td>Corollary 2
(Asymptotic normality)</td></tr>
</table>

Note: Here $\beta^\ddagger := C_\beta\sigma\left\{\frac{\log p}{n} + \frac{o^2\log^2 n}{n^2 s}\right\}^{1/2}$, $\theta^\ddagger := C_\theta\sigma\left\{\frac{\log n}{n} + \frac{s^2\log^2 p}{n^2 o}\right\}^{1/2}$, where $o = \|\theta^*\|_0 \vee 1$, and $C_\beta, C_\theta > 0$ are some absolute constants.

3.3 Minimax lower bounds

This subsection provides minimax lower bound guarantees for estimation and support recovery. For ease of display, we consider the model with Gaussian noise: assume that each $Y_i|X_{i\cdot}$ is drawn from $\mathcal{N}(X_{i\cdot}\beta^*,\ \sigma^2)$, where $X_{i\cdot} \in \mathbb{R}^{1\times p}$ follows from the random design as introduced in Assumption 1. Here, we only require a sparse eigenvalue assumption as

$$\sup_{S\subset[p],|S|\leq 2s} \|\Sigma_{S,S}\|_2 \leq C_{2s}, \tag{3.10}$$

where $C_{2s} > 0$ is an absolute constant. Denote the uncontaminated distribution of $(X_{i\cdot}, Y_i)$ as $\mathbf{P}_{X,Y}$ and write our model space as

$$\begin{aligned}\mathcal{M}(\beta, o) := \{&(n-k) \text{ observations are drawn from } \mathbf{P}_{X,Y},\\ &k \text{ observations are drawn from arbitrary } \mathbf{Q},\ \text{where } 0 \leq k \leq o\,\}.\end{aligned} \tag{3.11}$$

Much of the robust estimation literature considers the ϵ-Huber contamination model $(X_{i\cdot}, Y_i) \sim (1-\epsilon)\mathbf{P} + \epsilon\mathbf{Q}$ (Chen et al., 2018; Gao, 2020; Chinot, 2020). In contrast, our model setting (3.11) explicitly constrains the number of outliers (o) rather than the contamination probability (ϵ). This difference yields that our lower bound results hold independent significance.

Theorem 4 (Estimation lower bound)**.** *Assume that Assumption 2 and equation* (3.10) *hold and* $o \geq 9$. *Then for any* $q \in [1, 2]$ *with an absolute constant* $c_q \in$

$\left(0, \frac{(8C_{2s})^{-q/2}}{160}\right)$, *we have*

$$\inf_{\widehat{\beta}} \sup_{\beta^*:\|\beta^*\|_0 \le s} \sup_{\mathbf{R}\in\mathcal{M}(\beta^*,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \left(\left\| \widehat{\beta} - \beta^* \right\|_q^q \right) \ge c_q \sigma^q \left\{ s \left(\frac{\log(ep/s)}{n} \right)^{q/2} + s^{1-q/2} \frac{o^q}{n^q} \right\},$$

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

Combined with Theorem 2, it implies that our two-stage AC-IHT algorithm is minimax near-optimal (up to logarithmic factors), and can even surpass this minimax rate under a proper signal strength condition.

We next establish the minimax lower bound for variable selection. Define

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

as the s-sparse vector space with a minimal signal strength $a > 0$.

Theorem 5 (Selection lower bound)**.** *Assume that Assumption 2 and equation* (3.10) *hold and* $o \ge 8$*, then with an absolute constant* $c_2 \in (0, 1/5)$*, we have*

$$\inf_{\widehat{S}} \sup_{\beta^*\in\mathcal{B}(s,a)} \sup_{\mathbf{R}\in\mathcal{M}(\beta^*,o)} \mathbf{E}_{(X,Y)\sim\mathbf{R}} \left(\left| \widehat{S} \triangle \mathit{supp}(\beta^*) \right| \right) \ge c_2 s \tag{3.12}$$

holds for $a \le \frac{\sigma}{4\sqrt{2C_{2s}}} \left(\sqrt{\frac{\log(ep/s)}{n}} + \frac{1}{\sqrt{s}} \frac{o}{n} \right)$.

This result shows that no procedure can recover $\mathrm{supp}(\beta^*)$ with $o(s)$ selection error if the signals of β^* are relatively weak. It further demonstrates that the signal strength Assumption 3 is minimax near-optimal (up to logarithm terms) for the

variable selection task in a contamination setting.

4. Simulation Studies

This section presents numerical experiments that complement our theoretical findings. We set $p = 1000$, $n = 300$, $s = o = 10$, and assume X is generated from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$ for $i, j \in [p]$ and $\rho = 0.25$. The true parameter vectors β^* and θ^* have their first 10 entries set to 0.5 and all remaining entries set to 0. All simulations in this section are executed with 300 replications. Our two-stage AC-IHT algorithm involves two tuning parameters, $\lambda_{\beta,\infty}$ and $\lambda_{\theta,\infty}$, as suggested by Theorems 1 and 2 (see Supplementary S3.4 for the initial tuning of $(\lambda_{\beta,0}, \lambda_{\theta,0})$). Here, we determine their values using a Massart-type information criterion by minimizing the following objective

$$\left\| Y - X\tilde{\beta} - \sqrt{n}\tilde{\theta} \right\|_2^2 + A(\tilde{s}\log p + \tilde{o}\log n),$$

where $\tilde{\beta} = \tilde{\beta}(\lambda_{\beta,\infty}, \lambda_{\theta,\infty})$ and $\tilde{\theta} = \tilde{\theta}(\lambda_{\beta,\infty}, \lambda_{\theta,\infty})$ are the estimators obtained with the candidate parameters $(\lambda_{\beta,\infty}, \lambda_{\theta,\infty})$, and $\tilde{s} = |\mathrm{supp}(\tilde{\beta})|$ and $\tilde{o} = |\mathrm{supp}(\tilde{\theta})|$. In this section, we set $A = 2$ and search for the best pair $(\lambda_{\beta,\infty}, \lambda_{\theta,\infty})$ over the region $(10^{-2}, 1) \times (10^{-2}, 1)$. We set the decay rate $\kappa = 0.9$ and the learning rate $\eta = 0.75$.

For benchmarking, we compare our AC-IHT method with several existing robust estimation methods: IHT-ℓ_1 (Shen et al., 2025), the Progressive Iterative Quantile-Thresholding (PIQ) estimator (She et al., 2022), the Adaptive Huber (Ada-Huber)

estimator (Sun et al., 2020), AC-LASSO (Thompson, 2020), and AC-SCAD (which replaces the ℓ_1 penalty in AC-LASSO with the SCAD penalty). In addition, the Oracle estimator (defined in (3.8)) is included as an ideal reference for comparison.

Five metrics are used to evaluate the estimation performance: (i) the ℓ_2-error $\|\beta - \beta^*\|_2$; (ii) the ℓ_∞-error $\|\beta - \beta^*\|_\infty$; (iii) the Σ-norm-error $\|\beta - \beta^*\|_\Sigma = \big\{(\beta - \beta^*)^\top \Sigma(\beta - \beta^*)\big\}^{1/2}$; (iv) Matthews correlation coefficient (MCC) and (v) Symmetric difference (Sym_diff). Here, MCC and Sym_diff are defined as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})\}^{1/2}},$$

$$\mathrm{Sym\_diff} = \mathrm{FP} + \mathrm{FN},$$

where $\mathrm{TP} = |\widehat{S} \cap S^*|$, $\mathrm{TN} = |\widehat{S}^c \cap (S^*)^c|$, $\mathrm{FP} = |\widehat{S} \cap (S^*)^c|$, $\mathrm{FN} = |\widehat{S}^c \cap S^*|$ and $\widehat{S}$ is the support index of the estimator. Among these metrics, a smaller value indicates better estimation or variable selection performance for all except MCC. For MCC, values closer to 1 indicate better selection accuracy.

4.1 Estimation accuracy and selection consistency

In this subsection, we consider the contamination model with noise term ξ independently generated from the following three distributions, each with unit variance: the standard Gaussian distribution $\mathcal{N}(0,1)$, the Rademacher distribution, and the uniform distribution $\mathcal{U}(-\sqrt{3},\sqrt{3})$. Table 3 shows that AC-IHT achieves competitive performance: It yields the lowest estimation errors and the best support

recovery, with results close to the Oracle benchmark $\beta^{\dagger}$. Compared with other methods such as IHT-ℓ_1, PIQ, and Adaptive Huber, AC-IHT is more stable across different noise settings. Additionally, the AC-SCAD estimator shows intermediate performance between AC-IHT and AC-LASSO, consistent with the SCAD penalty being sandwiched between hard-thresholding and soft-thresholding penalties.

4.2 Oracle property

This subsection demonstrates that our AC-IHT method exhibits the oracle property. Under the standard Gaussian distribution, we first consider the convergence to the oracle estimator with increasing sample size n. Figure 1 shows that, as n increases from 200 to 800, AC-IHT performs well and gradually approaches the oracle estimator, illustrating high accuracy and precise support recovery. This empirical result aligns with the convergence guarantees established in Theorem 3. The simulation results illustrating the asymptotic normality and iteration-wise convergence behavior of AC-IHT are provided in Section S1.2 and S1.4 of the Supplementary Material.

Table 3: Comparison of estimation accuracy across different methods under contaminated data.

Method	$\|\beta-\beta^*\|_2$	$\|\beta-\beta^*\|_\infty$	$\|\beta-\beta^*\|_\Sigma$	**MCC**	**Sym_diff**
			Gaussian		
AC-IHT	0.222 (0.004)	0.140 (0.003)	0.213 (0.004)	0.989 (0.001)	0.243 (0.031)
IHT-ℓ_1	0.489 (0.014)	0.347 (0.010)	0.458 (0.013)	0.953 (0.003)	0.890 (0.050)
PIQ	0.535 (0.010)	0.341 (0.007)	0.522 (0.010)	0.912 (0.004)	2.007 (0.091)
Ada-Huber	0.531 (0.005)	0.276 (0.003)	0.616 (0.006)	0.697 (0.003)	10.760 (0.185)
AC-LASSO	0.484 (0.004)	0.229 (0.003)	0.540 (0.005)	0.556 (0.011)	38.993 (2.412)
AC-SCAD	0.350 (0.005)	0.182 (0.004)	0.333 (0.004)	0.561 (0.011)	32.257 (1.383)
Oracle	0.196 (0.003)	0.118 (0.002)	0.186 (0.002)	1.000 (0.000)	0.000 (0.000)
			Rademacher		
AC-IHT	0.214 (0.004)	0.136 (0.003)	0.202 (0.004)	0.992 (0.001)	0.170 (0.025)
IHT-ℓ_1	0.799 (0.004)	0.499 (0.001)	0.752 (0.004)	0.899 (0.002)	1.900 (0.031)
PIQ	0.543 (0.010)	0.346 (0.007)	0.530 (0.010)	0.913 (0.004)	1.987 (0.097)
Ada-Huber	0.528 (0.005)	0.279 (0.003)	0.607 (0.005)	0.659 (0.003)	13.083 (0.181)
AC-LASSO	0.489 (0.004)	0.238 (0.003)	0.548 (0.004)	0.588 (0.010)	32.317 (2.222)
AC-SCAD	0.368 (0.004)	0.198 (0.004)	0.345 (0.004)	0.591 (0.011)	30.363 (1.650)
Oracle	0.197 (0.003)	0.121 (0.002)	0.185 (0.002)	1.000 (0.000)	0.000 (0.000)
			Uniform		
AC-IHT	0.211 (0.004)	0.135 (0.003)	0.202 (0.004)	0.992 (0.001)	0.173 (0.025)
IHT-ℓ_1	0.680 (0.010)	0.463 (0.006)	0.638 (0.009)	0.919 (0.002)	1.533 (0.044)
PIQ	0.534 (0.010)	0.341 (0.007)	0.523 (0.010)	0.913 (0.004)	2.003 (0.091)
Ada-Huber	0.532 (0.005)	0.276 (0.003)	0.615 (0.006)	0.669 (0.003)	12.463 (0.206)
AC-LASSO	0.487 (0.004)	0.234 (0.003)	0.547 (0.004)	0.586 (0.010)	32.340 (2.141)
AC-SCAD	0.367 (0.004)	0.196 (0.004)	0.348 (0.004)	0.580 (0.011)	30.857 (1.556)
Oracle	0.193 (0.003)	0.117 (0.002)	0.183 (0.002)	1.000 (0.000)	0.000 (0.000)

Note. The numbers in parentheses denote the standard errors.

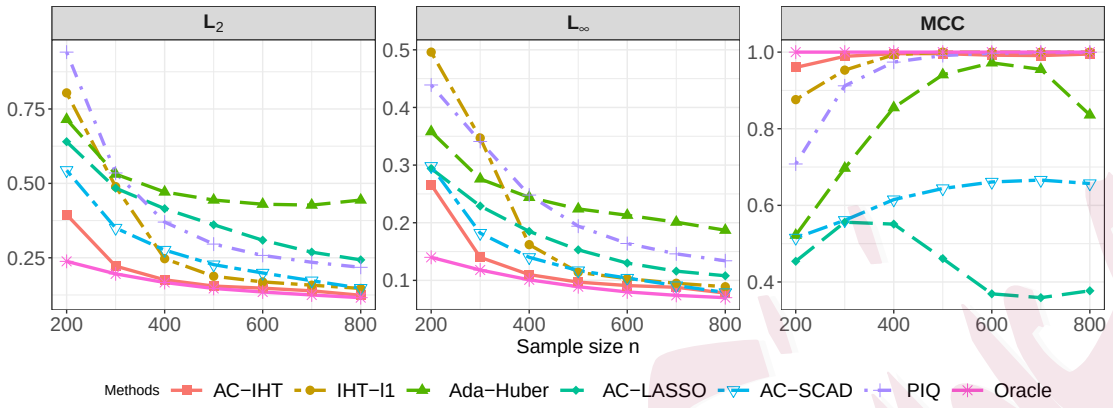

Figure 1: Estimation accuracy and support recovery performance with increasing sample size.

5. Discussion

We conclude by discussing the generalizability of our procedure and several directions for future research.

5.1 Extension to GLM

Here we extend our two-stage AC-IHT algorithm to the Generalized Linear Models (GLMs) setting, thereby allowing for a broader class of response distributions. Consider a GLM setup with n independent observations $\{(X_{i,\cdot}, Y_i)\}_{i=1}^n$, where $X_{i,\cdot} \in \mathbb{R}^{1\times p}$ is the i-th row of X. The distribution of each Y_i is assumed to follow an exponential family characterized by the natural parameter ζ_i^* and a scale parameter a, with the density function:

$$f(Y_i;\zeta_i^*) = \exp\left(\frac{Y_i\zeta_i^* - b(\zeta_i^*)}{a} + c(Y_i,a)\right), \tag{5.1}$$

where $b(\cdot)$ denotes a cumulant function. We adopt the canonical link function $b'^{-1}(\cdot)$ and define the linear predictor as $\zeta_i^* = X_{i,\cdot}\beta^* + \sqrt{n}\theta_i^*$ under our adversarial contamination setting. We consider minimizing the negative log-likelihood

$$-\log \prod_{i\in[n]} f(Y_i;\zeta_i^*) \propto \frac{1}{n}\sum_{i\in[n]} \Big\{ b(\zeta_i^*) - Y_i\zeta_i^* - a \times c(Y_i,a) \Big\}.$$

Therefore, only the **gradient update** step in our two-stage AC-IHT algorithm requires modification, while all other steps remain unchanged. The updated gradient step is as follows:

$$\textbf{Step G.1:} \quad \begin{aligned} H_\beta^{t+1} &\leftarrow \beta^t - \frac{\eta}{n}\sum_{i=1}^n X_{i,\cdot}^\top \left(b'(X_{i,\cdot}\beta^t + \sqrt{n}\theta_i^t) - Y_i \right), \\ H_\theta^{t+1} &\leftarrow \theta^t - \frac{\eta}{\sqrt{n}}\sum_{i=1}^n e_i \left(b'(X_{i,\cdot}\beta^t + \sqrt{n}\theta_i^t) - Y_i \right), \end{aligned} \tag{5.2}$$

where $\eta > 0$ denotes the learning rate, $e_i \in \mathbb{R}^{n\times 1}$ is a vector with its i-th component equals 1 and all other components are 0. To implement this variant, we replace **Step 1.1** in Algorithm 1 and Step 2.1 in Algorithm 2 with **Step G.1** (5.2). We introduce the following regularity assumption to establish the theoretical guarantees for the GLM extension.

Assumption 5. The function $b(\cdot)$ is twice-differentiable, and assume that each $\zeta_i^* \in \Theta$, where the parameter space $\Theta \subseteq \mathbb{R}$ is a closed (finite or infinite) interval. There exist two constants $0 < L \leq U < \infty$ such that the function $b''(\cdot)$ satisfies $L \leq \inf_{t\in\Theta} b''(t) \leq \sup_{t\in\mathbb{R}} b''(t) \leq U$.

This assumption controls the variance $\mathrm{Var}(Y_i) = ab''(\zeta_i^*)$ and provides uniform strong convexity (on Θ). It is a standard condition in high-dimensional GLM analyses (Abramovich and Grinshtein, 2016).

Theorem 6 (GLM). *Suppose that Assumptions 1, 2, and 5 hold. Let $\tilde{\beta}^{GLM}$ be the estimator obtained from the two-stage AC-IHT algorithm in the GLM setting. For appropriately chosen algorithmic parameters (refer to Supplementary Material S10), it holds with probability at least $1-\varrho-O(p^{-2}+n^{-3})$ that $\|\tilde{\beta}^{GLM}\|_0 \lesssim s$. Furthermore, the squared error $\|\tilde{\beta}^{GLM}-\beta^*\|_2^2$ achieves the same signal-adaptive rate as established in* (3.6).

Therefore, in the GLM setting, $\tilde{\beta}^{\mathrm{GLM}}$ maintains both sparsity and signal adaptivity as in (3.6), establishing the generality of our two-stage AC-IHT algorithm.

5.2 Relationship with heavy-tailed regression

Here, we establish a formal connection between the contaminated model (1.1) and heavy-tailed regression. Consider the high-dimensional heavy-tailed model

$$Y = X\beta^* + \epsilon \in \mathbb{R}^n, \tag{5.3}$$

where we assume the white noise ϵ is independent of X, and each ϵ_i is zero-mean and has a bounded $(1+\delta)$-th moment, i.e., $\mathbf{E}(|\epsilon_i|^{1+\delta}) \le v_\delta$ for every $i \in [n]$, where $\delta > 0$. For such a heavy-tailed regression, Sun et al. (2020) established the minimax

ℓ_2 estimation rate as

$$\begin{cases} v_\delta^{\frac{1}{1+\delta}} \sqrt{s} \left(\frac{\log p}{n}\right)^{\frac{\delta}{1+\delta}}, & \text{if } \delta \le 1, \\ v_1^{\frac{1}{2}} \sqrt{s} \left(\frac{\log p}{n}\right)^{\frac{1}{2}}, & \text{if } \delta > 1, \end{cases}$$

and achieved this rate via a regularized Huber estimator.

Now we reformulate model (5.3) as an adversarial contamination model through the following decomposition:

$$Y = X\beta^* + \big(\epsilon - \psi_\tau(\epsilon)\big) + \psi_\tau(\epsilon) = X\beta^* + \sqrt{n}\theta^* + \xi. \tag{5.4}$$

Here, each $\xi_i := \psi_\tau(\epsilon_i)$, $i \in [n]$ represents the τ-truncated noise, and $\psi_\tau(\cdot) := \max(\min(\cdot, \tau), -\tau)$ is the truncation operator. The remaining term, $\sqrt{n}\theta_i^* := \epsilon_i - \psi_\tau(\epsilon_i)$, is treated as an outlier component. Leveraging this decomposition, we apply the AC-IHT algorithm to the heavy-tailed regression and establish the following guarantees.

Theorem 7 (Heavy tailedness)**.** *Consider model* (5.3) *and suppose Assumption 1 holds with sample size* $n \gtrsim (s + \log n) \log p$*. Run Algorithm 1 and 2 with the learning rate* $\eta \in \left[\frac{2M}{4M^2+1}, \frac{4M}{4M^2+1}\right]$*, the decay rate* $\kappa \in \left(\frac{4M^2}{4M^2+1}, 1\right)$*, and the initial threshold* $\lambda_{\beta,0} > \|\beta^*\|_2/\sqrt{s}$*. Denote by* $(\tilde{\beta}, \tilde{\theta})$ *the output of Algorithm 2. Then:*

1. ***Case*** $\delta \in (0, 1)$*. With* $\lambda_{\theta,0} \gtrsim n^{-1/2}(n v_\delta/\varrho)^{\frac{1}{1+\delta}}$, $\lambda_{\beta,\infty} = \lambda_\beta \asymp v_\delta^{\frac{1}{1+\delta}} \left(\frac{\log p}{n}\right)^{\frac{\delta}{1+\delta}} \left(1 + \sqrt{\frac{\log n}{s}}\right)$, *and* $\lambda_{\theta,\infty} = \lambda_\theta \asymp n^{\frac{1-\delta}{2(1+\delta)}} \left(\frac{v_\delta}{\log p}\right)^{\frac{1}{1+\delta}}$*, under a probability at least* $1 - \varrho - O(p^{-2})$

we have:

$$\|\tilde{\beta}\|_0 \lesssim s, \quad \|\tilde{\beta} - \beta^*\|_2 \lesssim v_\delta^{\frac{1}{1+\delta}} \left(\frac{\log p}{n}\right)^{\frac{\delta}{1+\delta}} \sqrt{s + \log n}.$$

2. ***Case*** $\delta \geq 1$. *Take* $v_1 = \mathbf{E}(\epsilon_i^2)$. *With* $\lambda_{\theta,0} \gtrsim \sqrt{v_1/\varrho}$, $\lambda_{\beta,\infty} = \lambda_\beta \asymp \sqrt{\frac{v_1 \log p}{n}} \left(1 + \sqrt{\frac{\log n}{s}}\right)$, *and* $\lambda_{\theta,\infty} = \lambda_\theta \asymp \sqrt{\frac{v_1}{\log p}}$, *under a probability at least* $1 - \varrho - O(p^{-2})$ *we have:*

$$\|\tilde{\beta}\|_0 \lesssim s, \quad \|\tilde{\beta} - \beta^*\|_2 \lesssim \sqrt{\frac{v_1(s + \log n)\log p}{n}}.$$

The above result confirms that our two-stage AC-IHT algorithm remains minimax near-optimal (up to a $\log n$ term) under heavy-tailed settings, demonstrating the generality of the proposed procedure. Numerical simulations of our algorithm under heavy-tailed noise are presented in Supplementary Material S1.3, and are consistent with the theoretical result. Moreover, this theorem shows that heavy-tailed regression could be connected to the adversarial contamination framework through a truncation-based decomposition, thereby revealing the broader theoretical and practical significance of the contamination framework (1.1).

5.3 Limitations and future directions

This paper proposes an algorithmic regularization procedure that preserves signal adaptivity and achieves the strong oracle property. In our theoretical analysis, we explicitly balance optimization and statistical errors, yielding a computationally

efficient procedure with near-optimal guarantees up to logarithmic factors. However, several limitations remain, including the lack of theoretical guarantees for the adaptive tuning (used in Section 4), the remaining logarithmic gap between the upper and lower bounds, and the restriction to sub-Gaussian designs for X (discussed in Remark 2). These theoretical improvements are left for future investigation.

Supplementary Materials

The online supplementary materials contain the additional numerical experiments and all detailed proofs of our results.

Acknowledgements

The authors would like to thank the editor, associate editor, and reviewers for their helpful comments. We are also grateful to Dr. Zhifan Li and Dr. Jie Li for their constructive discussions.

References

Abramovich, F. and V. Grinshtein (2016). Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory 62*(6), 3721 – 3730.

Bellec, P. C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*.

Bhatia, K., P. Jain, P. Kamalaruban, and P. Kar (2017). Consistent robust regression. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Bhatia, K., P. Jain, and P. Kar (2015). Robust regression via hard thresholding. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 28. Curran Associates, Inc.

Blumensath, T. and M. E. Davies (2008). Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications 14*(5), 629 – 654.

Blumensath, T. and M. E. Davies (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis 27*(3), 265 – 274.

Butucea, C., M. Ndaoud, N. A. Stepanova, and A. B. Tsybakov (2018). Variable selection with Hamming loss. *The Annals of Statistics 46*(5), 1837 – 1875.

Chen, M., C. Gao, and Z. Ren (2016). A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics 10*(2), 3752 – 3774.

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.

Chen, Y., J. Fan, C. Ma, and Y. Yan (2021). Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *The Annals of statistics 49*(5), 2948 – 2971.

Chinot, G. (2020). ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electronic Journal of Statistics 14*(2), 3563 – 3605.

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.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Fan, J., H. Liu, Q. Sun, and T. Zhang (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics 46*(2), 814 – 841.

Fan, J., Z. Yang, and M. Yu (2023). Understanding implicit regularization in over-parameterized single index model. *Journal of the American Statistical Association 118*(544), 2315 – 2328.

Finocchio, G., A. Derumigny, and K. Proksch (2021). Robust-to-outliers square-root lasso, simultaneous inference with a mom approach. *arXiv preprint arXiv:2103.10420*.

Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing 17*, 293 – 310.

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli 26*(2), 1139 – 1170.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024). Contamination bias in linear regressions. *American Economic Review 114*(12), 4015 – 4051.

Hammouda, I., M. Ndaoud, and A.-K. Seghouane (2024). Outlier-bias removal with alpha divergence: A robust non-convex estimator for linear regression. *arXiv preprint arXiv:2412.19183*.

Han, R., L. Luo, Y. Luo, Y. Lin, and J. Huang (2026). Adaptive debiased lasso in high-dimensional generalized linear models with streaming data. *Journal of the American Statistical Association 0*(ja), 1 – 19.

Heng, A. and H. Soh (2025). Detecting covariate shifts with vision-language foundation models. In *ICLR*

2025 Workshop on Foundation Models in the Wild.

Huang, J., Y. Jiao, Y. Liu, and X. Lu (2018). A constructive approach to l_0 penalized regression. *Journal of Machine Learning Research 19*(10), 1 – 37.

Jain, P., A. Tewari, and P. Kar (2014). On iterative hard thresholding methods for high-dimensional m-estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 27. Curran Associates, Inc.

Kong, D., H. D. Bondell, and Y. Wu (2018). Fully efficient robust estimation, outlier detection and variable selection via penalized regression. *Statistica Sinica 28*(2), 1031 – 1052.

Lee, Y., S. N. MacEachern, and Y. Jung (2012). Regularization of Case-Specific Parameters for Robustness and Efficiency. *Statistical Science 27*(3), 350 – 372.

Liu, H. and R. Foygel Barber (2019). Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA 9*(4), 899 – 933.

Loh, P.-L. (2025). A theoretical review of modern robust statistics. *Annual Review of Statistics and Its Application 12*, 477 – 496.

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.

Minsker, S. and Y. Shen (2025). The impact of contamination and correlated design on the lasso: An average case analysis. *Statistics & Probability Letters 223*, 110417.

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.

Ndaoud, M. (2020). Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. *arXiv preprint arXiv:2008.12236*.

Nguyen, N. H. and T. D. Tran (2013). Robust lasso with missing and grossly corrupted observations. *IEEE Transactions on Information Theory 59*(4), 2036 – 2058.

Pensia, A., V. Jog, and P.-L. Loh (2025). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association 120*(550), 1002 – 1013.

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.

Sardy, S., P. Tseng, and A. Bruce (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing 49*(6), 1146 – 1152.

Sasai, T. and H. Fujisawa (2020). Robust estimation with lasso when outputs are adversarially contaminated. *arXiv preprint arXiv:2004.05990*.

Sasai, T. and H. Fujisawa (2025). Outlier robust and sparse estimation of linear regression coefficients. *Journal of Machine Learning Research 26*(93), 1 – 79.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626 – 639.

She, Y., J. Shen, and A. Barbu (2023). Slow kill for big data learning. *IEEE Transactions on Information Theory 69*(9), 5936 – 5955.

She, Y., Z. Wang, and J. Jin (2021). Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning. *The Annals of Statistics 49*(6), 3434 – 3459.

She, Y., Z. Wang, and J. Shen (2022). Gaining outlier resistance with progressive quantiles: Fast algorithms

and theoretical studies. *Journal of the American Statistical Association 117*(539), 1282 – 1295.

Shen, Y., J. Li, J.-F. Cai, and D. Xia (2025). Computationally efficient and statistically optimal robust high-dimensional linear regression. *The Annals of Statistics 53*(1), 374 – 399.

Suggala, A. S., K. Bhatia, P. Ravikumar, and P. Jain (2019). Adaptive hard thresholding for near-optimal consistent robust regression. In A. Beygelzimer and D. Hsu (Eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, Volume 99 of *Proceedings of Machine Learning Research*, pp. 2892 – 2897. PMLR.

Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive Huber regression. *Journal of the American Statistical Association 115*(529), 254 – 265.

Thompson, P. (2020). Outlier-robust sparse/low-rank least-squares regression and robust matrix completion. *arXiv preprint arXiv:2012.06750*.

Shixiang Liu

School of Statistics, Renmin University of China, Beijing, China.

E-mail: liushixiang_stat@ruc.edu.cn

Hanming Yang

Institute of Statistics and Big Data, Renmin University of China, Beijing, China.

E-mail: yanghanming@ruc.edu.cn