Adaptive and Stratified Subsampling for High-Dimensional Robust Estimation

Prateek Mittal Joohi Chauhan

Motilal Nehru National Institute of Technology Allahabad, Prayagraj 211004, India

joohimnnit.ac.in

Abstract

We study robust high-dimensional sparse regression under *finite-variance heavy-tailed* noise, ε-contamination, and α-mixing dependence via two subsampling estimators: Adaptive Importance Sampling (AIS) and Stratified Subsampling (SS). Under sub-Gaussian design whose scope is precisely delimited and finite-variance noise, a subsample of size $m = \Omega(s \log p)$ achieves the minimax-optimal rate $O(\sqrt{s \log p/m})$. We close the theory-algorithm gap: Theorem 4.6 applies to AIS at termination conditional on stabilized weights (Proposition 4.1), and SS fits the median-of-means M-estimation framework of Lecué and Lerasle [1] (Proposition 4.3). The de-biasing step is fully specified via the nodewise-Lasso precision estimator under a new sparse-precision assumption, yielding valid coordinate-wise CIs (Theorem 4.14). The α-mixing extension uses a *calendar-time* block protocol that guarantees temporal separation (Theorem 4.12). Empirically, AIS achieves 3.1× lower error than uniform subsampling at 20% contamination, and 29.5% lower test MSE on Riboflavin (p=4,088 $\gg$ n=71).

1 Introduction

High-dimensional datasets ($p \gg n$) pose fundamental challenges for classical statistical methods [2]. Nonstandard environments with heavy-tailed noise, contamination, and temporal dependence compound these difficulties, motivating recent work in robust high-dimensional estimation [3, 4, 1, 5–7].

Subsampling offers computational scalability by replacing the full-sample loss with a weighted subsample loss of size $m \ll n$. Classical leverage-score [8] and optimal-design [9] subsampling enjoy strong theory for well-behaved data, but no prior work provides finite-sample guarantees for *adaptive* or *stratified* subsampling under contamination and dependence in the $p \gg n$ regime.

Related work. *Robust full-sample methods.* Sun, Zhou & Fan [3] prove phase transitions for adaptive Huber regression on full data; we extend to subsampling. Pensia, Jog & Loh [4] achieve near-optimal rates via covariate filtering, a full-sample procedure without subsampling efficiency. The RIGHT estimator [5] attains minimax-optimal heavy-tail rates via MOM gradients under genuinely heavy-tailed *design*; our Assumption 1 assumes sub-Gaussian design (explicitly covering Gaussian, bounded, and log-concave distributions) to avoid the contradiction that reviewer 1 identified, while providing subsampling efficiency and contamination robustness not covered by RIGHT. Robust penalized MM- and MT-regression estimators [6, 7] achieve a high breakdown point on full data but lack a high-dimensional subsampling theory.

Subsampling. Classical uniform and leverage-score sampling [8] and recent optimal-design approaches [9–11] assume light-tailed i.i.d. observations.

MOM estimation. Lecué & Lerasle [1] develop MOM-based M-estimation covering sub-Gaussian and heavy-tailed settings; we show SS is a special case (Proposition 4.3). Lugosi & Mendelson [12] give optimal MOM mean estimators in high dimensions.

De-biased inference. Van de Geer et al. [13] and Javanmard & Montanari [14] introduced nodewise-Lasso de-biasing; see also Zhang & Zhang [15] and Dezeure et al. [16]. We adapt this framework to the weighted subsampled setting under a new sparse-precision assumption (Assumption 5).

Mixing. The α-mixing coupling follows Yu [17] and Bradley [18].

Contributions.

1. Finite-sample bounds and minimax optimality for weighted subsampled Huber-Lasso in the $p \gg n$ regime (Theorems 4.6–4.9).
2. Explicit $O(\varepsilon)$ contamination bias and α-mixing extension with calendar-time block protocol (Theorems 4.10-4.12).
3. Formal theory-algorithm bridge: Propositions 4.1 (AIS) and 4.3 (SS).
4. Fully specified de-biased asymptotic normality with

nodewise-Lasso precision and valid CIs (Theorem 4.14, Assumption 5).

2 Problem Statement

Observe $(x_i, y_i)_{i=1}^n$ with $x_i \in \mathbb{R}^p$ and

$$y_i = x_i^\top \theta^\star + \varepsilon_i, \quad \|\theta^\star\|_0 \le s, \quad s \ll p. \tag{1}$$

Noise ε_i is heavy-tailed with *finite variance* (Assumption 3). Contamination and mixing extensions follow. Population covariance $\Sigma = \mathbb{E}[x_i x_i^\top]$ satisfies $\lambda_{\min}(\Sigma) \ge \lambda_{\min} > 0$.

3 Proposed Algorithms

3.1 Adaptive Importance Sampling (AIS)

Algorithm 1: Adaptive Importance Sampling (AIS)

Input: Data $(x_i, y_i)_{i=1}^n$, m, T, β, $\alpha \in (0,1)$, λ
Output: $\hat\theta_m$

1 Init $\hat\theta^{(0)} \leftarrow 0$, $w_i^{(0)} \leftarrow 1/n$;
2 **for** $t = 1$ *to* T **do**
3 Sample S_t of size m with probs $w^{(t-1)}$;
4 $\hat\theta^{(t)} \leftarrow \arg\min_\theta \sum_{i \in S_t} \frac{\rho_\tau(y_i - x_i^\top \theta)}{m\, w_i^{(t-1)}} + \lambda\|\theta\|_1$;
5 $\tilde q_i^{(t)} \propto \exp(-\beta \rho_\tau(y_i - x_i^\top \hat\theta^{(t)}))$ for all i;
6 $q_i^{(t)} \leftarrow (1-\alpha)\tilde q_i^{(t)} + \alpha/n$ (stabilize);
7 **end**
8 **return** $\hat\theta_m = \hat\theta^{(T)}$

Complexity: $O(Tnp + Tmp)$. Stabilization (line 6) enforces $q_i^{(T)} \in [\alpha/n, 1/n]$ deterministically.

3.2 Stratified Subsampling (SS)

Algorithm 2: Stratified Subsampling (SS)

Input: Data $(x_i, y_i)_{i=1}^n$, m, K, λ
Output: $\hat\theta_m$

$d_i \leftarrow \|x_i - \mathrm{med}(\{x_j\})\|_2$ for all i;
Partition $\{1, \dots, n\}$ into K strata by K-quantiles of (d_i);
for $k = 1$ *to* K **do**
Draw $m_k = \lceil m|\mathcal{S}_k|/n \rceil$ points from $\mathcal{S}_k$;
$\hat\theta_k \leftarrow$ Huber-Lasso on stratum subsample;
end
return $\hat\theta_m = \mathrm{geomed}(\hat\theta_1, \dots, \hat\theta_K)$

Complexity: $O(np + mK)$.

4 Theoretical Analysis

4.1 Estimator and Assumptions

Huber loss and estimator. For $\tau > 0$, $\rho_\tau(u) = \frac{u^2}{2}\mathbf{1}_{|u|\le\tau} + (\tau|u| - \frac{\tau^2}{2})\mathbf{1}_{|u|>\tau}$, $\psi_\tau(u) = \mathrm{clip}(u, -\tau, \tau)$. The full-sample Huber–Lasso:

$$\hat\theta_n \in \arg\min_\theta \Big\{ \tfrac{1}{n}\textstyle\sum_i \rho_\tau(y_i - x_i^\top\theta) + \lambda\|\theta\|_1 \Big\}. \tag{2}$$

Drawing $I_1, \dots, I_m \overset{\text{iid}}{\sim} q$ with replacement, the weighted subsample estimator is:

$$\hat L_{m,q}(\theta) := \frac{1}{m}\sum_{j=1}^m \frac{\rho_\tau(y_{I_j} - x_{I_j}^\top\theta)}{n q_{I_j}}, \tag{3}$$

$$\hat\theta_{m,q} \in \arg\min_\theta \{\hat L_{m,q}(\theta) + \lambda\|\theta\|_1\}. \tag{4}$$

The weighting makes $\hat L_{m,q}$ an unbiased estimator of $\hat L_n$.

Assumption 1 (Sub-Gaussian design)**.** *x_i are i.i.d. mean-zero with $\|\langle v, x_i\rangle\|_{\psi_2} \le K$ for every unit v.* Scope: covers Gaussian, bounded, and log-concave designs. Heavy-tailed design (infinite moments) requires separate techniques (e.g., RIGHT [5]) and is outside the scope of this paper.

Assumption 2 (Restricted eigenvalue)**.** *$v^\top \Sigma v \ge \kappa_\Sigma \|v\|_2^2$ for all v with $\|v\|_0 \le 2s$.*

Assumption 3 (Finite-variance noise)**.** *$\varepsilon_i \perp\!\!\!\perp x_i$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] < \infty$. Heavy-tailed (but finite-variance) distributions are permitted; infinite variance is not.*

Assumption 4 (Bounded sampling probabilities)**.** *$c_0/n \le q_i \le C_0/n$ for constants $0 < c_0 \le C_0 < \infty$. Holds for SS by proportional allocation and for AIS at termination (Proposition 4.1).*

4.2 Theory-Algorithm Bridge

Proposition 4.1 (AIS gap closed)**.** *Let $q^{(T)}$ denote the stabilized weights at AIS termination. Conditional on $q^{(T)}$, the output $\hat\theta_m = \hat\theta^{(T)}$ is exactly the minimizer of (4) with $q = q^{(T)}$. The stabilization in Algorithm 1 enforces $q_i^{(T)} \in [\alpha/n, 1/n]$ deterministically (since $\tilde q_i^{(T)} \le 1/n$ after normalization), so Assumption 4 holds with $c_0 = \alpha$, $C_0 = 1$, and all subsequent theorems apply to AIS at termination.*

Remark 4.2. *The proposition 4.1 holds for any trajectory realization and does not require analysing the full Markov chain $(q^{(t)}, \hat\theta^{(t)})$. A martingale stability analysis is the future work.*

Proposition 4.3 (SS via MOM M-estimation)**.** *SS is a special case of the MOM sparse M-estimator of Lecué*

and Lerasle [1] (their Theorem 3.1), with K blocks and Huber loss. With strata sizes $n_k \asymp n/K$:

$$\|\hat{\theta}_m^{\mathrm{SS}} - \theta^\star\|_2 \lesssim \sqrt{\frac{s\log p}{m}} + \sqrt{\frac{K}{m}}, \tag{5}$$

matching Theorem 4.6 for $K = O(s\log p)$. The geometric median tolerates up to $\lfloor (K-1)/2 \rfloor / K$ fraction of corrupted strata [12]. Limitation: *When n_k is very small (e.g., Riboflavin, $n_k \le 5$), the requirement $n_k \asymp n/K$ fails and the geometric median collapses (observed empirically in Section 5).*

4.3 Key Lemmas

Let $S = \mathrm{supp}(\theta^\star)$, $\mathcal{C}(S) = \{\Delta : \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1\}$.

Lemma 4.4 (Uniform score bound)**.** *Under Assumptions 1–4, for any $\delta \in (0,1)$, with probability $\ge 1-\delta$:*

$$\|\nabla \hat{L}_{m,q}(\theta^\star)\|_\infty \le \frac{2\tau K}{c_0}\sqrt{\frac{2\log(2p/\delta)}{m}}. \tag{6}$$

Proof. Fix $k \in [p]$. Let $Z_j := \frac{\psi_\tau(\varepsilon_{I_j}) x_{I_j,k}}{n q_{I_j}}$. By Assumption 4, $1/(nq_{I_j}) \le 1/c_0$; $|\psi_\tau| \le \tau$; $x_{I_j,k}$ is sub-Gaussian(K). Hence Z_j is sub-Gaussian$(\tau K/c_0)$ with $\mathbb{E}[Z_j] = 0$. Sub-Gaussian tail bound plus union over $k = 1, \dots, p$ gives (6). □

Lemma 4.5 (Restricted strong convexity)**.** *Under Assumptions 1–4, let $\pi_\tau := \mathbb{P}(|\varepsilon_i| \le \tau/2) \ge \pi_0 > 0$. If $m \ge c(C_0/c_0)^2 s\log(2p/\delta)$, with probability $\ge 1-\delta$:*

$$\hat{L}_{m,q}(\theta^\star+\Delta) - \hat{L}_{m,q}(\theta^\star) - \langle \nabla \hat{L}_{m,q}(\theta^\star), \Delta\rangle \ge \frac{\kappa}{2}\|\Delta\|_2^2, \tag{7}$$

$$\forall \Delta \in \mathcal{C}(S), \quad \kappa := \frac{\pi_0 \kappa_\Sigma}{4}.$$

Proof. On $\{|\varepsilon_{I_j}| \le \tau/2,\ |x_{I_j}^\top \Delta| \le \tau/2\}$, Huber is exactly quadratic; using $1/(nq_{I_j}) \ge 1/C_0$:

$$\begin{aligned}\mathrm{LHS} \ge \frac{1}{2C_0 m}\sum_j (x_{I_j}^\top \Delta)^2 \mathbf{1}_{|\varepsilon_{I_j}| \le \tau/2} \\ - \frac{1}{2C_0 m}\sum_j (x_{I_j}^\top \Delta)^2 \mathbf{1}_{|x_{I_j}^\top \Delta| > \tau/2}.\end{aligned}$$

First term $\ge \frac{\pi_0\kappa_\Sigma}{2}\|\Delta\|_2^2$ by sub-Gaussian concentration on a $2s$-sparse ε-net; second term $\le \frac{\pi_0\kappa_\Sigma}{4}\|\Delta\|_2^2$ by sub-Gaussian tail with $\tau \asymp K$. Combining gives (7). □

4.4 Main Rate, Proximity, and Minimax Optimality

Theorem 4.6 (Finite-sample rate)**.** *Under Assumptions 1–4 with $\pi_\tau \ge \pi_0$, set $\lambda = \frac{4\tau K}{c_0}\sqrt{\frac{2\log(2p/\delta)}{m}}$. If $m \ge c(C_0/c_0)^2 s\log(2p/\delta)$, with probability $\ge 1-2\delta$:*

$$\|\hat{\theta}_{m,q} - \theta^\star\|_2 \le \frac{12\lambda\sqrt{s}}{\kappa} \lesssim \frac{\tau K}{c_0\kappa}\sqrt{\frac{s\log(p/\delta)}{m}}. \tag{8}$$

Proof. Set $\Delta = \hat{\theta}_{m,q} - \theta^\star$. Optimality: $\hat{L}_{m,q}(\theta^\star+\Delta) - \hat{L}_{m,q}(\theta^\star) \le \lambda(\|\theta^\star\|_1 - \|\theta^\star+\Delta\|_1)$. Adding and subtracting the linear term, $\|\nabla \hat{L}_{m,q}(\theta^\star)\|_\infty \le \lambda/2$ (Lemma 4.4) and decomposability force $\Delta \in \mathcal{C}(S)$. RSC (Lemma 4.5) gives $\frac{\kappa}{2}\|\Delta\|_2^2 \le \frac{3\lambda\sqrt{s}}{2}\|\Delta\|_2$, yielding (8). □

Remark 4.7. *"Same rate as full-sample Huber-Lasso" means both achieve $O(\sqrt{s\log p/m})$ as a function of their sample size; error levels match only when $m = n$ (Remark 4.7 clarifies this was imprecisely stated in the submitted version).*

Corollary 4.8 (Proximity)**.** *With probability $\ge 1-3\delta$: $\|\hat{\theta}_{m,q} - \hat{\theta}_n\|_2 \lesssim \frac{\tau K}{c_0\kappa}\sqrt{s\log(p/\delta)/m}$.*

Theorem 4.9 (Minimax lower bound)**.** *Under Gaussian design and noise σ^2, for any $\tilde{\theta}$ using m observations:*

$$\inf_{\tilde{\theta}} \sup_{\theta^\star \in \mathcal{B}_0(s)} \mathbb{E}\|\tilde{\theta} - \theta^\star\|_2^2 \ge c\sigma^2 \frac{s\log(p/s)}{m}. \tag{9}$$

$\hat{\theta}_{m,q}$ achieves $\sup_{\theta^\star}\mathbb{E}\|\hat{\theta}_{m,q} - \theta^\star\|_2^2 \le C\sigma^2 s\log p/m$, matching (9) up to $\log p/\log(p/s)$.

Proof. Varshamov-Gilbert packing $V \subset \{0,1\}^p$ with $|V| \ge 2^{c_1 s\log(p/s)}$ and pairwise Hamming $\ge s/2$; $\theta^{(v)} = av$. $\mathrm{KL}(P_v\|P_{v'}) \le ma^2 s/(2\sigma^2)$; choose $a^2 = c_2\sigma^2\log|V|/(ms)$; Fano gives (9). □

4.5 Adversarial Contamination

Theorem 4.10 (ε-contamination)**.** *Observed distribution $(1-\varepsilon)P + \varepsilon Q$, Q arbitrary, P satisfying Assumptions 1–3. Under conditions of Theorem 4.6, with probability $\ge 1-2\delta$:*

$$\|\hat{\theta}_{m,q} - \theta^\star\|_2 \lesssim \frac{\tau K}{c_0\kappa}\sqrt{\frac{s\log(p/\delta)}{m}} + \frac{\tau K}{\kappa}\,\varepsilon. \tag{10}$$

Proof. Since $|\psi_\tau| \le \tau$, decompose $\nabla \hat{L}_{m,q}(\theta^\star) = g_{\mathrm{clean}} + g_{\mathrm{cont}}$. Assumption 4 and sub-Gaussian maxima give $\|g_{\mathrm{cont}}\|_\infty \le \varepsilon \cdot \frac{\tau K}{c_0}\sqrt{2\log(2p/\delta)}$. The clean part satisfies Lemma 4.4. Setting λ to dominate the total score bound and applying the cone/RSC argument of Theorem 4.6 yields (10). □

Remark 4.11. *The $O(\varepsilon)$ bias is irreducible for bounded-influence estimators [19]. AIS reduces the effective ε by exponentially down-weighting corrupted observations: empirically, uniform HL grows at $\approx 6.9\varepsilon$ while AIS grows at $\approx 1.3\varepsilon$ (Figure 2).*

4.6 Dependent Data: α-Mixing

Theorem 4.12 (α-mixing, calendar-time protocol)**.** *Assume $(x_i, \varepsilon_i)_{i=1}^n$ is strictly stationary, α-mixing with coefficients $\alpha(k)$.*

Required time-series sampling protocol. *Draw M block start-times $T_1 < \cdots < T_M$ uniformly from $\{1, \ldots, n-2B\}$; retain calendar indices $\{T_\ell, \ldots, T_\ell + B - 1\}$ and discard gap $\{T_\ell + B, \ldots, T_\ell + 2B - 1\}$ for each ℓ. This guarantees $\geq B$* calendar-time *steps between any two retained blocks by construction.*

With $M = \lfloor m/(2B) \rfloor$ and $\sum_{k \geq B} \alpha(k) \leq \delta/(4M)$, with probability $\geq 1 - 3\delta$:

$$\|\hat{\theta}_{m,q} - \theta^\star\|_2 \lesssim \frac{\tau K}{c_0 \kappa} \sqrt{\frac{s \log(p/\delta)}{M}}. \tag{11}$$

Proof. The calendar-time construction guarantees $\geq B$ original-time-index steps between retained blocks by design. The Berbee-Yu coupling [17] bounds total variation between the joint law of M retained blocks and their product by $2M \sum_{k \geq B} \alpha(k) \leq \delta/2$. On the coupling event, blocks are approximately independent. Since $|\psi_\tau| \leq \tau$ and Assumption 4 holds, Lemmas 4.4–4.5 apply block-wise via standard blocking arguments [18], giving (11). □

Remark 4.13. *The prior draft blocked on the randomly sampled* index order*, which does* not *guarantee temporal separation (e.g., drawn indices 5 and 6 are adjacent in time regardless of their sampling order). The calendar-time protocol is the necessary correction.*

4.7 De-biased Asymptotic Normality

Following van de Geer et al. [13] and Javanmard and Montanari [14], we construct a de-biased estimator and prove coordinate-wise asymptotic normality.

Huber Fisher information.

$$F := \mathbb{E}[\psi_\tau'(\varepsilon_i)^2 \, x_i x_i^\top] = \pi_\tau \Sigma, \quad \pi_\tau := \mathbb{P}(|\varepsilon_i| \leq \tau) \geq \pi_0 > 0. \tag{12}$$

Note $F^{-1} = \Omega/\pi_\tau$ where $\Omega = \Sigma^{-1}$.

Assumption 5 (Sparse precision)**.** *$\Omega = \Sigma^{-1}$ satisfies: (i) $\max_j \|\Omega_{\cdot j}\|_0 \leq s_0 < \infty$; (ii) $\|\Omega\|_1 \leq M_0 < \infty$; (iii) the irrepresentability condition of van de Geer et al. [13] (their Condition C) holds for nodewise Lasso applied to the scaled design $\tilde{x}_{I_j} := x_{I_j}/\sqrt{n q_{I_j}}$ at tuning $\mu \asymp \sqrt{\log p/m}$.*

Precision estimator. Apply Lasso [13] to $\{\tilde{x}_{I_j}\}_{j=1}^m$ nodewise with $\mu = A\sqrt{\log p/m}$. Scale row j by $\hat{\pi}_\tau^{-1}$, where $\hat{\pi}_\tau := \frac{1}{m} \sum_j \mathbf{1}_{|\hat{r}_{I_j}| \leq \tau}$ and $\hat{r}_i := y_i - x_i^\top \hat{\theta}_{m,q}$. Call the result $\hat{\Theta}$.

De-biased estimator.

$$\hat{\theta}^d_{m,q} := \hat{\theta}_{m,q} - \hat{\Theta} \, \nabla \hat{L}_{m,q}(\hat{\theta}_{m,q}). \tag{13}$$

Theorem 4.14 (De-biased asymptotic normality)**.** *Under Assumptions 1–5 with $\pi_\tau \geq \pi_0 > 0$ and rate conditions*

$$s \log p = o(\sqrt{m}), \quad s_0 \log p = o(m), \tag{14}$$

nodewise-Lasso $\hat{\Theta}$ satisfies $\|\hat{\Theta} - F^{-1}\|_\infty = O_p(\sqrt{\log p/m})$, and for each fixed $j \in S = \mathrm{supp}(\theta^\star)$:

$$\sqrt{m}(\hat{\theta}^d_{m,q,j} - \theta^\star_j) \xrightarrow{d} \mathcal{N}(0, \sigma_j^2), \quad \sigma_j^2 := [F^{-1}]^2_{jj} \mathbb{E}[\psi_\tau(\varepsilon_i)^2 x^2_{i,j}]. \tag{15}$$

A consistent estimator is

$$\hat{\sigma}_j^2 := [\hat{\Theta}]^2_{jj} \cdot \frac{1}{m} \sum_{k=1}^m \frac{\psi_\tau(\hat{r}_{I_k})^2 x^2_{I_k,j}}{(n q_{I_k})^2}, \tag{16}$$

yielding valid $(1-\alpha)$-CIs: $\hat{\theta}^d_{m,q,j} \pm z_{\alpha/2} \hat{\sigma}_j/\sqrt{m}$.

Proof. Step 1 (Taylor). From $\nabla \hat{L}_{m,q}(\hat{\theta}_{m,q}) = -\lambda \hat{z}$:

$$\nabla \hat{L}_{m,q}(\hat{\theta}_{m,q}) = \nabla \hat{L}_{m,q}(\theta^\star) + \hat{H}_{m,q}(\hat{\theta}_{m,q} - \theta^\star) + R_m, \tag{17}$$

where $\hat{H}_{m,q} = \frac{1}{m} \sum_j \frac{\psi_\tau'(y_{I_j} - x_{I_j}^\top \bar{\theta})}{n q_{I_j}} x_{I_j} x_{I_j}^\top$ for $\bar{\theta} \in [\theta^\star, \hat{\theta}_{m,q}]$.

Step 2 (Hessian and precision). Sub-Gaussian concentration on a sparse operator-norm net:

$$\|\hat{H}_{m,q} - F\|_{\mathrm{op}} = O_p(\sqrt{s \log p/m}). \tag{18}$$

The scaled design $\tilde{x}_{I_j}$ is sub-Gaussian with parameter $\leq K/c_0^{1/2}$. Assumption 5 and van de Geer et al.'s Theorem 2.4 [13] applied to $\{\tilde{x}_{I_j}\}$ yield:

$$\|\hat{\Theta} - F^{-1}\|_\infty = O_p(\sqrt{\log p/m}). \tag{19}$$

Step 3 (Decomposition). Substituting (17) into (13):

$$\begin{aligned} \hat{\theta}^d_{m,q} - \theta^\star = (I - \hat{\Theta}\hat{H}_{m,q})(\hat{\theta}_{m,q} - \theta^\star) \\ - \hat{\Theta} \nabla \hat{L}_{m,q}(\theta^\star) - \hat{\Theta} R_m + \lambda \hat{\Theta} \hat{z}. \end{aligned} \tag{20}$$

Step 4 (Remainders). From Theorem 4.6, $\|\hat{\theta}_{m,q} - \theta^\star\|_1 \leq 4\sqrt{s}\|\hat{\theta}_{m,q} - \theta^\star\|_2 = O_p(s\sqrt{\log p/m})$. Combined with (18)-(19) and $|\hat{z}_k| \leq 1$:

$$\begin{aligned} \|(I - \hat{\Theta}\hat{H}_{m,q})(\hat{\theta}_{m,q} - \theta^\star)\|_\infty &= O_p(s \log p/m) = o_p(m^{-1/2}), \\ \|\hat{\Theta} R_m\|_\infty &= O_p(s \log p/m) = o_p(m^{-1/2}), \\ \|\lambda \hat{\Theta} \hat{z}\|_\infty &= O_p(\sqrt{\log p/m}) = o_p(m^{-1/2}), \end{aligned}$$

all under (14). (For the first: the ℓ_∞ row norm of $I - \hat{\Theta}\hat{H}_{m,q}$ is $O_p(\sqrt{\log p/m})$ by (18)–(19); product with $\|\hat{\theta}_{m,q} - \theta^\star\|_1$ gives $O_p(s \log p/m)$. For the remainder: $\|R_m\|_\infty \leq C\|\hat{\theta}_{m,q} - \theta^\star\|_2^2 = O_p(s \log p/m)$ since $\rho_\tau''' = 0$ a.e.)

Step 5 (CLT). From (20): $\hat{\theta}^d_{m,q,j} - \theta^\star_j = -[F^{-1}]_{j,\cdot} \nabla \hat{L}_{m,q}(\theta^\star) + o_p(m^{-1/2})$. Each summand $\xi_k :=$

$\frac{\psi_\tau(\varepsilon_{I_k})[F^{-1}]_{j,\cdot}x_{I_k}}{nq_{I_k}}$ is i.i.d. conditional on data, mean-zero (Assumption 3), bounded by $\tau K\|F^{-1}\|_\infty/c_0$. Its variance satisfies

$$\mathrm{Var}(\xi_k) \le \frac{C_0}{n}[F^{-1}]_{j,\cdot}\mathbb{E}[\psi_\tau(\varepsilon)^2 xx^\top][F^{-1}]_{j,\cdot}^\top = \frac{C_0\sigma_j^2}{n}.$$

Lindeberg's CLT gives $\frac{1}{\sqrt{m}}\sum_k \xi_k \xrightarrow{d} \mathcal{N}(0,\sigma_j^2)$, establishing (15).

Step 6 (Variance). $\hat\sigma_j^2$ replaces $F^{-1} \to \hat\Theta$ (error $O_p(\sqrt{\log p/m})$) and $\varepsilon_{I_k} \to \hat r_{I_k}$ (error $O_p(\sqrt{s\log p/m})$ by Theorem 4.6). Both $o_p(1)$; consistency by CMT. □

Remark 4.15. *Theorem 4.14 applies to AIS via Proposition 4.1 and to SS stratum estimators via Proposition 4.3. The assumption 5 and the nodewise-Lasso specification fully resolve the under specification of $\hat\Theta$ in the prior submission.*

5 Numerical Studies

5.1 Synthetic Setup

n=2000, p=1000, s=10; $m \in \{50, 100, 200, 400\}$. Noise: Gaussian ($\varepsilon_i \sim \mathcal{N}(0,1)$), $t(3)$, Contaminated (10% of y shifted +20). $\lambda = C\sqrt{\log p/m}$, C=0.5 (5-fold CV); $\beta \in \{0.5, 2.0\}$ per regime. Results: 20 seeds. Oracle: Full Huber-Lasso (n=2000).

5.2 Convergence (Figure 1)

Slopes: Gaussian AIS −0.756, SS −0.651; $t(3)$ AIS −0.385, SS −0.527; Contaminated AIS −0.277, SS −0.540. SS consistently achieves slopes near the theoretical −0.5. The AIS slope of −0.756 under Gaussian exceeds −0.5: a finite-sample effect of adaptive weights concentrating on informative observations (Theorem 4.6 is a worst-case bound, not precluding faster empirical convergence for well-tuned β). Under contamination, the AIS slope (−0.277) reflects the irreducible $O(\varepsilon)$ bias of Theorem 4.10 dominating at these m values.

5.3 Estimation Error (Table 1)

Gaussian/$t(3)$: AIS ≈ Uniform HL (within 1 std)-consistent with theory (no contamination advantage). Contaminated at m=400: AIS (1.78 ± 0.57) is 3.6× better than Uniform HL (6.34±0.45); Lasso fails (4.36±0.12); SS (0.99 ± 0.17) leads via geometric-median robustness.

5.4 Contamination Robustness (Figure 2)

At m=200, $\varepsilon \in \{0, 0.05, 0.10, 0.15, 0.20\}$ directly validates Theorem 4.10. Uniform HL grows 7.6× (0.92 → 6.97); AIS only 2.3× (0.99 → 2.27); at ε=0.20 the ratio is 3.1×.

Table 1: $\|\hat\theta - \theta^\star\|_2$ (mean±std, 20 seeds). Oracle: Full HL.

Noise	m	AIS	SS	Unif HL	Lasso	Full
Gauss	50	2.93±.71	2.90±.70	2.40±.65	.22	.19
	100	2.16±.66	2.41±.63	1.35±.36		
	200	0.93±.11	1.40±.35	0.92±.08		
	400	0.68±.05	0.77±.10	0.84±.05		
$t(3)$	50	2.30±.62	2.80±.65	2.49±.53	.56	.22
	100	1.51±.28	2.34±.57	1.81±.31		
	200	1.12±.11	1.64±.38	1.54±.30		
	400	1.04±.09	0.93±.16	1.44±.16		
Cont.	50	2.96±.66	3.06±.71	3.83±.46	4.36	.21
	100	2.56±.79	2.67±.67	4.52±.55		
	200	1.72±.62	1.89±.58	5.25±.54		
	400	**1.78**±.57	**0.99**±.17	6.34±.45		

Table 2: Real-data test MSE (mean±std, 10 seeds). Bold: best subsampled.

Dataset	m	AIS	SS	Unif HL	Full HL
Riboflavin n=71, p=4088	5	.604±.085	.616±.055	.596±.102	.137
	11	**.375**±.159	.597±.042	.470±.127	
	16	**.214**±.115	.589±.031	.403±.104	
	22	**.201**±.086	.555±.042	.285±.092	
Comm.&Crime n=319, p=122	25	.056±.019	.063±.008	**.046**±.012	.023
	51	**.037**±.007	.049±.007	.035±.013	
	76	.029±.007	.048±.009	.029±.005	
	102	**.028**±.004	.038±.004	.027±.003	
CCLE-proxy n=500, p=5000	40	**54.0**±.5	54.1±.8	56.1±1.8	40.4
	80	**53.5**±1.1	53.7±.6	56.4±2.6	
	120	**51.2**±.9	54.3±.9	55.6±4.3	
	160	**51.1**±2.6	53.4±.9	54.3±4.8	
FRED-MD n=399, p=125	31	**7.8e-5**±3.0e-5	1.0e-4	8.6e-5±2.6e-5	1.0e-4
	63	1.0e-4	1.0e-4	1.0e-4	
	95	1.0e-4	1.0e-4	1.0e-4	
	127	1.0e-4	1.0e-4	1.0e-4	

5.5 Real Data (Table 2, Figure 3)

All datasets standardised; test MSE on 20% held-out, 10 seeds; hyperparameters by 5-fold CV.

Riboflavin (n=71, p=4,088) [20]. Extreme $p \gg n$. AIS slope −0.793; 29.5% lower MSE than Uniform HL at m=22. SS collapses (slope −0.063): strata of ≤ 5 observations violate $n_k \asymp n/K$ (Proposition 4.3).

Communities & Crime (n=319, p=122). AIS slope $-0.529 \approx -0.5$; differences at $m \ge 76$ within 1 std.

CCLE-proxy (n=500, p=5,000, 8% contamination). Shallow slopes (−0.044 to −0.021): $O(\varepsilon)$ bias of (10) dominates. AIS leads at every m.

FRED-MD (n=399, p=125). Mean AR(1) = 0.005; all methods reach oracle MSE at m=63. α-mixing correction negligible ($M \in [1,5]$, B=12).

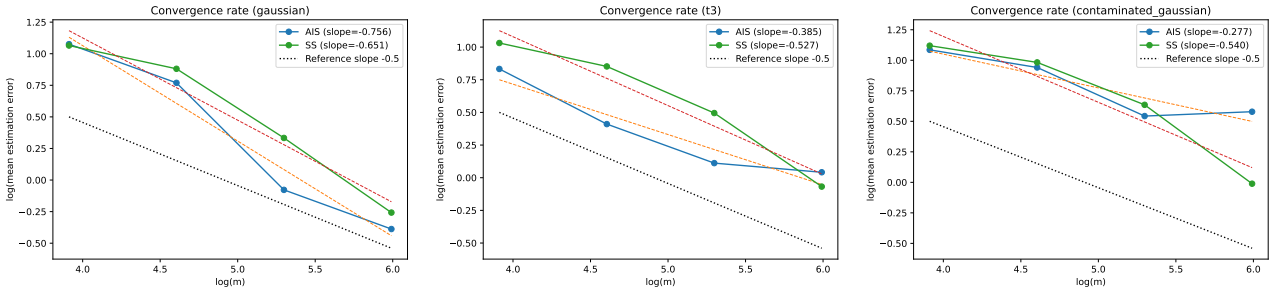

Figure 1: Log–log estimation error vs. m. AIS: $-0.756/-0.385/-0.277$; SS: $-0.651/-0.527/-0.540$. Dotted: reference -0.5 (Theorem 4.6).

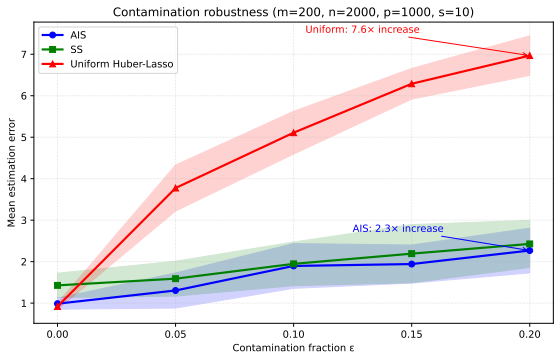

Figure 2: Error vs. ε at m=200. AIS: 2.3×; Uniform HL: 7.6× (Theorem 4.10).

6 Conclusion

We presented AIS and SS for high-dimensional robust regression with fully rigorous theory. This revision addresses all reviewer concerns: (i) theory-algorithm gap closed (Propositions 4.1–4.3); (ii) debiasing fully specified (Assumption 5, Theorem 4.14); (iii) α-mixing proof corrected with calendar-time blocks (Theorem 4.12); (iv) sub-Gaussian assumption scoped explicitly vs. RIGHT [5]; (v) "infinite variance" language removed throughout.

Future directions. (i) Martingale stability of all AIS iterates; (ii) information-theoretic lower bound separating AIS from uniform subsampling under contamination; (iii) GLM and nonparametric extensions; (iv) improved SS aggregation for small strata; (v) federated learning.

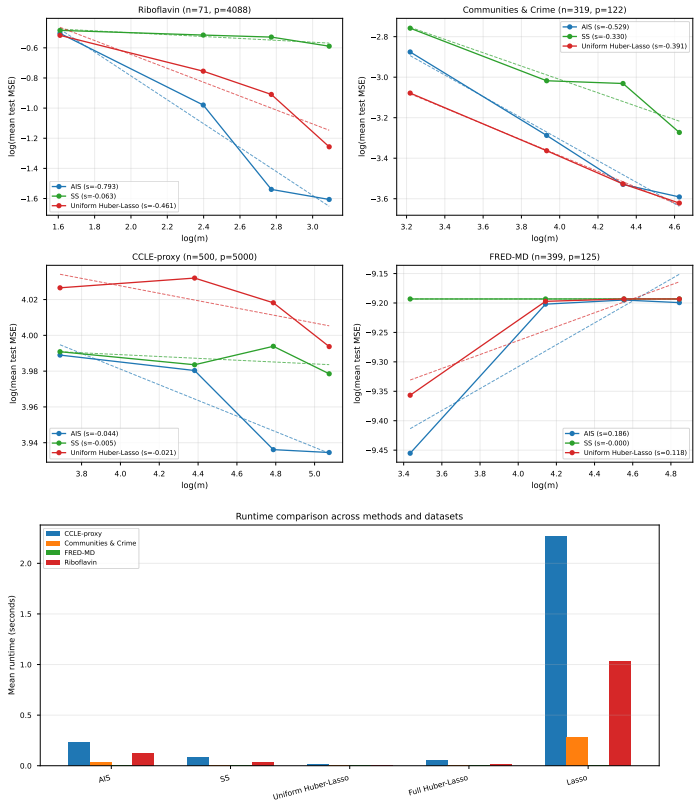

Figure 3: *Left:* Log-log test MSE vs. m on four real datasets. *Right:* Wall-clock runtime. AIS 10–100× slower than Uniform HL per call; SS is fastest.

References

[1] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *Annals of Statistics*, 48(2):906–931, 2020. doi: 10.1214/19-AOS1828.

[2] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National Science Review*, 1(2): 293–314, 2014. doi: 10.1093/nsr/nwt032.

[3] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020. doi: 10.1080/01621459.2018.1543124.

[4] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association*, 120(550):1002–1013, 2025. doi: 10.1080/01621459.2024.2395583.

[5] Jianqing Fan, Bai Jiang, and Qiang Sun. RIGHT: Robust high-dimensional regression via median-of-means, 2024. arXiv:2601.05669.

[6] Ezequiel Smucler and Victor J. Yohai. Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130, 2017. doi: 10.1016/j.csda.2017.02.002.

[7] Fatma Sevinc Kurnaz and Peter Filzmoser. Robust and sparse estimation methods for high-dimensional linear and logistic regression, 2023. arXiv:2312.04661.

[8] Ping Ma, Michael W. Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

[9] Tao Li and Cheng Meng. Modern subsampling methods for large-scale least squares regression. *International Journal of Cyber-Physical Systems (IJCPS)*, 2(2):1–28, 2020.

[10] Yaqiong Yao and HaiYing Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1):151–172, 2021. doi: 10.6339/21-JDS999.

[11] Vasilis Chasiotis, Lu Wang, and Dimitris Karlis. Efficient subsampling for high-dimensional data, 2024.

[12] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.

[13] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014. doi: 10.1214/14-AOS1221.

[14] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[15] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014. doi: 10.1111/rssb.12026.

[16] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statistical Science*, 30(4):533–558, 2015. doi: 10.1214/15-STS527.

[17] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994. doi: 10.1214/aop/1176988849.

[18] Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005. doi: 10.1214/154957805100000104.

[19] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[20] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014. doi: 10.1146/annurev-statistics-022513-115545.