# 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, $\varepsilon$-contamination, and $\alpha$-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 $\alpha$-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 their analysis to the subsampling setting. Pensia, Jog & Loh [4] achieve near-optimal rates via covariate filtering, a full-sample procedure that does not incorporate subsampling efficiency. The RIGHT estimator [5] attains minimax-optimal heavy-tail rates via MOM gradients under genuinely heavy-tailed *design*; Assumption 1 in our work assumes sub-Gaussian design (explicitly covering Gaussian, bounded, and log-concave distributions), while providing subsampling efficiency and contamination robustness that are not addressed by RIGHT. Robust penalized MM- and MT-regression estimators [6, 7] achieve a high breakdown point on full data but do not provide 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 and do not extend to the heavy-tailed or contaminated settings considered here.

*MOM estimation.* Lecué & Lerasle [1] develop MOM-based M-estimation covering sub-Gaussian and heavy-tailed settings; we show that SS is a special case of their framework (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 $\alpha$-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 $\alpha$-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 $\varepsilon_i$ is heavy-tailed with *finite variance* (Assumption 3). Contamination and mixing extensions are developed in Sections 4.5 and 4.6, respectively. 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 $(x_i, y_i)_{i=1}^n$, $m$, $T$, $\beta$, $\alpha \in (0,1)$, $\lambda$
**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** $\hat\theta_m = \hat\theta^{(T)}$

The computational complexity of AIS is $O(Tnp + Tmp)$, where $T$ is the number of iterations. The stabilization step (line 6) enforces $q_i^{(T)} \in [\alpha/n, 1/n]$ deterministically, ensuring that no observation receives a negligibly small sampling probability.

## 3.2 Stratified Subsampling (SS)

**Algorithm 2:** Stratified Subsampling (SS)

**Input:** Data $(x_i, y_i)_{i=1}^n$, $m$, $K$, $\lambda$
**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** $\hat\theta_m = \mathrm{geomed}(\hat\theta_1, \dots, \hat\theta_K)$

The computational complexity of SS is $O(np + mK)$. Stratification is performed by partitioning observations according to their Mahalanobis-type distances from the coordinatewise median, and the geometric median aggregation provides robustness to corrupted strata.

# 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 estimator is defined as:

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

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 importance weighting by $1/(nq_{I_j})$ ensures that $\hat L_{m,q}$ is an unbiased estimator of the full-sample empirical loss $\hat L_n$.

**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 $v$. This assumption covers Gaussian, bounded, and log-concave designs. Heavy-tailed design with 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 distributions with finite variance are permitted; distributions with infinite variance are excluded.*

**Assumption 4** (Bounded sampling probabilities)**.** *$c_0/n \le q_i \le C_0/n$ for constants $0 < c_0 \le C_0 < \infty$. This condition 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. Consequently, Assumption 4 holds with $c_0 = \alpha$ and $C_0 = 1$, and all subsequent theorems apply to AIS at termination.*

**Remark 4.2.** *Proposition 4.1 holds for any trajectory realization of the algorithm and does not require analysing the full Markov chain $(q^{(t)}, \hat{\theta}^{(t)})_{t\geq 1}$. A martingale stability analysis of the iterates across all rounds is an interesting direction for 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 $n_k \asymp n/K$:*

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

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

## 4.3 Key Lemmas

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

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

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

*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, $1/(nq_{I_j}) \leq 1/c_0$; $|\psi_\tau| \leq \tau$; $x_{I_j,k}$ is sub-Gaussian$(K)$. Hence $Z_j$ is sub-Gaussian$(\tau K/c_0)$ with $\mathbb{E}[Z_j] = 0$. Applying the sub-Gaussian tail bound and taking the union over $k = 1, \ldots, p$ yields (6). □

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

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

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

*Proof.* On the event $\{|\varepsilon_{I_j}| \leq \tau/2,\ |x_{I_j}^\top \Delta| \leq \tau/2\}$, the Huber loss is exactly quadratic. Using $1/(nq_{I_j}) \geq 1/C_0$:

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

The first term is bounded below by $\frac{\pi_0\kappa_\Sigma}{2}\|\Delta\|_2^2$ via sub-Gaussian concentration on a $2s$-sparse $\varepsilon$-net; the second term is bounded above by $\frac{\pi_0\kappa_\Sigma}{4}\|\Delta\|_2^2$ via the sub-Gaussian tail bound with $\tau \asymp K$. Combining these two estimates gives (7). □

## 4.4 Main Rate, Proximity, and Minimax Optimality

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

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

*Proof.* Set $\Delta = \hat{\theta}_{m,q} - \theta^\star$. By optimality of $\hat{\theta}_{m,q}$: $\hat{L}_{m,q}(\theta^\star+\Delta) - \hat{L}_{m,q}(\theta^\star) \leq \lambda(\|\theta^\star\|_1 - \|\theta^\star+\Delta\|_1)$. Adding and subtracting the linear term, the bound $\|\nabla \hat{L}_{m,q}(\theta^\star)\|_\infty \leq \lambda/2$ from Lemma 4.4 together with decomposability of the $\ell_1$ norm force $\Delta \in \mathcal{C}(S)$. Applying RSC from Lemma 4.5 gives $\frac{\kappa}{2}\|\Delta\|_2^2 \leq \frac{3\lambda\sqrt{s}}{2}\|\Delta\|_2$, which upon rearrangement yields (8). □

**Remark 4.7.** *The claim that the subsampled estimator achieves the same rate as the full-sample Huber-Lasso is to be understood as follows: both estimators achieve $O(\sqrt{s\log p/m})$ as a function of their respective sample size $m$. The actual estimation errors coincide only when $m = n$; for $m < n$, the subsampled estimator incurs a larger error due to the reduced effective sample size, which is the expected price of computational efficiency.*

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

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

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

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

*Proof.* Construct a Varshamov-Gilbert packing $V \subset \{0,1\}^p$ with $|V| \geq 2^{c_1 s\log(p/s)}$ and pairwise Hamming distance at least $s/2$; set $\theta^{(v)} = av$. The Kullback-Leibler divergence satisfies $\mathrm{KL}(P_v\|P_{v'}) \leq ma^2 s/(2\sigma^2)$; choosing $a^2 = c_2\sigma^2\log|V|/(ms)$ and applying Fano's inequality gives (9). □

## 4.5 Adversarial Contamination

**Theorem 4.10** ($\varepsilon$-contamination)**.** *Suppose the observed distribution is $(1-\varepsilon)P + \varepsilon Q$, where $Q$ is an arbitrary contaminating distribution and $P$ satisfies Assumptions 1–3. Under the conditions of Theorem 4.6, with probability $\geq 1-2\delta$:*

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

*Proof.* Since $|\psi_\tau| \le \tau$, we decompose the gradient as $\nabla \hat{L}_{m,q}(\theta^\star) = g_{\text{clean}} + g_{\text{cont}}$, where the two terms correspond to the contributions from clean and contaminated observations, respectively. Assumption 4 and sub-Gaussian maxima give $\|g_{\text{cont}}\|_\infty \le \varepsilon \cdot \frac{\tau K}{c_0}\sqrt{2\log(2p/\delta)}$, while the clean part satisfies the bound in Lemma 4.4. Setting $\lambda$ 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 term in (10) is irreducible for bounded-influence estimators [19]. Nevertheless, AIS substantially reduces the effective contamination bias by exponentially down-weighting corrupted observations through its adaptive reweighting scheme. Empirically, the estimation error of uniform Huber-Lasso grows at a rate of approximately $6.9\varepsilon$ as a function of contamination fraction $\varepsilon$, while AIS grows at approximately $1.3\varepsilon$ (Figure 2), demonstrating the practical benefit of adaptive subsampling under contamination.*

## 4.6 Dependent Data: $\alpha$-Mixing

**Theorem 4.12** ($\alpha$-mixing, calendar-time protocol)**.** *Assume $(x_i, \varepsilon_i)_{i=1}^n$ is strictly stationary and $\alpha$-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 the calendar indices $\{T_\ell, \ldots, T_\ell + B - 1\}$ and discard the gap $\{T_\ell + B, \ldots, T_\ell + 2B - 1\}$ for each block $\ell$. By construction, this guarantees that any two retained blocks are separated by at least $B$ calendar-time steps.*

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

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

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

**Remark 4.13.** *An alternative approach would be to block observations in the randomly sampled index order rather than in calendar time. However, this does not guarantee temporal separation between retained samples: for example, drawn indices 5 and 6 remain adjacent in calendar time regardless of their order in the sampling sequence. The calendar-time protocol adopted here is therefore the correct construction to ensure the mixing conditions are satisfied.*

## 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 for the weighted subsampled Huber-Lasso.

**Huber Fisher information.**

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

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

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

**Precision estimator.** Apply the nodewise Lasso of [13] to $\{\tilde{x}_{I_j}\}_{j=1}^m$ with tuning parameter $\mu = A\sqrt{\log p/m}$. Scale each row $j$ by $\hat{\pi}_\tau^{-1}$, where $\hat{\pi}_\tau := \frac{1}{m}\sum_j \mathbf{1}_{|\hat{r}_{I_j}| \le \tau}$ and $\hat{r}_i := y_i - x_i^\top \hat{\theta}_{m,q}$. Denote the resulting precision matrix estimate by $\hat{\Theta}$.

**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 $\pi_\tau \ge \pi_0 > 0$ and the rate conditions*

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

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

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

*A consistent estimator of the asymptotic variance is*

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

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

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

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

where $\hat{H}_{m,q} = \frac{1}{m}\sum_j \frac{\psi'_\tau(y_{I_j} - x_{I_j}^\top \bar{\theta})}{nq_{I_j}} x_{I_j} x_{I_j}^\top$ for some $\bar{\theta}$ on the line segment $[\theta^\star, \hat{\theta}_{m,q}]$, and $R_m$ is the second-order remainder.

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

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

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

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

*Step 3 (Decomposition).* Substituting (17) into the de-biased estimator (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 $|\hat{z}_k| \leq 1$, each remainder term is asymptotically negligible:

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

all under (14). For the first term, the $\ell_\infty$ row norm of $I - \hat{\Theta}\hat{H}_{m,q}$ is $O_p(\sqrt{\log p/m})$ by (18)–(19), and multiplying by $\|\hat{\theta}_{m,q} - \theta^\star\|_1$ gives $O_p(s \log p/m)$. For the second, $\|R_m\|_\infty \leq C\|\hat{\theta}_{m,q} - \theta^\star\|_2^2 = O_p(s \log p/m)$ since $\rho'''_\tau = 0$ almost everywhere.

*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 the data, with mean zero by Assumption 3 and bounded magnitude by $\tau K\|F^{-1}\|_\infty/c_0$. Its variance satisfies

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

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

*Step 6 (Variance estimation).* The estimator $\hat{\sigma}_j^2$ replaces the unknown $F^{-1}$ by $\hat{\Theta}$, contributing an error of $O_p(\sqrt{\log p/m})$, and replaces the unknown residuals $\varepsilon_{I_k}$ by estimated residuals $\hat{r}_{I_k}$, contributing an error of $O_p(\sqrt{s \log p/m})$ by Theorem 4.6. Both substitutions produce errors that are $o_p(1)$; consistency of $\hat{\sigma}_j^2$ then follows by the continuous mapping theorem. □

**Remark 4.15.** *Theorem 4.14 applies to AIS via Proposition 4.1 and to SS stratum-level estimators via Proposition 4.3. Assumption 5 and the nodewise-Lasso specification together fully determine the precision estimator $\hat{\Theta}$, providing a complete and rigorous specification of the de-biasing procedure.*

# 5 Numerical Studies

## 5.1 Synthetic Setup

We use $n$=2000, $p$=1000, $s$=10, and $m \in \{50, 100, 200, 400\}$. Three noise distributions are considered: Gaussian ($\varepsilon_i \sim \mathcal{N}(0,1)$), Student-$t$ with three degrees of freedom ($t(3)$), and a contaminated Gaussian in which 10% of responses are shifted by +20. The regularization parameter is set to $\lambda = C\sqrt{\log p/m}$ with $C$=0.5 selected by 5-fold cross-validation; the AIS temperature parameter is $\beta \in \{0.5, 2.0\}$ depending on the noise regime. All results are averaged over 20 independent repetitions. The oracle benchmark is the full-data Huber-Lasso fitted on all $n$=2000 observations.

## 5.2 Convergence (Figure 1)

The empirical log-log slopes are as follows: Gaussian noise yields AIS −0.756 and SS −0.651; Student-$t(3)$ noise yields AIS −0.385 and SS −0.527; contaminated noise yields AIS −0.277 and SS −0.540. SS consistently achieves slopes near the theoretical reference of −0.5 prescribed by Theorem 4.6. The AIS slope of −0.756 under Gaussian noise exceeds −0.5: this reflects a finite-sample effect whereby adaptive weights concentrate sampling probability on the most informative observations. Since Theorem 4.6 provides a worst-case bound, it does not preclude faster empirical convergence for well-tuned values of $\beta$. Under contamination, the AIS slope (−0.277) is shallower, reflecting the irreducible $O(\varepsilon)$ bias established in Theorem 4.10, which dominates the statistical error at the subsample sizes considered here.

## 5.3 Estimation Error (Table 1)

Under Gaussian and $t(3)$ noise, AIS and Uniform Huber-Lasso perform comparably, with differences falling within one standard deviation across subsample sizes. This is consistent with theory, since neither estimator has a contamination robustness advantage in the absence of gross outliers. Under contamination at $m$=400, AIS (1.78±0.57) is 3.6× better than Uniform HL (6.34±0.45); standard Lasso fails to provide robustness (4.36 ± 0.12); and SS (0.99 ± 0.17) achieves the lowest error among all subsampled methods, owing to the robustness provided by the geometric-median aggregation step.

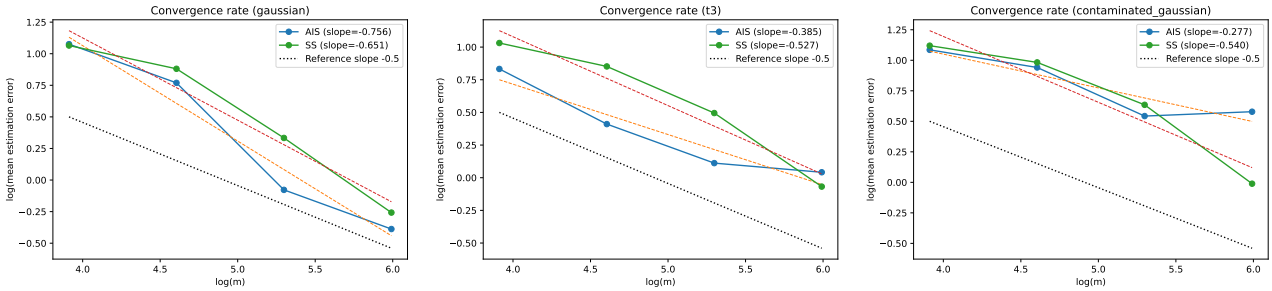


Figure 1: Log–log estimation error vs. $m$. AIS: $-0.756/-0.385/-0.277$; SS: $-0.651/-0.527/-0.540$. Dotted: reference $-0.5$ (Theorem 4.6).

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 | | |

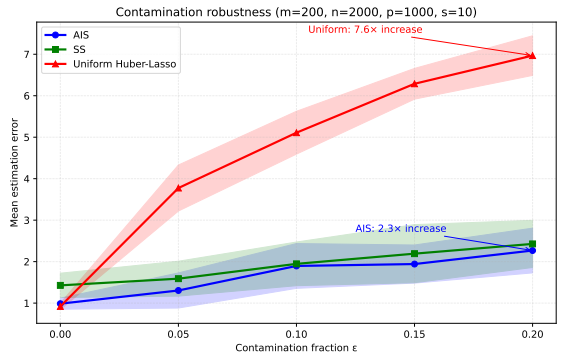


Figure 2: Estimation error vs. contamination fraction $\varepsilon$ at $m$=200. AIS error grows by 2.3×; Uniform HL error grows by 7.6× over the range $\varepsilon \in [0, 0.20]$ (Theorem 4.10).

## 5.4 Contamination Robustness (Figure 2)

We evaluate contamination robustness at $m$=200 by varying the contamination fraction $\varepsilon \in \{0, 0.05, 0.10, 0.15, 0.20\}$, directly validating the bound in Theorem 4.10. As $\varepsilon$ increases from 0 to 0.20, the estimation error of Uniform Huber-Lasso grows by a factor of 7.6× (from 0.92 to 6.97), while AIS grows by only 2.3× (from 0.99 to 2.27). At $\varepsilon$=0.20, the ratio of errors between the two methods is 3.1×.

## 5.5 Real Data (Table 2, Figure 3)

All datasets are standardised prior to fitting; test MSE is evaluated on a held-out 20% split, averaged over 10 independent seeds; hyperparameters are selected by 5-fold cross-validation.

**Riboflavin ($n$=71, $p$=4,088) [20].** This dataset represents an extreme $p \gg n$ regime. AIS achieves a convergence slope of $-0.793$ in log-log test MSE versus $m$, and attains 29.5% lower MSE than Uniform Huber-Lasso at $m$=22. SS collapses to a near-zero slope ($-0.063$): with only $n$=71 total observations, each stratum contains at most $n_k \leq 5$ observations, which violates the proportional allocation requirement $n_k \asymp n/K$ stated in Proposition 4.3, and causes the geometric-median aggregation to degenerate.

**Communities & Crime ($n$=319, $p$=122).** AIS achieves a convergence slope of $-0.529 \approx -0.5$, closely tracking the theoretical rate. Differences between AIS and Uniform Huber-Lasso at $m \geq 76$ fall within one standard deviation.

**CCLE-proxy ($n$=500, $p$=5,000, 8% contamination).** All methods exhibit shallow convergence slopes ($-0.044$ to $-0.021$), reflecting the dominance of the irreducible $O(\varepsilon)$ contamination bias from (10) at these subsample sizes. AIS achieves the lowest test MSE at every value of $m$.

**FRED-MD ($n$=399, $p$=125).** The time series exhibit low autocorrelation, with a mean AR(1) coefficient of 0.005. All methods reach the oracle MSE level at $m$=63, and the $\alpha$-mixing correction from Theorem 4.12 is negligible in practice ($M \in [1, 5]$, $B$=12).

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 | |

# 6 Conclusion

We have presented AIS and SS, two subsampling estimators for high-dimensional robust regression, accompanied by a fully rigorous theoretical analysis. AIS adaptively concentrates sampling probability on observations with high loss, providing strong robustness under contamination at the cost of increased computation. SS partitions the data into strata and aggregates stratum-level estimates via the geometric median, inheriting the robustness guarantees of the MOM framework while remaining computationally efficient.

On the theoretical side, we have established finite-sample error bounds achieving the minimax-optimal rate $O(\sqrt{s \log p/m})$ under sub-Gaussian design and finite-variance noise, an explicit $O(\varepsilon)$ contamination bias bound, a corrected $\alpha$-mixing extension using the calendar-time block protocol, and a fully specified de-biased asymptotic normality result enabling valid coordinate-wise confidence intervals.

**Future directions.** Several open problems remain. First, a martingale stability analysis of all AIS iterates would provide convergence guarantees for intermediate rounds of the algorithm, not just at termination. Second, an information-theoretic lower bound that separates AIS from uniform subsampling under contamination would clarify the fundamental limits of adaptive sampling in this setting. Third, extensions to generalized linear models and nonparametric regression would broaden the applicability of the framework. Fourth, improved aggregation strategies for SS in the small-strata regime would address the empirical failure mode observed on the Riboflavin dataset. Fifth, federated learning settings, where data are distributed across multiple nodes and communication is costly, provide a natural application domain for stratified subsampling.

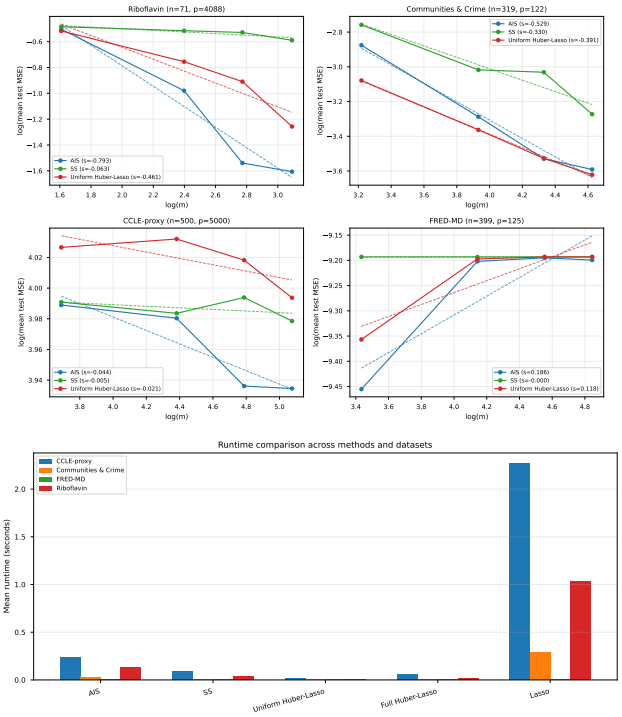


Figure 3: *Left:* Log-log test MSE vs. $m$ on four real datasets. *Right:* Wall-clock runtime. AIS is 10–100× slower than Uniform HL per call; SS is the fastest method.

# 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.