# Optimal Linear Regression Without a Variance

**Artin Tajdini**
University of Washington
artin@cs.washington.edu

**Jasper C.H. Lee**
University of California, Davis
jasperlee@ucdavis.edu

**Jonathan Scarlett**
National University of Singapore
scarlett@comp.nus.edu.sg

**Kevin Jamieson**
University of Washington
jamieson@cs.washington.edu

## Abstract

We study high-probability linear regression in the infinite variance regime, where the regression noise has only a finite conditional $\alpha$-moment $\nu$ for some unknown $\alpha \in (1, 2)$, and the covariates may only have finite moments up to a given order (e.g., 4th moment). We give a black-box reduction that converts finite-variance regression procedures into weak-moment procedures. The reduction is based on empirical residual pruning and is agnostic to the moment exponent, the noise scale, the covariance, and the design moment. Under a finite fourth-moment condition on the covariates, any finite-variance subroutine attaining the sub-Gaussian rate yields the following when the number of samples $n$ is above a suitably defined threshold: Without knowing $\alpha$ or $\nu$, we have with probability at least $1 - \delta$ that $\|\widehat{\theta} - \theta^*\|_\Sigma \lesssim_\alpha \nu^{1/\alpha} \big(\frac{d+\log(1/\delta)}{n}\big)^{\frac{\alpha-1}{\alpha}}$, where $d$ is the dimension, $\Sigma$ is the covariate covariance matrix, and $\|v\|_\Sigma := (v^\top \Sigma v)^{1/2}$. We further prove a matching minimax lower bound that holds under the same conditions. The reduction is computationally efficient whenever the finite-variance subroutine is. We also obtain a distribution-free prediction guarantee under a bounded raw $\alpha$-moment assumption on the response, and give an analogous $\alpha$-agnostic reduction for mean estimation.

## 1 Introduction

Statistical estimation under heavy-tailed distributions has received sustained attention since Catoni's work on robust mean and variance estimation (Catoni, 2012). A central theme in this line of work is that finite moment assumptions exhibit poor concentration of empirical averages, so high-probability guarantees require estimators designed for heavier tails. The mean-estimation problem is now well understood under finite (co)variance, both in one dimension and in high dimensions (Lugosi and Mendelson, 2019c; Cherapanamjeri et al., 2019; Hopkins, 2020; Lee and Valiant, 2022a,b), and more recent work gives optimal rates when the variance is infinite (Bubeck et al., 2013; Devroye et al., 2016; Cherapanamjeri et al., 2022; Whitehouse et al., 2024; Lee et al., 2025). In this paper, we study the corresponding question for linear regression.

Linear regression adds the geometry of the covariates to the heavy-tailed estimation problem, and robust procedures typically require some control of the covariate distribution. Sharp finite-variance guarantees are known under conditions such as finite fourth moments, hypercontractivity, or small-ball compatibility; see Table 1 and the related work below. Concretely, with $d$ parameters, $n$ data points, and noise variance $\sigma^2$, the prediction-norm rate is

$$\sigma \left( \frac{d + \log(1/\delta)}{n} \right)^{1/2}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

up to constants.[1] Importantly, ordinary least squares is insufficient when only finite variance noise is assumed, and prior works have developed a number of robust regression procedures in such settings. However, the "below-variance" regime, where the regression error has only an $\alpha$-moment for $1 < \alpha < 2$, is less developed.[2] We also refer to this setting as *weak-moment regression*.

Formally, our modeling assumptions are stated as follows.

**Assumption 1** (Linear model and error moment)**.** *There is a parameter $\theta^* \in \mathbb{R}^d$ such that*

$$y = x^\top \theta^* + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid x] = 0, \qquad \mathbb{E}[|\varepsilon|^\alpha \mid x] \le \nu,$$

*for some unknown $1 < \alpha \le 2$ and $\nu > 0$. The conditional distribution of $\varepsilon$ (given $x$) may depend on $x$, i.e., the noise need not be "purely additive".*

Given $n$ i.i.d. samples, a natural pessimistic baseline is to reduce the problem coordinatewise to one-dimensional weak-moment estimation (using scalar robust mean estimators such as (Bubeck et al., 2013; Devroye et al., 2016)), which gives the rate $\widetilde{\mathcal{O}}\Big(\nu^{1/\alpha}\sqrt{d}\big(\frac{\log(1/\delta)}{n}\big)^{(\alpha-1)/\alpha}\Big)$ and pays a multiplicative $\sqrt{d}$ penalty. Under sub-Gaussian covariates, Huber-loss methods already achieve sharp high-probability guarantees (Sun et al., 2020), improving the $d$ dependence to $d^{\frac{\alpha-1}{\alpha}}$. The question here is whether one can remove the coordinatewise loss without such light-tailed design assumptions. The central question of this paper is:

*Can one improve on the coordinatewise rate $\widetilde{\mathcal{O}}\Big(\nu^{1/\alpha}\sqrt{d}\big(\frac{\log(1/\delta)}{n}\big)^{(\alpha-1)/\alpha}\Big)$ without light-tailed assumptions on covariates?*

We answer this question affirmatively: under the standard finite fourth-moment covariate assumption, which is common even in finite-variance noise settings, we give matching upper and lower bounds on the estimation error, provided that the sample size is sufficiently large, without knowing $\alpha$ or $\nu$.

Our main algorithm is a black-box reduction from weak-moment regression to finite-variance regression. Starting from a sufficiently accurate initialization, our method prunes the most extreme sample points using an estimated threshold, and passes the resulting sample set to a black-box finite-variance regression subroutine. The role of pruning is to make the sample distribution effectively have a finite variance. The pruning threshold is chosen to balance between the variance reduction and the introduced bias, while our particular empirically-driven threshold-estimation procedure ensures that the threshold can be computed without knowledge of $\alpha$ and $\nu$. On top of this pruning procedure, we also employ a standard iterative approach for regression: at every stage with a current regression vector estimate $\widehat{\theta}$, we compute the residual $\tilde{y} = y - \widehat{\theta}^\top x$ and run regression on the residual to further reduce the error in $\widehat{\theta}$.

Our reduction is computationally efficient, guaranteeing the polynomial runtime of the overall regression algorithm whenever the black-box finite-variance regression subroutine has polynomial runtime. The main guarantees are summarized in the following proposition; here and subsequently, we write $\Sigma := \mathbb{E}[xx^\top] - \mathbb{E}[x]\mathbb{E}[x]^\top$ and $\|v\|_\Sigma := (v^\top \Sigma v)^{1/2}$.

**Proposition 1** (Informal)**.** *Under covariance $\Sigma$, and finite fourth-moment covariates, our black-box reduction turns any finite-variance regression routine with sub-Gaussian rate into an estimator $\widehat{\theta}$ satisfying, above an explicit sample-size threshold,*

$$\|\widehat{\theta} - \theta^*\|_\Sigma \lesssim_\alpha \nu^{1/\alpha}\left(\frac{d + \log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}} \tag{1}$$

*with probability at least $1 - \delta$. A matching lower bound under the same conditions shows that this rate is minimax optimal up to constants.*

Without covariate assumptions, parameter estimation is not identifiable in general. However, we can still obtain a parallel black-box reduction for prediction if the *non-central* weak-moment of the response is bounded by $\nu$.

[1] This result assumes positive-definite covariance.

[2] To our knowledge, the best known lower bound for $\Sigma = I$ is from Sun et al. (2020), and equals $\left(\frac{\log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}$ as the dimension-dependent terms cancel out; this is equal to mean estimation rate in $\mathbb{R}$.

Table 1: Representative high-probability robust regression guarantees.

| Work | Heavy-tailed covariates | Infinite variance ($\alpha \in (1,2)$) moment | Adaptive | Efficient | **Statistical optimality** |
|---|---|---|---|---|---|
| Sun et al. (2020) | ✗ | ✓ | ✗ | ✓ | ✓ |
| El Hanchi et al. (2024) | 4th moment | ✗ | - | ✗ | ✓ |
| Lugosi and Mendelson (2019b); Oliveira and Resende (2025) | SB or $L_2/L_1$ | ✗ | - | ✗ | ✓ |
| Hsu and Sabato (2014); Depersin (2020); Cherapanamjeri et al. (2020) | $L_4/L_2$ or stronger | ✗ | - | ✓ | ✓ |
| Pensia et al. (2025) | $L_4/L_2$ | independent noise[b] | - | ✓ | ✓ |
| Mourtada et al. (2022) | arbitrary | raw 2nd | - | ✗ | ?[a] |
| Sun et al. (2020) | bounded + Gram | ✓ | ✗ | ✓ | ✗ |
| Coordinatewise | covariance | ✓ | ✓ | ✓ | ✗ |
| This paper | 4th moment | ✓ | ✓ | ✗ | ✓[c] |
| | $L_4/L_2$ or stronger | ✓ | ✓ | ✓[c] | ✓[c] |
| | arbitrary | raw $\alpha$ | ✓ | ✗ | ?[a] |

"Adaptive" means adaptation to the weak-moment exponent and scale.
[a] Distribution-free prediction has an additional $\log(n/d)$ factor relative to the sub-Gaussian prediction rate.
[b] Assumes i.i.d. errors independent of the covariates and calibrates the guarantee through a noise quantile.
[c] Assuming an efficient, optimal black-box subroutine.

**Proposition 2** (Informal). *For arbitrary covariates, under the stronger assumption $\mathbb{E}[|y|^\alpha \mid x] \le \nu$, our reduction combined with a distribution-free finite-variance prediction subroutine returns a predictor $\widehat{h}$ satisfying, for sample size threshold $n \ge 2d$,*

$$\left(\mathbb{E}_x(\widehat{h}(x) - x^\top \theta^*)^2\right)^{1/2} \lesssim_\alpha \nu^{1/\alpha} \left(\frac{d\log(n/d) + \log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}, \tag{2}$$

*where the additional $\log(n/d)$ factor is inherited from the finite-variance distribution-free subroutine.*

**Contributions.** We summarize our main contributions as follows.

- We introduce a black-box reduction for solving regression with only a weak error moment via finite-variance regression. The reduction treats the finite-variance estimator as a subroutine and does not require knowledge of the moment exponent $\alpha$, scale $\nu$, or covariance $\Sigma$.
- Under a finite fourth-moment condition on the design, any subroutine for the finite-variance setting attaining a sub-Gaussian rate yields the rate in Eq. (1) when $n$ is above a suitably-defined threshold. The resulting estimator is computationally efficient whenever the finite-variance subroutine is computationally efficient.
- We prove matching lower bounds under the same assumptions. In our construction, the covariates have covariance $\Sigma$ and satisfy the fourth-to-second moment comparison used by the upper bound, showing that the rate Eq. (1) is minimax optimal up to constants (at least for $n$ exceeding the above-mentioned threshold).
- For distribution-free prediction with arbitrary covariates, when the weak moment condition is imposed on the full response rather than only on the regression error, our reduction gives the infinite-variance analogue of the best known finite-variance guarantee in Eq. (2); the additional $\log(n/d)$ factor is inherited from that finite-variance benchmark.

We also provide a simple experiment in Appendix H regarding known vs. unknown parameters $(\alpha, \nu)$.

### 1.1 Related Work

We outline the most relevant prior works here, and give a more detailed overview in Appendix G.

**Mean estimation.** Robust mean estimation provides a canonical setting for high-probability estimation under heavy tails; see Lugosi and Mendelson (2019a) for a survey. Median and median-of-means

ideas have a long history (Small, 1990; Minsker, 2015), and Catoni's work gives exponential-type deviation bounds under weak moment assumptions (Catoni, 2012). The finite-variance case is now well understood in one dimension and high dimensions (Bubeck et al., 2013; Devroye et al., 2016; Lee and Valiant, 2022a; Lugosi and Mendelson, 2019c; Cherapanamjeri et al., 2019; Hopkins, 2020; Lee and Valiant, 2022b). The infinite-variance case has more recently been developed for both optimal rates and adaptation to the unknown moment exponent or scale (Cherapanamjeri et al., 2022; Lee et al., 2025; Dang et al., 2023; Whitehouse et al., 2024). Related weak-moment ideas also appear in streaming estimation and bandits (Bubeck et al., 2013; Das et al., 2024; Genalti et al., 2024). The regression problem considered here adds a covariate distribution and requires controlling prediction or parameter error rather than only a mean.

**Heavy-tailed Regression.** Least squares can be highly suboptimal under heavy tails (Brownlees et al., 2015; Lugosi and Mendelson, 2019a). Median-of-means-based methods provide one route to high-probability regression under weak moments (Hsu and Sabato, 2014, 2016; Lugosi and Mendelson, 2019b; Lecué and Lerasle, 2020); however, the statistically optimal tournament-style guarantees are not computationally efficient, while computationally tractable variants typically require stronger structure or lose in the rate and/or confidence dependence.

Truncated least-squares and Catoni-style robust-risk objectives control heavy-tailed losses by replacing empirical means in the objective (Audibert and Catoni, 2011; Brownlees et al., 2015); trimmed-mean procedures give uniform guarantees for robust mean estimation and quadratic regression under weak moment and small-ball assumptions (Oliveira and Resende, 2025). The finite-variance quantile-risk benchmark under fourth moments is due to El Hanchi et al. (2024). On the computational side, Huber-type methods give sharp dense rates under sub-Gaussian designs and extend to heavy-tailed covariates through covariate filtering (Sun et al., 2020; Pensia et al., 2025). Spectral and semidefinite methods give computationally efficient high-probability guarantees under $L_4/L_2$, certifiable hypercontractivity, bounded covariance, or related assumptions (Depersin, 2020; Oliveira et al., 2022; Cherapanamjeri et al., 2020). Thus, the finite-variance literature separates statistical optimality under weak assumptions from computational efficiency under stronger covariate regularity.

## 2 Black-Box Reduction Algorithms

### 2.1 Estimation Under Finite Fourth-Moment Design

This section gives a black-box reduction from finite-variance regression to regression with only a conditional $\alpha$-moment of the noise.

**Assumption 2** (Fourth-to-second moment comparison)**.** *The covariance $\Sigma = \mathbb{E}[xx^\top]$ is positive definite, and*

$$K_4 := \sup_{\|v\|_\Sigma = 1} \mathbb{E}[(v^\top x)^4] < \infty.$$

We will see that a "pruning-bias" step in our analysis requires control of $\mathbb{E}|v^\top x|^{2\alpha}$ for $\|v\|_\Sigma = 1$, and Assumption 2 controls these quantities uniformly over $\alpha \in (1, 2]$, giving a common covariate assumption for adaptation to unknown $\alpha$. Note that $K_4$ may depend on $d$, unlike the constant hypercontractivity assumption used in prior work (Cherapanamjeri et al., 2020; Depersin, 2020; Pensia et al., 2025). In particular, for covariates sampled uniformly from the signed standard basis vectors, this constant grows linearly with dimension.

**Assumption 3** (Finite-variance estimation primitive)**.** *There is a regression routine* EstFV *with a rate function $R_{\mathrm{est}}$ satisfying the following: Whenever*

$$z = x^\top g + \xi, \qquad \mathbb{E}[x\xi] = 0, \qquad \mathbb{E}[\xi^2] \le \sigma^2,$$

*the output of* EstFV *on $N$ i.i.d. samples with failure probability $\eta$ satisfies*

$$\|\widehat{g} - g\|_\Sigma \le \sigma R_{\mathrm{est}}(N, \eta)$$

*with probability at least $1 - \eta$, for some rate function $R_{\mathrm{est}}$ that is nonincreasing in both arguments.*

Our procedure, described in Algorithm 1, does not require knowing $\alpha$, $\nu$, $\Sigma$, or $K_4$. The unknown moment exponent and scale enter only through the analysis; the empirical thresholds are estimated

**Algorithm 1** Black-Box Weak-Moment Regression

**Input:** Samples $\{(x_i, y_i)\}_{i=1}^n$, confidence $\delta \in (0, 1/10)$, finite-variance routine EstFV, and rate function $R_{\text{est}}$

**Output:** Estimator $\widehat{\theta}$

$S_0, \{(T_j, P_j)\}_{j=1}^{\log n+1} \leftarrow \mathsf{Split}(\{(x_i, y_i)\}_{i=1}^n)$ `// Satisfying Eq. (3)`

$\delta_j \leftarrow \delta/(8 \log n)$ $(j \le \log n)$, $\delta_{\log n+1} \leftarrow \delta/8$

$\widehat{\theta}_0 \leftarrow \mathsf{InitLAD}(S_0, \delta)$ (see Algorithm 2);

**for** $j = 1, \ldots, \log n + 1$ **do**

  $\widehat{\tau}_j \leftarrow \inf\left\{t : \frac{1}{|T_j|}\sum_{(x_i,y_i)\in T_j} \mathbf{1}\{|y_i - x_i^\top \widehat{\theta}_{j-1}| \le t\} \ge 1 - 64\left(R_{\text{est}}(|P_j|, \delta_j)^2 + \frac{\log(1/\delta_j)}{|T_j|}\right)\right\}$

  **foreach** $(x_i, y_i) \in P_j$ **do**

    $z_i \leftarrow (y_i - x_i^\top \widehat{\theta}_{j-1})\mathbf{1}\{|y_i - x_i^\top \widehat{\theta}_{j-1}| \le \widehat{\tau}_j\}$

  $\widehat{g}_j \leftarrow \mathsf{EstFV}\left(\{(x_i, z_i) : (x_i, y_i) \in P_j\}, \delta_j\right)$

  $\widehat{\theta}_j \leftarrow \widehat{\theta}_{j-1} + \widehat{g}_j$

**return** $\widehat{\theta}_{\log n+1}$

**Algorithm 2** Initial Estimate via Least Absolute Deviations (InitLAD)

**Input:** Initialization block $S_0$ and confidence $\delta \in (0, 1/10)$

**Output:** Initial estimate $\widehat{\theta}_0$

$M \leftarrow \lceil 512 \log \frac{1}{\delta} \rceil$

$\{B_k\}_{k=1}^M, \{V_g\}_{g=1}^M \leftarrow \mathsf{Split}(S_0)$ `// Equal size blocks`

**for** $k = 1, \ldots, M$ **do**

  $\widehat{\theta}_k \leftarrow \underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{(x_i,y_i)\in B_k} |y_i - x_i^\top \beta|$

  $\widehat{L}_k \leftarrow \operatorname{median}_{g\in[M]} \frac{1}{|V_g|} \sum_{(x_i,y_i)\in V_g} |y_i - x_i^\top \widehat{\theta}_k|$

**return** $\widehat{\theta}_{k^*}$ *where* $k^* = \operatorname{argmin}_{k\in[M]} \widehat{L}_k$

from the data. The split in Algorithm 1 is chosen so that

$$|S_0| \ge \frac{n}{4}, \qquad |T_j|, |P_j| \ge \frac{n}{8 \log n} \quad (j \le \log n), \qquad |T_{\log n+1}|, |P_{\log n+1}| \ge \frac{n}{4}. \tag{3}$$

Throughout the paper, we refer to $T_j$ and $P_j$ as *threshold blocks* and *pruning blocks* respectively. The main algorithmic steps are as follows:

- **Warm-start initialization.** The black-box reduction only needs an initial estimate at constant scale. We use Algorithm 2, which computes Least Absolute Deviations (LAD) estimates on independent subblocks and returns the candidate with the smallest median validation LAD loss.
- **Threshold estimation.** At each iteration, we use an independent threshold block to choose a threshold for the absolute residuals under the current estimate. The fraction of residuals to remove is chosen using the error bound for the finite-variance estimator. This balances the bias from pruning with the estimation error on the retained observations, without knowing $\alpha$ or $\nu$.
- **Residual pruning.** On the corresponding pruning block, residuals above the estimated threshold are set to zero. Conditional on earlier rounds and the current threshold block, the transformed observations in the independent pruning block remain i.i.d., as required by Assumption 3, whereas reusing the same observations would require a guarantee that also accounts for the resulting dependence. The pruned residuals define a finite-variance regression problem whose regression parameter is close to the remaining error $\theta^* - \widehat{\theta}$.
- **Finite-variance correction.** The routine EstFV estimates this correction term $\theta^* - \widehat{\theta}$, and the estimate is updated. The early iterations reduce the warm-start error to a constant scale; the final iteration gives the target $\nu^{1/\alpha} R_{\text{est}}^{2(\alpha-1)/\alpha}$ rate.

**Theorem 1** (Generic estimation theorem). *Under Assumption 1, Assumption 2, and Assumption 3, there exist constants $A_0, B_0 > 0$ such that if*

$$n \geq A_0 K_4^2 d \log(1/\delta), \tag{4}$$

*and*

$$R_{\text{est}}\left(\frac{n}{8\log n}, \frac{\delta}{8\log n}\right)^2 + \frac{\log n \, \log((\log n)/\delta)}{n} \leq \left(\frac{1}{A_0\sqrt{K_4}}\right)^{\frac{\alpha}{\alpha-1}}, \tag{5}$$

*then Algorithm 1 satisfies, with probability at least $1-\delta$,*

$$\|\widehat{\theta} - \theta^*\|_\Sigma \leq B_0 \nu^{1/\alpha} \left(R_{\text{est}}\left(\frac{n}{4}, \frac{\delta}{8}\right)^2 + \frac{\log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}. \tag{6}$$

**Corollary 1** (Fourth-moment plug-in). *Assume* EstFV *is the fourth-moment finite-variance estimator of El Hanchi et al. (2024). Let $N_{\text{EH}}(d, \delta, \mathcal{D}_X)$ denote the sample-size lower bound in their theorem at confidence $\delta$ for design distribution $\mathcal{D}_X$, and let $C_{\text{est}}$ be such that*

$$R_{\text{est}}(N, \eta) \leq C_{\text{est}} \sqrt{\frac{d + \log(1/\eta)}{N}}.$$

*There exist constants $A_{\alpha, C_{\text{est}}}, B_{\alpha, C_{\text{est}}} > 0$, depending only on $\alpha$ and $C_{\text{est}}$, such that if*[3]

$$n \geq A_{\alpha, C_{\text{est}}} \Big[ K_4^{3 \vee \left(1 + \frac{\alpha}{2(\alpha-1)}\right)} (d + \log(1/\delta)) \log\left(\frac{d + \log(1/\delta)}{\delta}\right) + N_{\text{EH}}(d, \delta, \mathcal{D}_X) \log\left(\frac{N_{\text{EH}}(d, \delta, \mathcal{D}_X)}{\delta}\right) \Big],$$

*then Algorithm 1 satisfies, with probability at least $1-\delta$,*

$$\|\widehat{\theta} - \theta^*\|_\Sigma \leq B_{\alpha, C_{\text{est}}} \nu^{1/\alpha} \left(\frac{d + \log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}. \tag{7}$$

**Remark 1** (Computational efficiency). *Algorithm 2 is computationally efficient since it solves only convex LAD optimizations. The pruning rounds evaluate empirical quantiles at the current estimate and call* EstFV. *Hence Algorithm 1 is computationally efficient whenever the finite-variance subroutine is computationally efficient. Using the finite-variance regression algorithm of Cherapanamjeri et al. (2020) as* EstFV *therefore gives a computationally efficient version under their stronger certifiable hypercontractivity assumptions, whereas using the regression algorithm of El Hanchi et al. (2024) as finite-variance subroutine gives a computationally inefficient algorithm.*

**Remark 2** (Empirical tuning). *Our algorithm adapts to the unknown moment exponent and scale through the threshold blocks. If $\alpha$ and $\nu$ are known, one may replace these empirical thresholds by the deterministic thresholds used in the proof. The same tuning idea can be used for weak-moment mean estimation: the deterministic pruning radius in Algorithm 5 of Cherapanamjeri et al. (2022) can be replaced by an empirical radius from an independent block. The four properties used in their Lemma C.7 continue to hold, up to an additional threshold-estimation event; consequently, the condition $\delta \geq e^{-\mathcal{O}(n)}$ only changes by numerical constants. See Appendix E for more details.*

### 2.2 Prediction Under Arbitrary Covariates

The distribution-free prediction version of our algorithm uses the same pruning step with the zero initial predictor. In this setting, we adopt a weak moment assumption on the response itself, rather than on the regression noise, as formalized in the following.

**Assumption 4** (Raw response moment). *There is a parameter $\theta^* \in \mathbb{R}^d$ such that*

$$\mathbb{E}[y \mid x] = x^\top \theta^*, \qquad \mathbb{E}\left[|y|^\alpha \mid x\right] \leq \nu$$

*for some $1 < \alpha \leq 2$ and $\nu > 0$.*

[3] The notation $a \vee b$ means $\max\{a, b\}$.

**Algorithm 3** Black-Box Distribution-Free Weak-Moment Regression

**Input:** $2n$ samples $\{(x_i, y_i)\}_{i=1}^{2n}$, confidence $\delta \in (0,1)$, class $\mathcal{F}$, finite-variance prediction routine PredFV, and rate function $R_{\mathrm{pred}}$

**Output:** Predictor $\widehat{h}$

$T, P \leftarrow \mathsf{Split}(\{(x_i, y_i)\}_{i=1}^{2n}), |T| = |P| = n$

**if** $R_{\mathrm{pred}}(n, \delta/2)^2 + \frac{\log(4/\delta)}{n} > 1/128$ **then**

  **return** $\widehat{h} \equiv 0$

$\widehat{\tau} \leftarrow \inf\left\{t : \frac{1}{|T|}\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| \le t\} \ge 1 - 64\left(R_{\mathrm{pred}}(n, \delta/2)^2 + \frac{\log(4/\delta)}{n}\right)\right\}$

**foreach** $(x_i, y_i) \in P$ **do**

  $z_i \leftarrow y_i \mathbf{1}\{|y_i| \le \widehat{\tau}\}$

$\widetilde{h} \leftarrow \mathsf{PredFV}\left(\{(x_i, z_i) : (x_i, y_i) \in P\}, \delta/2\right)$

**return** $x \mapsto \max\{-\widehat{\tau}, \min\{\widehat{\tau}, \widetilde{h}(x)\}\}$

This is stronger than Assumption 1 in the sense that a bound on the raw moment implies an analog on the centered moment (with at most a constant factor increase), but not vice versa. The reason for this stronger assumption is that in the distribution-free setting, a weak-moment assumption on the noise is not enough, since there exist noise-free instances such that the prediction error of every estimator is lower bounded by an absolute constant with constant probability (see Proposition 2 in Mourtada et al. (2022)). We will see that the raw response moment directly controls the bias from pruning $y$, so no warm start or iterative correction is needed. This is in contrast with Algorithm 1, where only the noise moment is bounded, and successive residual corrections progressively reduce the contribution of $x^\top(\theta^* - \widehat{\theta})$ to the pruning bias.

**Assumption 5** (Finite-variance prediction primitive)**.** *There is a prediction routine* PredFV *with rate function* $R_{\mathrm{pred}}$ *such that the following holds. For any distribution of* $(x, z)$ *with* $\mathbb{E}[z^2 \mid x] \le \sigma^2$ *a.s., the output* $\widetilde{h}$ *of* PredFV *on* $N$ *i.i.d. samples with failure probability* $\eta$ *satisfies*

$$\mathbb{E}\big(\widetilde{h}(x) - z\big)^2 - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}\left(x^\top\theta - z\right)^2 \le \sigma^2 R_{\mathrm{pred}}(N, \eta)^2$$

*with probability at least* $1 - \eta$*, for some rate function* $R_{\mathrm{pred}}$ *that is nonincreasing in both arguments.*

**Theorem 2** (Adaptive distribution-free prediction)**.** *Suppose that Assumption 4 and Assumption 5 hold, with no condition imposed on the distribution of* $x$*. Then Algorithm 3 satisfies with probability at least* $1 - \delta$ *that*

$$\left(\mathbb{E}_x(\widehat{h}(x) - x^\top\theta^*)^2\right)^{1/2} \le \sqrt{514}\,\nu^{1/\alpha}\left(R_{\mathrm{pred}}(n, \delta/2)^2 + \frac{\log(4/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}. \tag{8}$$

If the squared risks $\mathbb{E}_{x,y}(\widehat{h}(x) - y)^2$ and $\mathbb{E}_{x,y}(x^\top\theta^* - y)^2$ are finite, then Eq. (8) also implies the same bound on the squared excess risk, since $\mathbb{E}_{x,y}(\widehat{h}(x) - y)^2 - \mathbb{E}_{x,y}(x^\top\theta^* - y)^2 = \mathbb{E}_x\big(\widehat{h}(x) - x^\top\theta^*\big)^2$.

**Corollary 2** (Distribution-free weak-moment prediction)**.** *Suppose that Assumption 4 holds, and let* PredFV *be the distribution-free finite-variance predictor of Mourtada et al. (2022). For* $n \ge 2d$*, their theorem gives Assumption 5 with*

$$R_{\mathrm{pred}}(n, \eta)^2 \le C_{\mathrm{df}}\frac{d\log(n/d) + \log(1/\eta)}{n}$$

*for some numerical constant* $C_{\mathrm{df}}$*. Then Algorithm 3 satisfies, with probability at least* $1 - \delta$*,*

$$\left(\mathbb{E}_x(\widehat{h}(x) - x^\top\theta^*)^2\right)^{1/2} \le \sqrt{514}\,(C_{\mathrm{df}} + 1)^{\frac{\alpha-1}{\alpha}}\nu^{1/\alpha}\left(\frac{d\log(n/d) + \log(4/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}. \tag{9}$$

### 2.3 Proofs

We first state two key lemmas, whose proofs are deferred to Appendices A and B.

**Lemma 1** (Warm-start initialization). *Under Assumption 1 and Assumption 2, if*

$$|S_0| \geq 2^{31} K_4^2 d \log(1/\delta), \tag{10}$$

*then the output of Algorithm 2 satisfies, with probability at least* $1 - \delta/8$,

$$\|\widehat{\theta}_0 - \theta^*\|_\Sigma \leq 2^{24} K_4^2 \nu^{1/\alpha}. \tag{11}$$

**Lemma 2** (Residual-pruning step). *Suppose that Assumption 1 and Assumption 2 hold. Let* $\bar{\theta} \in \mathbb{R}^d$ *be deterministic, and let* $T$ *and* $P$ *be independent blocks of size* $N$. *Let* $\widehat{\tau}$ *be the empirical quantile of* $\{|y_i - x_i^\top \bar{\theta}| : (x_i, y_i) \in T\}$ *at level*

$$1 - 64\left(R_{\mathrm{est}}(N,\delta)^2 + \frac{\log(1/\delta)}{N}\right). \tag{12}$$

*For* $(x_i, y_i) \in P$, *define*

$$z_i := (y_i - x_i^\top \bar{\theta})\mathbf{1}\{|y_i - x_i^\top \bar{\theta}| \leq \widehat{\tau}\}, \qquad \widehat{g} := \mathsf{EstFV}\left(\{(x_i, z_i) : (x_i, y_i) \in P\}, \delta\right).$$

*If*

$$R_{\mathrm{est}}(N,\delta)^2 + \frac{\log(1/\delta)}{N} \leq \frac{1}{128}, \tag{13}$$

*and* EstFV *satisfies Assumption 3 at block size* $N$, *then with probability at least* $1 - 2\delta$,

$$\|\bar{\theta} + \widehat{g} - \theta^*\|_\Sigma \leq 32\left(1 + \sqrt{K_4}\frac{\|\bar{\theta} - \theta^*\|_\Sigma}{\nu^{1/\alpha}}\right)\nu^{1/\alpha}\left(R_{\mathrm{est}}(N,\delta)^2 + \frac{\log(1/\delta)}{N}\right)^{\frac{\alpha-1}{\alpha}}. \tag{14}$$

*Proof of Theorem 1.* We prove the theorem with $A_0 = 2^{40}$ and $B_0 = 2^8$. Define the normalized radii

$$r_j = \frac{\|\widehat{\theta}_j - \theta^*\|_\Sigma}{\nu^{1/\alpha}}.$$

By Lemma 1 and Eq. (4), with probability at least $1 - \delta/8$, the choice of $\delta_j$ and the sizes in Eq. (3) give

$$r_0 \leq 2^{24} K_4^2. \tag{15}$$

For each iterative round, Lemma 2 is applied with block size at least $\frac{n}{8\log n}$ and confidence $\delta_j = \delta/(8\log n)$. For every such round, we have

$$\begin{aligned} R_{\mathrm{est}}(|P_j|,\delta_j)^2 + \frac{\log(1/\delta_j)}{|T_j|} &\leq 32\left[R_{\mathrm{est}}\left(\frac{n}{8\log n}, \frac{\delta}{8\log n}\right)^2 + \frac{\log n\,\log((\log n)/\delta)}{n}\right] \\ &\leq 32\left(\frac{1}{A_0\sqrt{K_4}}\right)^{\frac{\alpha}{\alpha-1}} \qquad \text{(by Eq. (5))} \\ &\leq \frac{32}{A_0^2} \leq \frac{1}{128}. \qquad (K_4 \geq 1, \alpha \leq 2) \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned} 32\sqrt{K_4}\left(R_{\mathrm{est}}(|P_j|,\delta_j)^2 + \frac{\log(1/\delta_j)}{|T_j|}\right)^{\frac{\alpha-1}{\alpha}} &\leq \frac{32^{1+\frac{\alpha-1}{\alpha}}}{A_0} \\ &\leq \frac{32\sqrt{32}}{A_0} \leq \frac{1}{8}. \qquad (\alpha \leq 2) \end{aligned} \tag{17}$$

Then, by Lemma 2 and a union bound over the iterative rounds, we have with probability at least $1 - \delta/4$ that

$$r_{j+1} \leq 32\left(1 + \sqrt{K_4} r_j\right)\left(R_{\mathrm{est}}(|P_j|,\delta_j)^2 + \frac{\log(1/\delta_j)}{|T_j|}\right)^{\frac{\alpha-1}{\alpha}}$$

in each iterative round. Combining with Eq. (17), we obtain

$$r_{j+1} \leq \frac{1}{8\sqrt{K_4}} + \frac{r_j}{8}. \tag{18}$$

Let $r_{\text{iter}}$ denote the normalized radius after the $\log n$ iterative rounds. By repeated application of Eq. (18) over these rounds (along with $\sum_{i=1}^{\infty} 8^{-i} = 1/7$), we have

$$r_{\text{iter}} \le \frac{1}{7\sqrt{K_4}} + 8^{-\log_2 n} r_0 \le \frac{2}{7\sqrt{K_4}}. \qquad \text{(Eq. (15) and Eq. (4))} \tag{19}$$

Lemma 2 is applied once more to the final threshold/pruning pair with confidence $\delta/8$. These blocks have a larger size of at least $n/4$, so the monotonicity of $R_{\text{est}}$ and Eq. (16) imply that the final-round quantity in Eq. (13) is at most $1/128$. The final-round concentration term is at most $8\log(1/\delta)/n$. Using Eq. (19) and $B_0 = 2^8$, the final pruning step yields

$$\|\widehat{\theta} - \theta^*\|_\Sigma \le B_0 \nu^{1/\alpha} \left( R_{\text{est}}\left(\frac{n}{4}, \frac{\delta}{8}\right)^2 + \frac{\log(1/\delta)}{n} \right)^{\frac{\alpha-1}{\alpha}}$$

with probability at least $1 - \delta$. □

The proof of Theorem 2 uses a similar empirical-threshold argument from Theorem 1, but using the raw response rather than the residual. After pruning, $y\mathbf{1}\{|y| \le \widehat{\tau}\}$ has bounded conditional second moment, so the finite-variance prediction primitive applies to its conditional mean. The target shift caused by pruning is then controlled directly from the conditional raw $\alpha$-moment. The full details, as well as the proofs of Corollary 1 and Corollary 2, are deferred to Appendices C and D.

## 3 Lower Bound

Our main lower bound is stated as follows, and is proved in Appendix F.

**Theorem 3** (Random-design lower bound)**.** *Let $1 < \alpha < 2$, $\nu > 0$, and let $\Sigma \in \mathbb{R}^{d\times d}$ be positive definite. If $n \ge d$, then there exists a finite parameter set $\Theta$ and distributions $(\mathcal{D}_\theta)_{\theta\in\Theta}$ satisfying Assumption 1, and Assumption 2 with covariance $\Sigma$ and $K_4 \le d$, such that*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \Pr_{\mathcal{D}_\theta^{\otimes n}} \left( \|\widehat{\theta} - \theta\|_\Sigma \ge \frac{1}{8}\nu^{1/\alpha} \left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}} \right) \ge \frac{1}{6},$$

*where the infimum is over all measurable estimators based on the $n$ samples.*

*Moreover, the same family satisfies Assumption 4, and*

$$\inf_{\widehat{f}} \sup_{\theta\in\Theta} \Pr_{\mathcal{D}_\theta^{\otimes n}} \left( \left(\mathbb{E}_{x\sim\mathcal{D}_X} (\widehat{f}(x) - x^\top\theta)^2\right)^{1/2} \ge \frac{1}{8}\nu^{1/\alpha} \left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}} \right) \ge \frac{1}{6}$$

*where $\mathcal{D}_X$ is the common marginal distribution of the covariates, and the infimum is over measurable predictors based on the $n$ samples.*

*Proof Sketch.* The lower bound uses a rare-response construction. The parameter set becomes a hypercube after applying the map $\theta \mapsto \Sigma^{1/2}\theta$, with side length chosen so that $|x^\top\theta|$ is of order $\nu^{1/\alpha}(d/n)^{(\alpha-1)/\alpha}$ on every design point. The design chooses one coordinate uniformly and with a random sign, $x \in \{\pm\sqrt{d}\,\Sigma^{1/2}e_j : j \in [d]\}$. This gives covariance $\Sigma$ and satisfies the fourth-moment condition. Conditional on $x$, the response is usually zero and is equal to $(4n/d)x^\top\theta$ with probability $d/(4n)$. It follows that $\mathbb{E}[y \mid x] = x^\top\theta$, while the rare nonzero value is large enough to make the conditional $\alpha$-moment of order $\nu$.

The statistical hardness comes from hiding each coordinate sign behind a rare event. If two parameters differ only in the $j$-th coordinate after applying $\theta \mapsto \Sigma^{1/2}\theta$, their one-sample laws differ only when the design selects coordinate $j$ and the nonzero response occurs. Hence, the total variation distance between the $n$-sample laws is at most $1/4$. Assouad's method then implies that, under the uniform prior on the hypercube, any estimator must make a constant fraction of coordinate sign mistakes on average. Each wrong sign costs one side length in the corresponding coordinate after applying $\theta \mapsto \Sigma^{1/2}\theta$. Summing these errors gives the $\Sigma$-norm lower bound at the stated rate. The same sign mistakes give the direct $L_2(\mathcal{D}_X)$ prediction lower bound.

The complete proof is given in Appendix F. □

**Remark 3** (Confidence term)**.** *Theorem 3 proves a dimension-dependent lower bound but does not capture the dependence on the confidence parameter $\delta$. To obtain a confidence-dependent term, we can make use of the one-dimensional weak-moment mean-estimation lower bound. Specifically, take $Z$ uniform on $\{\pm 1\}^d$, set $x = \Sigma^{1/2} Z$, and restrict the parameter set to $\theta_t = t\Sigma^{-1/2} e_1$. Then $y = tZ_1 + \varepsilon$, and multiplying by the observed sign $Z_1$ gives the one-dimensional mean-estimation problem $Z_1 y = t + Z_1 \varepsilon$ with the same conditional $\alpha$-moment. Therefore, there exists a constant $c_\alpha > 0$, depending only on $\alpha$, such that the lower bound of Devroye et al. (2016) implies, for $0 < \delta < c_\alpha$,*

$$\inf_{\widehat{\theta}} \sup_{t} \Pr\left( \|\widehat{\theta} - \theta_t\|_\Sigma \geq c_\alpha \nu^{1/\alpha} \left( \frac{\log(1/\delta)}{n} \right)^{\frac{\alpha-1}{\alpha}} \right) \geq \delta.$$

*Combining this one-dimensional reduction with Theorem 3 gives the overall lower-bound rate with $d + \log(1/\delta)$ replacing the above numerator.*

**Remark 4.** *When $n < d$, the same lower bounds hold with $(d/n)^{(\alpha-1)/\alpha}$ replaced by 1. This follows by applying the theorem at sample size $d$ and using that minimax risk cannot increase when additional samples are given.*

For the constructed lower bound instance, $K_4 \leq d$ as stated in Theorem 3. Hence, for any sample size of at least $n = \widetilde{\Omega}(d^3)$, our upper bound is tight up to constants on this hard instance.

## 4 Conclusion

We introduced a black-box reduction for using finite-variance regression procedures in the infinite-variance setting without knowing the moment parameters $(\alpha, \nu)$. The reduction does not require knowledge of the covariance matrix, and it preserves the computational efficiency of the chosen subroutine. Under the fourth-moment covariate condition, an optimal finite-variance subroutine yields the minimax estimation rate. An analogous reduction gives adaptive prediction guarantees for arbitrary covariates under a moment bound on the raw response. For estimation, it remains open whether the fourth-moment condition on the covariates can be weakened while retaining the same estimation rate. For distribution-free prediction, the additional $\log(n/d)$ factor is inherited from the finite-variance subroutine, and it remains open whether it can be removed.

## Acknowledgement

This work is supported by the Singapore National Research Foundation (NRF) and the Ministry of Digital Development and Information (MDDI) under the AI Visiting Professorship programme (Award AIVP-2024-003).

## References

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794.

Bartlett, P. and Lugosi, G. (1999). An inequality for uniform deviations of sample averages from their means. *Statistics & probability letters*, 44(1):55–62.

Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et Statistiques*, volume 48, pages 1148–1185.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806.

Cherapanamjeri, Y., Hopkins, S. B., Kathuria, T., Raghavendra, P., and Tripuraneni, N. (2020). Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *ACM SIGACT Symposium on Theory of Computing*, pages 601–609.

Cherapanamjeri, Y., Tripuraneni, N., Bartlett, P., and Jordan, M. (2022). Optimal mean estimation without a variance. In *Conference on Learning Theory*, pages 356–357.

Dang, T., Lee, J., Song, M., and Valiant, P. (2023). Optimality in mean estimation: Beyond worst-case, beyond sub-Gaussian, and beyond $1 + \alpha$ moments. *Advances in Neural Information Processing Systems*, 36:4150–4176.

Das, A., Nagaraj, D., Pal, S., Suggala, A., and Varshney, P. (2024). Near-optimal streaming heavy-tailed statistical estimation with clipped SGD. *Advances in Neural Information Processing Systems*, 37:8834–8900.

Depersin, J. (2020). A spectral algorithm for robust regression with subgaussian rates. *arXiv preprint arXiv:2007.06072*.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695 – 2725.

El Hanchi, A., Maddison, C., and Erdogdu, M. (2024). Minimax linear regression under the quantile risk. In *Conference on Learning Theory*, pages 1516–1572.

Genalti, G., Marsigli, L., Gatti, N., and Metelli, A. M. (2024). $(\epsilon, u)$-adaptive regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 1882–1915.

Hopkins, S., Li, J., and Zhang, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912.

Hopkins, S. B. (2020). Mean estimation with sub-gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213.

Hsu, D. and Sabato, S. (2014). Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931.

Lee, J. C., McKelvie, W., Song, M., and Valiant, P. (2025). All-purpose mean estimation over r: Optimal sub-Gaussianity with outlier robustness and low moments performance. In *International Conference on Machine Learning*.

Lee, J. C. and Valiant, P. (2022a). Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE.

Lee, J. C. and Valiant, P. (2022b). Optimal sub-gaussian mean estimation in very high dimensions. In *Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 98–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Lugosi, G. and Mendelson, S. (2019b). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965.

Lugosi, G. and Mendelson, S. (2019c). Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, pages 2308–2335.

Mourtada, J., Vaškevičius, T., and Zhivotovskiy, N. (2022). Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 4(3):253–292.

Oliveira, R. I. and Resende, L. (2025). Trimmed sample means for robust uniform mean estimation and regression. *The Annals of Statistics*, 53(5):2153–2178.

Oliveira, R. I., Rico, Z. F., and Thompson, P. (2022). A spectral least-squares-type method for heavy-tailed corrupted regression with unknown covariance & heterogeneous noise. *arXiv preprint arXiv:2209.02856*.

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

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277.

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

Whitehouse, J., Chugg, B., Martinez-Taboada, D., and Ramdas, A. (2024). Mean estimation in Banach spaces under infinite variance and martingale dependence. *arXiv preprint arXiv:2411.11271*.

## A Proof of Lemma 1 (Warm-Start Initialization)

*Proof.* We refer to $B_k$ and $V_g$ in Algorithm 2 as LAD blocks and validation blocks, respectively.

The Paley–Zygmund inequality applied to $(x^\top v)^2$ implies, for every $v \in \mathbb{R}^d$,

$$\Pr\left(|x^\top v| \geq \frac{1}{2}\|v\|_\Sigma\right) \geq \frac{9}{16K_4}. \tag{20}$$

Let $M = \lceil 512 \log(1/\delta) \rceil$. The split in Algorithm 2 is chosen so that $|B_k|, |V_g| \geq |S_0|/(4M)$; by (10), $|B_k|, |V_g| \geq 2^{20} K_4^2 d$ for all $k, g \in [M]$.

For a LAD block $B_k$, let $\mathcal{E}_k$ be the event that

$$\frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} |y_i - x_i^\top \theta^*| \leq 16\nu^{1/\alpha} \quad \text{and} \quad \inf_{v\neq 0} \frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} \mathbf{1}\left\{|x_i^\top v| \geq \frac{1}{2}\|v\|_\Sigma\right\} \geq \frac{9}{32K_4}. \tag{21}$$

By Assumption 1 and Markov's inequality, we have

$$\Pr\left(\frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} |y_i - x_i^\top \theta^*| > 16\nu^{1/\alpha}\right) \leq \frac{\mathbb{E}|\varepsilon|}{16\nu^{1/\alpha}} \leq \frac{1}{16}. \tag{22}$$

For the empirical small-ball part of (21), let

$$\mathcal{C} = \left\{\left\{x : |x^\top v| \geq \frac{1}{2}\|v\|_\Sigma\right\} : v \neq 0\right\}.$$

Each $C \in \mathcal{C}$ is a union of two halfspaces. Halfspaces in $\mathbb{R}^d$ have VC dimension $d+1$, so Sauer's lemma bounds the number of subsets induced by $\mathcal{C}$ on $2|B_k|$ points by $(2e|B_k|/(d+1))^{2d+2}$. Letting $p = 9/(16K_4)$, (20) gives $\Pr(C) \geq p$ for every $C \in \mathcal{C}$. Using $|B_k| \geq 2^{20} K_4^2 d$, the VC inequality for relative deviation (Bartlett and Lugosi, 1999, Eq. (1)) gives

$$\begin{aligned}\Pr\left(\exists C \in \mathcal{C} : \frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} \mathbf{1}\{x_i \in C\} < \frac{1}{2}\Pr(C)\right) \\ \leq 8\left(\frac{6|B_k|}{d+1}\right)^{2d+2} \exp\left(-\frac{|B_k|p}{32}\right) \leq \frac{1}{64}.\end{aligned} \tag{23}$$

Hence, $\Pr(\mathcal{E}_k) \geq 59/64$.

We claim that on $\mathcal{E}_k$, every $\beta$ with $\|\beta - \theta^*\|_\Sigma \geq 2^{12} K_4 \nu^{1/\alpha}$ has larger empirical LAD loss than $\theta^*$. Let $r = \|\beta - \theta^*\|_\Sigma$. By (21), at least a fraction $9/(32K_4)$ of the block satisfies $|x_i^\top(\beta - \theta^*)| \geq r/2$, while

$$\frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} \mathbf{1}\{|y_i - x_i^\top \theta^*| > r/4\} \leq \frac{4}{r} \frac{1}{|B_k|} \sum_{(x_i,y_i)\in B_k} |y_i - x_i^\top \theta^*| \leq \frac{64\nu^{1/\alpha}}{r} \leq \frac{1}{64K_4}.$$

Hence, at least a fraction $1/(4K_4)$ of the block satisfies both $|x_i^\top(\beta - \theta^*)| \geq r/2$ and $|y_i - x_i^\top \theta^*| \leq r/4$. On these samples, $|y_i - x_i^\top \beta| \geq r/4$, so the empirical LAD loss of $\beta$ is at least $r/(16K_4)$, whereas that of $\theta^*$ is at most $16\nu^{1/\alpha}$. Since $r \geq 2^{12} K_4 \nu^{1/\alpha}$, the former loss is larger as claimed.

Since $\widehat{\theta}_k$ is a LAD minimizer, it follows that $\mathcal{E}_k$ implies

$$\|\widehat{\theta}_k - \theta^*\|_\Sigma \leq 2^{12} K_4 \nu^{1/\alpha}. \tag{24}$$

The events $\mathcal{E}_1, \ldots, \mathcal{E}_M$ are independent, and our choice of $M$ gives

$$\Pr\left(\bigcap_{k=1}^{M} \mathcal{E}_k^c\right) \leq \left(\frac{5}{64}\right)^M \leq \frac{\delta}{16}.$$

On $\bigcup_{k=1}^{M} \mathcal{E}_k$, let $k_0$ be the first index satisfying (24). For every validation block $V_g$ and every $\beta \in \mathbb{R}^d$, let

$$L_g(\beta) = \frac{1}{|V_g|} \sum_{(x_i,y_i)\in V_g} |y_i - x_i^\top \beta|, \qquad \widehat{L}(\beta) = \mathrm{median}_{g=1,\dots,M}\, L_g(\beta).$$

After the LAD-block observations are realized on $\bigcup_{k=1}^{M} \mathcal{E}_k$, the candidate $\widehat{\theta}_{k_0}$ is fixed and the validation blocks remain independent. For a fresh validation block,

$$\mathbb{E}L_g(\widehat{\theta}_{k_0}) \le \mathbb{E}|\varepsilon| + \mathbb{E}|x^\top(\widehat{\theta}_{k_0} - \theta^*)| \le 2^{13} K_4 \nu^{1/\alpha}.$$
(by Eq. (24), $\mathbb{E}|\varepsilon| \le \nu^{1/\alpha}$, and $\mathbb{E}|x^\top v| \le \|v\|_\Sigma$)

Then, over the validation blocks,

$$\Pr\left(\widehat{L}(\widehat{\theta}_{k_0}) > 2^{18} K_4 \nu^{1/\alpha}\right) \le \exp(-M/128) \le \frac{\delta}{32}. \quad \text{(Markov on each } V_g \text{ then Hoeffding)} \tag{25}$$

For each validation block $V_g$, define $\mathcal{F}_g$ by replacing $B_k$ by $V_g$ in (21). The analysis of $\mathcal{E}_k$ applies verbatim to $\mathcal{F}_g$, yielding $\Pr(\mathcal{F}_g) \ge 59/64$, and hence

$$\Pr\left(\sum_{g=1}^{M} \mathbf{1}\{\mathcal{F}_g\} < \frac{3}{4}M\right) \le \exp(-M/128) \le \frac{\delta}{32}. \qquad \text{(Hoeffding)} \tag{26}$$

The analysis of $\mathcal{E}_k$ further establishes the following analog on $\mathcal{F}_g$: Every $\beta$ with $\|\beta - \theta^*\|_\Sigma \ge 2^{24} K_4^2 \nu^{1/\alpha}$ satisfies $L_g(\beta) \ge 2^{20} K_4 \nu^{1/\alpha}$. Similarly, for $r = \|\beta - \theta^*\|_\Sigma$, at least a fraction $9/(32K_4)$ of the validation block has $|x_i^\top(\beta - \theta^*)| \ge r/2$, while

$$\frac{1}{|V_g|} \sum_{(x_i,y_i)\in V_g} \mathbf{1}\{|y_i - x_i^\top \theta^*| > r/4\} \le \frac{64\nu^{1/\alpha}}{r} \le \frac{1}{64K_4}.$$

Hence, at least a fraction $1/(4K_4)$ of the validation block satisfies $|y_i - x_i^\top \beta| \ge r/4$, and so $L_g(\beta) \ge r/(16K_4) \ge 2^{20} K_4 \nu^{1/\alpha}$.

On $\bigcup_{k=1}^{M} \mathcal{E}_k$ and outside the failure events in (25) and (26), every $\beta$ with $\|\beta - \theta^*\|_\Sigma \ge 2^{24} K_4^2 \nu^{1/\alpha}$ has $\widehat{L}(\beta) > 2^{18} K_4 \nu^{1/\alpha}$, while $\widehat{L}(\widehat{\theta}_{k_0}) \le 2^{18} K_4 \nu^{1/\alpha}$. The minimizer in Algorithm 2 therefore satisfies (11). The failure probabilities in (25), (26), and $\bigcap_{k=1}^{M} \mathcal{E}_k^c$ sum to at most $\delta/8$. □

## B Proof of Lemma 2 (Residual-Pruning Step)

*Proof.* Recall that we refer to $T_j$ and $P_j$ in Algorithm 1 as *threshold blocks* and *pruning blocks* respectively. Let

$$\gamma = R_{\mathrm{est}}(N,\delta)^2 + \frac{\log(1/\delta)}{N}, \qquad \Delta = \bar{\theta} - \theta^*. \tag{27}$$

By (13), $\gamma \le 1/128$. For a fresh sample $(x, y)$, the definition of $\Delta$ gives $y - x^\top \bar{\theta} = \varepsilon - x^\top \Delta$. Using $1 < \alpha \le 2$ and Jensen's inequality, we have

$$\begin{aligned} \mathbb{E}|\varepsilon - x^\top \Delta|^\alpha &\le 2\left(\mathbb{E}|\varepsilon|^\alpha + \mathbb{E}|x^\top \Delta|^\alpha\right) \\ &\le 2\left(\nu + \|\Delta\|_\Sigma^\alpha\right). \end{aligned} \tag{28}$$

For every $v$ with $\|v\|_\Sigma = 1$, we have

$$\begin{aligned} &\mathbb{E}\left[|v^\top x|^\alpha |\varepsilon - x^\top \Delta|^\alpha\right] \\ &\le 2\mathbb{E}\left[|v^\top x|^\alpha |\varepsilon|^\alpha\right] + 2\mathbb{E}\left[|v^\top x|^\alpha |x^\top \Delta|^\alpha\right] \\ &\le 2\nu \mathbb{E}|v^\top x|^\alpha + 2\mathbb{E}\left[|v^\top x|^\alpha |x^\top \Delta|^\alpha\right] && \text{(tower property, noise moment)} && (29) \\ &\le 2\nu + 2\left(\mathbb{E}|v^\top x|^{2\alpha}\right)^{1/2}\left(\mathbb{E}|x^\top \Delta|^{2\alpha}\right)^{1/2} && (\|v\|_\Sigma = 1,\ \alpha \le 2, \text{ Cauchy–Schwarz}) && (30) \\ &\le 2\nu + 2K_4^{\alpha/2}\|\Delta\|_\Sigma^\alpha && (2\alpha \le 4, \text{ definition of } K_4) && (31) \end{aligned}$$

Let $R = |y - x^\top \bar{\theta}|$. By the definition of $\widehat{\tau}$ in the lemma statement, if $\widehat{\tau} < \inf\{t : \Pr(R \le t) \ge 1 - 128\gamma\}$, then at least $(1 - 64\gamma)N$ threshold-block residuals are smaller than this population quantile, whose strict lower tail is at most $1 - 128\gamma$. Hence

$$\Pr\left(\widehat{\tau} < \inf\{t : \Pr(R \le t) \ge 1 - 128\gamma\}\right) \le \exp(-16N\gamma) \le \delta/2. \qquad \text{(Chernoff)}$$

Moreover, Markov's inequality gives $\Pr(R > (\mathbb{E}R^\alpha/(16\gamma))^{1/\alpha}) \le 16\gamma$. If $\widehat{\tau} > (\mathbb{E}R^\alpha/(16\gamma))^{1/\alpha}$, then more than $64\gamma N$ threshold-block residuals exceed this deterministic level; it follows that

$$\Pr\left(\widehat{\tau} > \left(\frac{\mathbb{E}R^\alpha}{16\gamma}\right)^{1/\alpha}\right) \le \exp(-16N\gamma) \le \delta/2. \qquad \text{(Chernoff)}$$

Hence, with probability at least $1 - \delta$ over the threshold block,

$$\Pr(|y - x^\top \bar{\theta}| > \widehat{\tau}) \le 128\gamma, \qquad \widehat{\tau}^\alpha \le \frac{\mathbb{E}|y - x^\top \bar{\theta}|^\alpha}{16\gamma}. \tag{32}$$

We henceforth condition on the threshold block and assume (32). Define

$$g = \Sigma^{-1}\mathbb{E}\left[x(y - x^\top \bar{\theta})\mathbf{1}\{|y - x^\top \bar{\theta}| \le \widehat{\tau}\}\right].$$

The identity $\mathbb{E}[x(y - x^\top \bar{\theta})] = \Sigma(\theta^* - \bar{\theta})$ gives

$$g - (\theta^* - \bar{\theta}) = -\Sigma^{-1}\mathbb{E}\left[x(y - x^\top \bar{\theta})\mathbf{1}\{|y - x^\top \bar{\theta}| > \widehat{\tau}\}\right].$$

Therefore,

$$\begin{aligned}
&\|g - (\theta^* - \bar{\theta})\|_\Sigma \\
&= \sup_{\|v\|_\Sigma = 1} \left|\mathbb{E}\left[v^\top x(y - x^\top \bar{\theta})\mathbf{1}\{|y - x^\top \bar{\theta}| > \widehat{\tau}\}\right]\right| \\
&\le \sup_{\|v\|_\Sigma = 1} \left(\mathbb{E}\left[|v^\top x|^\alpha |y - x^\top \bar{\theta}|^\alpha\right]\right)^{1/\alpha} \Pr(|y - x^\top \bar{\theta}| > \widehat{\tau})^{\frac{\alpha-1}{\alpha}} && \text{(Hölder)} \\
&\le \left(2\nu + 2K_4^{\alpha/2}\|\Delta\|_\Sigma^\alpha\right)^{1/\alpha} (128\gamma)^{\frac{\alpha-1}{\alpha}} && \text{(Eq. (31), Eq. (32))} \\
&\le 24\left(\nu^{1/\alpha} + \sqrt{K_4}\|\Delta\|_\Sigma\right)\gamma^{\frac{\alpha-1}{\alpha}}. && (1 < \alpha \le 2)
\end{aligned} \tag{33}$$

On the pruning block, the pruned response decomposes as $z = x^\top g + \xi$. By construction, $\mathbb{E}[x\xi] = 0$. Since $g$ is the population least-squares coefficient for the pruned response,

$$\mathbb{E}[\xi^2] \le \mathbb{E}[z^2].$$

By the definition of $z$ in Algorithm 1, we have $z^2 \le \widehat{\tau}^{2-\alpha}|y - x^\top \bar{\theta}|^\alpha$, and combining this with Eq. (32) and Eq. (28), we obtain

$$\mathbb{E}[z^2] \le \widehat{\tau}^{2-\alpha}\mathbb{E}|y - x^\top \bar{\theta}|^\alpha \le 4\left(\nu^{1/\alpha} + \|\Delta\|_\Sigma\right)^2 \gamma^{-\frac{2-\alpha}{\alpha}}. \tag{34}$$

The finite-variance guarantee then implies, with probability at least $1 - \delta$,

$$\begin{aligned}
\|\widehat{g} - g\|_\Sigma &\le 2\left(\nu^{1/\alpha} + \|\Delta\|_\Sigma\right) R_{\text{est}}(N, \delta)\gamma^{-\frac{2-\alpha}{2\alpha}} \\
&\le 2\left(\nu^{1/\alpha} + \|\Delta\|_\Sigma\right)\gamma^{\frac{\alpha-1}{\alpha}}, && (R_{\text{est}}^2 \le \gamma)
\end{aligned} \tag{35}$$

Moreover, since $K_4 \ge 1$ by definition,

$$\nu^{1/\alpha} + \|\Delta\|_\Sigma \le \left(1 + \sqrt{K_4}\frac{\|\Delta\|_\Sigma}{\nu^{1/\alpha}}\right)\nu^{1/\alpha}.$$

Combining (33) and (35) proves (14). The bound in (32) fails with probability at most $\delta$, and the finite-variance call fails with probability at most $\delta$, for a total of $2\delta$. □

## C Proof of Theorem 2 (Distribution-Free Prediction)

To simplify notation, we define

$$\gamma = R_{\text{pred}}(n, \delta/2)^2 + \frac{\log(4/\delta)}{n}. \tag{36}$$

If $\gamma > 1/128$, the algorithm returns $\widehat{h}(x) = 0$. By Assumption 4,

$$|x^\top \theta^*|^\alpha = |\mathbb{E}[y \mid x]|^\alpha \le \mathbb{E}[|y|^\alpha \mid x] \le \nu,$$
$$\mathbb{E}_x\left(\widehat{h}(x) - x^\top\theta^*\right)^2 = \mathbb{E}_x\left(x^\top\theta^*\right)^2 \le \nu^{2/\alpha} \le 128\,\nu^{2/\alpha}\gamma^{\frac{2(\alpha-1)}{\alpha}}.$$

This proves Eq. (8) in the case $\gamma > 1/128$. Assume now that $\gamma \le 1/128$. Since $|T| = n$, we have $n\gamma \ge \log(4/\delta)$. Define the deterministic thresholds

$$\tau_{\text{quant}} := \inf\{t : \Pr(|y| \le t) \ge 1 - 128\gamma\}, \qquad \tau_{\text{moment}} := \left(\frac{\nu}{16\gamma}\right)^{1/\alpha},$$

noting that $\Pr(|y| < \tau_{\text{quant}}) \le 1 - 128\gamma$ by definition. By the definition of $\widehat{\tau}$ in Algorithm 3, if $\widehat{\tau} < \tau_{\text{quant}}$, then at most $64\gamma n$ observations in $T$ satisfy $|y_i| \ge \tau_{\text{quant}}$. Combining this with $\Pr(|y| \ge \tau_{\text{quant}}) \ge 128\gamma$, we have

$$\Pr\left(\widehat{\tau} < \tau_{\text{quant}}\right) \le \Pr\left(\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| \ge \tau_{\text{quant}}\} \le \frac{1}{2}\mathbb{E}\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| \ge \tau_{\text{quant}}\}\right). \tag{37}$$

We further upper bound the right-hand side as follows:

$$\begin{aligned}
\Pr\left(\widehat{\tau} < \tau_{\text{quant}}\right) &\le \exp\left(-\frac{1}{8}\mathbb{E}\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| \ge \tau_{\text{quant}}\}\right) && \text{(Eq. (37) and Chernoff)} \\
&= \exp\left(-\frac{n}{8}\Pr(|y| \ge \tau_{\text{quant}})\right) \\
&\le \exp(-16n\gamma) && (\Pr(|y| \ge \tau_{\text{quant}}) \ge 128\gamma) \\
&\le \delta/4. && \text{(choice of } \gamma) \qquad (38)
\end{aligned}$$

The upper deterministic threshold satisfies

$$\begin{aligned}
\Pr(|y| > \tau_{\text{moment}}) &\le \frac{\mathbb{E}|y|^\alpha}{\tau_{\text{moment}}^\alpha} && \text{(Markov)} \\
&\le 16\gamma. && \text{(moment bound, choice of } \tau_{\text{moment}})
\end{aligned}$$

Next, using the definition of $\widehat{\tau}$, we observe that if $\widehat{\tau} > \tau_{\text{moment}}$, then more than $64\gamma n$ observations in $T$ exceed $\tau_{\text{moment}}$. It follows that

$$\begin{aligned}
\Pr\left(\widehat{\tau} > \tau_{\text{moment}}\right) &\le \Pr\left(\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| > \tau_{\text{moment}}\} > 64\gamma n\right) \\
&\le \left(\frac{e\,\mathbb{E}\sum_{(x_i,y_i)\in T} \mathbf{1}\{|y_i| > \tau_{\text{moment}}\}}{64\gamma n}\right)^{64\gamma n} && \text{(Chernoff)} \\
&= \left(\frac{e\,\Pr(|y| > \tau_{\text{moment}})}{64\gamma}\right)^{64\gamma n} \\
&\le \left(\frac{e}{4}\right)^{64\gamma n} && \text{(Markov)} \\
&\le \exp(-16n\gamma) \\
&\le \delta/4. && \text{(choice of } \gamma) \qquad (39)
\end{aligned}$$

Let $\mathcal{E}_{\text{thresh}} := \{\tau_{\text{quant}} \le \widehat{\tau} \le \tau_{\text{moment}}\}$. The bounds Eq. (38) and Eq. (39) imply

$$\Pr(\mathcal{E}_{\text{thresh}}^c) \le \delta/2.$$

By the definition of $\tau_{\text{quant}}$, $\Pr(|y| > \tau_{\text{quant}}) \le 128\gamma$. On $\mathcal{E}_{\text{thresh}}$,

$$\Pr(|y| > \widehat{\tau}) \le 128\gamma, \qquad \widehat{\tau}^\alpha \le \frac{\nu}{16\gamma}. \tag{40}$$

Note also that on $\mathcal{E}_{\text{thresh}}$, the threshold is fixed before the independent pruning block is used. Let $z$ denote the corresponding population pruned response, and let $g(x) = \mathbb{E}[z \mid x]$. We then have

$$\begin{aligned}
|g(x) - x^\top\theta^*| &= |\mathbb{E}\left[y\,\mathbf{1}\{|y| > \widehat{\tau}\} \mid x\right]| \\
&\le \left(\mathbb{E}[|y|^\alpha \mid x]\right)^{1/\alpha} \Pr(|y| > \widehat{\tau} \mid x)^{\frac{\alpha-1}{\alpha}} && \text{(Hölder)} \\
&\le \nu^{1/\alpha} \Pr(|y| > \widehat{\tau} \mid x)^{\frac{\alpha-1}{\alpha}}. && \text{(response moment)}
\end{aligned}$$

Taking the squared average, we obtain

$$\begin{aligned}
\mathbb{E}_x\left(g(x) - x^\top\theta^*\right)^2 &\le \nu^{2/\alpha}\mathbb{E}_x\left[\Pr(|y| > \widehat{\tau} \mid x)^{\frac{2(\alpha-1)}{\alpha}}\right] \\
&\le \nu^{2/\alpha} \Pr(|y| > \widehat{\tau})^{\frac{2(\alpha-1)}{\alpha}} && \text{(Jensen and } 2(\alpha-1)/\alpha \le 1) \\
&\le 128\,\nu^{2/\alpha}\gamma^{\frac{2(\alpha-1)}{\alpha}}. && \text{(by Eq. (40))}
\end{aligned} \tag{41}$$

The pruned response satisfies

$$\mathbb{E}[z^2 \mid x] \le \widehat{\tau}^{2-\alpha}\mathbb{E}[|y|^\alpha \mid x] \le \nu\widehat{\tau}^{2-\alpha}.$$

Since the pruning block is independent of the threshold block, Assumption 5 gives, with probability at least $1 - \delta/2$ over the pruning block,

$$\begin{aligned}
\mathbb{E}_x\left(\widetilde{h}(x) - g(x)\right)^2 &\le \mathbb{E}_x\left(x^\top\theta^* - g(x)\right)^2 + \nu\widehat{\tau}^{2-\alpha}R_{\text{pred}}(n, \delta/2)^2 \\
&\le \mathbb{E}_x\left(x^\top\theta^* - g(x)\right)^2 + \nu\widehat{\tau}^{2-\alpha}\gamma && \text{(choice of } \gamma) \\
&\le 129\,\nu^{2/\alpha}\gamma^{\frac{2(\alpha-1)}{\alpha}} && \text{(Eq. (41) and Eq. (40))}
\end{aligned} \tag{42}$$

Since $|z| \le \widehat{\tau}$, we also have $g(x) \in [-\widehat{\tau}, \widehat{\tau}]$. Clipping $\widetilde{h}$ to this interval cannot increase its squared error to $g$, so Eq. (42) holds with $\widehat{h}$ in place of $\widetilde{h}$. Combining Eq. (41) and Eq. (42),

$$\begin{aligned}
\mathbb{E}_x\left(\widehat{h}(x) - x^\top\theta^*\right)^2 &\le 2\mathbb{E}_x\left(\widehat{h}(x) - g(x)\right)^2 + 2\mathbb{E}_x\left(g(x) - x^\top\theta^*\right)^2 \\
&\le 514\,\nu^{2/\alpha}\gamma^{\frac{2(\alpha-1)}{\alpha}}.
\end{aligned}$$

The threshold event fails with probability at most $\delta/2$, and the finite-variance prediction call in Eq. (42) fails with probability at most $\delta/2$ conditionally on the threshold block. By the union bound, we conclude that Eq. (8) holds with probability at least $1 - \delta$.

# D Proofs of Corollaries

## D.1 Proof of Corollary 1 (Fourth-Moment Plug-in)

*Proof of Corollary 1.* The assumed finite-variance rate gives

$$R_{\text{est}}\left(\frac{n}{8\log n}, \frac{\delta}{8\log n}\right)^2 \le 32C_{\text{est}}^2\frac{\log n\,(d + \log(\log n/\delta))}{n}.$$

The additional term in (5) is bounded by the same expression with $C_{\text{est}}^2$ replaced by 1. Choose $A_{\alpha,C_{\text{est}}} = 2^{24}(C_{\text{est}}^2 + 1)\left[A_0 + A_0^{\alpha/(\alpha-1)}\right]$. The assumed lower bound on $n$ implies (4) and

$$R_{\text{est}}\left(\frac{n}{8\log n}, \frac{\delta}{8\log n}\right)^2 + \frac{\log n\,\log(\log n/\delta)}{n} \le \frac{2^{19}(C_{\text{est}}^2 + 1)}{A_{\alpha,C_{\text{est}}}K_4^{2\vee\frac{\alpha}{2(\alpha-1)}}}.$$

By the choice of $A_{\alpha,C_{\mathrm{est}}}$ and $K_4 \geq 1$, this is at most

$$\left(\frac{1}{A_0\sqrt{K_4}}\right)^{\frac{\alpha}{\alpha-1}},$$

which gives Eq. (5).

In (El Hanchi et al., 2024, Theorem 11), replacing $\delta$ by $\eta$ only changes the sample-size lower bound through $\log(1/\eta)$. Using the second term in the assumed lower bound on $n$, together with the preceding control of the logarithmic factors, we have

$$\frac{n}{8\log n} \geq N_{\mathrm{EH}}\left(d, \frac{\delta}{8\log n}, \mathcal{D}_X\right), \qquad \frac{n}{4} \geq N_{\mathrm{EH}}\left(d, \frac{\delta}{8}, \mathcal{D}_X\right).$$

Hence, every block used by Algorithm 1 is large enough to apply (El Hanchi et al., 2024, Theorem 11) at the corresponding confidence level. On the final block,

$$R_{\mathrm{est}}\left(\frac{n}{4}, \frac{\delta}{8}\right)^2 + \frac{\log(1/\delta)}{n} \leq 8(C_{\mathrm{est}}^2+1)\frac{d+\log(1/\delta)}{n}. \tag{43}$$

Substituting Eq. (43) into Eq. (6) gives Eq. (7) with $B_{\alpha,C_{\mathrm{est}}} = B_0\big(8(C_{\mathrm{est}}^2+1)\big)^{(\alpha-1)/\alpha}$. □

### D.2 Proof of Corollary 2 (Distribution-Free Weak-Moment Prediction)

*Proof of Corollary 2.* The assumed prediction rate gives

$$\begin{aligned} R_{\mathrm{pred}}(n,\delta/2)^2 + \frac{\log(4/\delta)}{n} &\leq C_{\mathrm{df}}\frac{d\log(n/d)+\log(2/\delta)}{n} + \frac{\log(4/\delta)}{n} \\ &\leq (C_{\mathrm{df}}+1)\frac{d\log(n/d)+\log(4/\delta)}{n}. \end{aligned}$$

Substituting this into Eq. (8) gives Eq. (9). □

## E Mean Estimation with Unknown $\alpha$

We first show that an empirical pruning threshold, estimated from $n$ additional samples, yields the high-probability guarantees needed for the subsequent mean estimation procedure. This serves as an empirical counterpart to a deterministic pruning threshold used in Cherapanamjeri et al. (2022).

**Lemma 3** (Empirical pruning counterpart of Lemma C.7 in Cherapanamjeri et al. (2022))**.** *Let $1 < \alpha < 2$, and let $X$ be a mean-zero random vector in $\mathbb{R}^d$ such that*

$$\sup_{\|v\|_2=1} \mathbb{E}|\langle X, v\rangle|^\alpha \leq 1. \tag{44}$$

*Let $\delta \in (0,1)$, and set*

$$q = \min\left\{\frac{1}{64}, \frac{6(d+\log(16/\delta))}{n}\right\}.$$

*Let $x^\dagger$ be independent of the threshold and pruning samples (denoted $T$ and $X_1,\dots,X_n$ below) and satisfy $\|x^\dagger\|_2 \leq 24\sqrt{d}$. From a threshold block $T$ of size $n$, define*

$$\widehat{\tau} = \max\left\{\sqrt{d}, \inf\left\{t \geq 0 : \frac{1}{n}\sum_{X_i \in T} \mathbf{1}\{\|X_i - x^\dagger\|_2 \leq t\} \geq 1-q\right\}\right\}.$$

*Let $X_1,\dots,X_n$ be an independent pruning block, set $\mathcal{Y} = \{X_i : \|X_i - x^\dagger\|_2 \leq \widehat{\tau}\}$, and let $Y$ have the conditional law of $X$ given $\|X - x^\dagger\|_2 \leq \widehat{\tau}$. Denote the mean and covariance of $Y$ by $\widetilde{\mu}$ and $\widetilde{\Sigma}$. With probability at least $1 - \delta/8 - 3e^{-n/1000}$, the following hold:*

(i) $|\mathcal{Y}| \geq 3n/4$, (ii) $\|\widetilde{\mu}\|_2 \leq 2q^{\frac{\alpha-1}{\alpha}}$,

(iii) $\sup_{\|v\|_2=1} \mathbb{E}|\langle Y - \widetilde{\mu}, v\rangle|^\alpha \leq 20$, (iv) $\mathrm{Tr}(\widetilde{\Sigma}) \leq 20d\,q^{-\frac{2-\alpha}{\alpha}}$.

*Proof.* For $u \in (0,1)$, let $\tau(u) = \inf\{t : \Pr(\|X - x^\dagger\|_2 \le t) \ge 1-u\}$. Since $\Pr(\|X - x^\dagger\|_2 < \tau(2q)) \le 1 - 2q$,

$$\begin{aligned}\Pr\left(\widehat{\tau} < \max\{\sqrt{d}, \tau(2q)\}\right) &\le \Pr\left(\frac{1}{n}\sum_{X_i \in T} \mathbf{1}\{\|X_i - x^\dagger\|_2 \ge \tau(2q)\} \le q\right) \\ &\le e^{-nq/4}. \qquad \text{(Chernoff)}\end{aligned}$$

Moreover, $\Pr(\|X - x^\dagger\|_2 > \tau(q/2)) \le q/2$, so

$$\begin{aligned}\Pr\left(\widehat{\tau} > \max\{\sqrt{d}, \tau(q/2)\}\right) &\le \Pr\left(\frac{1}{n}\sum_{X_i \in T} \mathbf{1}\{\|X_i - x^\dagger\|_2 > \tau(q/2)\} > q\right) \\ &\le e^{-nq/6}. \qquad \text{(Chernoff)}\end{aligned}$$

Let $\mathcal{E}$ denote the event that both displayed threshold bounds hold. Under $\mathcal{E}$, the definition of $\tau(2q)$ gives

$$\Pr(\|X - x^\dagger\|_2 > \widehat{\tau}) \le 2q. \tag{45}$$

Moreover, Markov's inequality gives

$$\Pr\left(\|X\|_2 > \left(\frac{2}{q}\mathbb{E}\|X\|_2^\alpha\right)^{1/\alpha}\right) \le \frac{q}{2}. \tag{46}$$

We also make use of the following moment upper bound from (Cherapanamjeri et al., 2022, Lemma 4.1):

$$\mathbb{E}\|X\|_2^\alpha \le \frac{4}{3}d^{\alpha/2}. \tag{47}$$

Combining the above, we have

$$\begin{aligned}\tau(q/2) &\le \|x^\dagger\|_2 + \left(\frac{2}{q}\mathbb{E}\|X\|_2^\alpha\right)^{1/\alpha} && \text{(definition of } \tau\text{, triangle inequality, Eq. (46))} \\ &\le 24\sqrt{d} + \left(\frac{8}{3q}d^{\alpha/2}\right)^{1/\alpha} && (\|x^\dagger\|_2 \le 24\sqrt{d}\text{, Eq. (47))} \\ &\le 6\sqrt{d}\,q^{-1/\alpha}. && (q \le 1/64,\ 1 < \alpha < 2)\end{aligned}$$

Combining with $\widehat{\tau} \le \max\{\sqrt{d}, \tau(q/2)\}$ therefore gives

$$\widehat{\tau} \le 6\sqrt{d}\,q^{-1/\alpha}. \tag{48}$$

On $\mathcal{E}$, each pruning-block point is retained with probability at least $1 - 2q \ge 31/32$. Therefore

$$\begin{aligned}\Pr\left(|\mathcal{Y}| < \frac{3n}{4}\right) &\le \exp\left(-2n\left(\frac{31}{32} - \frac{3}{4}\right)^2\right) \qquad \text{(Hoeffding)} \\ &\le e^{-n/11}.\end{aligned}$$

It remains to verify the three population claims. Since Eq. (45) and $q \le 1/64$ imply $\Pr(\|X - x^\dagger\|_2 \le \widehat{\tau}) \ge 31/32$, we have for any $\|v\|_2 = 1$ that

$$\begin{aligned}|\langle \widetilde{\mu}, v\rangle| &= \frac{\left|\mathbb{E}[\langle X, v\rangle \mathbf{1}\{\|X - x^\dagger\|_2 > \widehat{\tau}\}]\right|}{\Pr(\|X - x^\dagger\|_2 \le \widehat{\tau})} && \text{(definition of } \widetilde{\mu}\text{)} \\ &\le \frac{32}{31}\Pr(\|X - x^\dagger\|_2 > \widehat{\tau})^{\frac{\alpha-1}{\alpha}} && \text{(Hölder and } \Pr(\cdot) \ge 31/32\text{)} \\ &\le 2q^{\frac{\alpha-1}{\alpha}}. && \text{(by Eq. (45))}\end{aligned}$$

Taking the supremum over $\|v\|_2 = 1$ gives the bound on $\|\widetilde{\mu}\|_2$. The same $31/32$ lower bound on the retained probability also gives

$$\mathbb{E}|\langle Y - \widetilde{\mu}, v\rangle|^\alpha \le 2^{\alpha-1}\mathbb{E}|\langle Y, v\rangle|^\alpha + 2^{\alpha-1}|\langle \widetilde{\mu}, v\rangle|^\alpha \qquad (|a-b|^\alpha \le 2^{\alpha-1}(|a|^\alpha + |b|^\alpha))$$

$$\leq 2^{\alpha-1}\frac{32}{31} + 2^{\alpha-1}\|\widetilde{\mu}\|_2^\alpha \qquad \text{(conditioning defining } Y\text{)}$$
$$\leq 20. \qquad (\|\widetilde{\mu}\|_2 \text{ bound and } q \leq 1/64)$$

Finally, we have

$$\begin{aligned}
\mathrm{Tr}(\widetilde{\Sigma}) &\leq \mathbb{E}\|Y\|_2^2 \\
&\leq \frac{32}{31}\mathbb{E}\left[\|X\|_2^2 \mathbf{1}\{\|X - x^\dagger\|_2 \leq \widehat{\tau}\}\right] && \text{(conditioning defining } Y\text{)} \\
&\leq \frac{32}{31}\mathbb{E}\left[\|X\|_2^2 \mathbf{1}\{\|X\|_2 \leq \widehat{\tau} + 24\sqrt{d}\}\right] && (\|x^\dagger\|_2 \leq 24\sqrt{d}) \\
&\leq \frac{32}{31}(\widehat{\tau} + 24\sqrt{d})^{2-\alpha}\mathbb{E}\|X\|_2^\alpha && (u^2\mathbf{1}\{u \leq R\} \leq R^{2-\alpha}u^\alpha) \\
&\leq \frac{32}{31}\cdot\frac{4}{3}d^{\alpha/2}(\widehat{\tau} + 24\sqrt{d})^{2-\alpha} && \text{(by Eq. (47))} \\
&\leq 20d\,q^{-\frac{2-\alpha}{\alpha}}. && \text{(by Eq. (48))}
\end{aligned}$$

Recall that $q = \min\left\{\frac{1}{64}, \frac{6(d+\log(16/\delta))}{n}\right\}$. If $q = 6(d+\log(16/\delta))/n$, then $e^{-nq/6} \leq \delta/16$. If $q = 1/64$, then $e^{-nq/6} \leq e^{-n/1000}$. The same bound controls $e^{-nq/4}$, and $e^{-n/11} \leq e^{-n/1000}$. This proves all four claimed properties with failure probability at most $\delta/8 + 3e^{-n/1000}$. □

**Corollary 3** (Empirically tuned weak-moment mean estimation)**.** *Let $1 < \alpha < 2$, and let $X$ have mean $\mu \in \mathbb{R}^d$ and satisfy*

$$\sup_{\|v\|_2=1} \mathbb{E}|\langle X - \mu, v\rangle|^\alpha \leq 1.$$

*Assume $\delta \in (0, 1/2)$ and $\delta \geq 64e^{-n/1000}$. There is a modification of the mean estimation algorithm of Cherapanamjeri et al. (2022) that does not use $\alpha$ in the pruning step and satisfies, with probability at least $1-\delta$,*

$$\|\widehat{\mu} - \mu\|_2 \leq C\left[\sqrt{\frac{d}{n}} + \left(\frac{d+\log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}\right],$$

*where $C > 0$ is a numerical constant. The modification uses an independent threshold block of size $n$ to compute the empirical radius in Lemma 3, then replaces the deterministic pruning radius in Cherapanamjeri et al. (2022, Algorithm 5) by this empirical radius.*

*Proof.* The initial estimate is computed as in Cherapanamjeri et al. (2022, Algorithm 4). (Cherapanamjeri et al., 2022, Lemma B.1) gives

$$\|x^\dagger - \mu\|_2 \leq 24\sqrt{d}$$

with probability at least $1 - e^{-n/50}$. On this event, Lemma 3, applied to $X - \mu$ with failure probability $\delta/2$, gives the four pruning properties used in Cherapanamjeri et al. (2022, Lemma C.7): at least $3n/4$ retained samples, pruning bias at most $2q^{(\alpha-1)/\alpha}$, retained directional $\alpha$-moments bounded by 20, and retained trace covariance at most $20dq^{-(2-\alpha)/\alpha}$, where $q = \min\{1/64, 6(d+\log(32/\delta))/n\}$.

The rest of their proof uses Cherapanamjeri et al. (2022, Lemma C.7) only through these four properties. Replacing their deterministic pruning radius by the empirical radius from Lemma 3 therefore leaves the testing and descent steps unchanged, and yields

$$\|\widehat{\mu} - \mu\|_2 \leq C\left[\sqrt{\frac{d}{n}} + \left(\frac{d+\log(1/\delta)}{n}\right)^{\frac{\alpha-1}{\alpha}}\right].$$

The only additional cost is the threshold-block failure probability in Lemma 3. The condition $\delta \geq 64e^{-n/1000}$ absorbs this term, together with the initialization failure, into the final failure probability $\delta$. □

## F Proof of Theorem 3 (Lower Bound)

*Proof.* Let

$$\Delta = \frac{1}{4}\nu^{1/\alpha} d^{-1/2} \left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}}$$

and set

$$\Theta = \{\theta \in \mathbb{R}^d : (\Sigma^{1/2}\theta)_j \in \{\pm\Delta\} \text{ for every } j \in [d]\}.$$

For each $\theta \in \Theta$, define $\mathcal{D}_\theta$ as follows. The covariate is uniform on the $2d$ points

$$\left\{\pm\sqrt{d}\,\Sigma^{1/2} e_j : j \in [d]\right\}.$$

Conditionally on $x$, define the response by

$$y = \begin{cases} \dfrac{4n}{d} x^\top \theta, & \text{with probability } \dfrac{d}{4n}, \\ 0, & \text{otherwise.} \end{cases}$$

Since $n \geq d$, the probability $d/(4n)$ is at most one. Let $P_\theta = \mathcal{D}_\theta^{\otimes n}$ be the law of the sample under parameter $\theta$.

We first verify the assumptions. The covariance is

$$\mathbb{E}[xx^\top] = \frac{1}{d}\sum_{j=1}^{d} d\,\Sigma^{1/2} e_j e_j^\top \Sigma^{1/2} = \Sigma.$$

Moreover, for every $v \in \mathbb{R}^d$,

$$\mathbb{E}[(x^\top v)^4] = d\sum_{j=1}^{d} (\Sigma^{1/2} v)_j^4 \leq d\left(v^\top \Sigma v\right)^2.$$

Hence, the design satisfies Assumption 2 with $K_4 \leq d$. For any $\theta \in \Theta$ and any point $x$ in the support of the design, we have

$$|x^\top \theta| = \sqrt{d}\,\Delta = \frac{1}{4}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}}. \tag{49}$$

Moreover,

$$\mathbb{E}[y \mid x] = \frac{d}{4n} \cdot \frac{4n}{d} x^\top \theta = x^\top \theta.$$

The raw response moment is

$$\mathbb{E}[|y|^\alpha \mid x] = \frac{d}{4n}\left(\frac{4n}{d}\right)^\alpha |x^\top \theta|^\alpha = \frac{\nu}{4}. \tag{50}$$

This also verifies Assumption 4 for the parameter $\theta$. To bound the centered response moment, we write

$$\begin{aligned} |x^\top \theta|^\alpha &= \frac{\nu}{4^\alpha}\left(\frac{d}{n}\right)^{\alpha-1} && \text{(by Eq. (49))} \\ &\leq \frac{\nu}{4}, && (n \geq d,\ \alpha \geq 1) \end{aligned}$$

and hence

$$\begin{aligned} \mathbb{E}[|y - x^\top\theta|^\alpha \mid x] &\leq 2^{\alpha-1}\left(\mathbb{E}[|y|^\alpha \mid x] + |x^\top\theta|^\alpha\right) && \text{(triangle inequality)} \\ &\leq 2^{\alpha-2}\nu && \text{(Eq. (49) and Eq. (50))} \\ &\leq \nu. \end{aligned}$$

We now show that neighboring parameters are hard to distinguish. For $j \in [d]$ and $\theta \in \Theta$, define $\theta^{(j)} \in \Theta$ by flipping only the $j$-th whitened coordinate:

$$(\Sigma^{1/2}\theta^{(j)})_j = -(\Sigma^{1/2}\theta)_j, \qquad (\Sigma^{1/2}\theta^{(j)})_k = (\Sigma^{1/2}\theta)_k \quad k \neq j.$$

The one-sample laws under $\theta$ and $\theta^{(j)}$ differ only when $x$ selects coordinate $j$, which has probability $1/d$. Conditional on that event, the two reward laws agree at zero and place mass $d/(4n)$ on opposite nonzero atoms. Hence,

$$\begin{aligned} \mathrm{TV}(P_\theta, P_{\theta^{(j)}}) &\le n\,\mathrm{TV}(\mathcal{D}_\theta, \mathcal{D}_{\theta^{(j)}}) && \text{(TV tensorization upper bound)} \\ &\le n \cdot \frac{1}{d} \cdot \frac{d}{4n} = \frac{1}{4}. \end{aligned} \tag{51}$$

Consider an arbitrary estimator $\widehat{\theta}$ and set $A_j = \{(\Sigma^{1/2}\widehat{\theta})_j < 0\}$. For each coordinate $j$, define the average probability of a wrong sign decision by

$$e_j = \frac{1}{|\Theta|} \sum_{\theta:(\Sigma^{1/2}\theta)_j = \Delta} P_\theta(A_j) + \frac{1}{|\Theta|} \sum_{\theta:(\Sigma^{1/2}\theta)_j = -\Delta} P_\theta(A_j^c).$$

Reindexing the second sum in the definition of $e_j$ by the flip $\theta \mapsto \theta^{(j)}$ gives

$$\begin{aligned} e_j &= \frac{1}{|\Theta|} \sum_{\theta:(\Sigma^{1/2}\theta)_j = \Delta} \left[P_\theta(A_j) + P_{\theta^{(j)}}(A_j^c)\right] \\ &\ge \frac{1}{|\Theta|} \sum_{\theta:(\Sigma^{1/2}\theta)_j = \Delta} [1 - \mathrm{TV}(P_\theta, P_{\theta^{(j)}})] \ge \frac{3}{8}. \qquad \text{(by Eq. (51))} \end{aligned} \tag{52}$$

Let

$$M(\theta) = \sum_{j=1}^{d} \left[\mathbf{1}\{(\Sigma^{1/2}\theta)_j = \Delta\}\mathbf{1}_{A_j} + \mathbf{1}\{(\Sigma^{1/2}\theta)_j = -\Delta\}\mathbf{1}_{A_j^c}\right]$$

be the number of wrong signs. Summing Eq. (52) over $j$ gives

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{E}_{P_\theta} M(\theta) = \sum_{j=1}^{d} e_j \ge \frac{3d}{8}. \tag{53}$$

If a sign is wrong in coordinate $j$, then the $j$-th coordinate of $\Sigma^{1/2}(\widehat{\theta} - \theta)$ has absolute value at least $\Delta$. Hence,

$$\begin{aligned} \|\widehat{\theta} - \theta\|_\Sigma = \|\Sigma^{1/2}(\widehat{\theta} - \theta)\|_2 &\ge \Delta M(\theta)^{1/2} \\ &\ge \frac{\Delta}{\sqrt{d}} M(\theta). \qquad (M(\theta) \le d) \end{aligned} \tag{54}$$

The fact that $0 \le M(\theta) \le d$ further implies

$$\begin{aligned} \mathbb{E}_{P_\theta} M(\theta) &\le \frac{d}{4}\left(1 - P_\theta\Big(M(\theta) \ge \frac{d}{4}\Big)\right) + d\,P_\theta\Big(M(\theta) \ge \frac{d}{4}\Big) \\ &= \frac{d}{4} + \frac{3d}{4} P_\theta\left(M(\theta) \ge \frac{d}{4}\right). \end{aligned}$$

Averaging over $\theta \in \Theta$ and using Eq. (53), we obtain

$$\frac{3d}{8} \le \frac{d}{4} + \frac{3d}{4} \sup_{\theta \in \Theta} P_\theta\left(M(\theta) \ge \frac{d}{4}\right),$$

and hence

$$\sup_{\theta \in \Theta} P_\theta\left(M(\theta) \ge \frac{d}{4}\right) \ge \frac{3d/8 - d/4}{3d/4} = \frac{1}{6}. \tag{55}$$

On this event, Eq. (54) gives

$$\|\widehat{\theta} - \theta\|_\Sigma \geq \Delta\sqrt{\frac{d}{4}} = \frac{1}{8}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}}.$$

It follows that

$$\sup_{\theta\in\Theta} P_\theta\left(\|\widehat{\theta} - \theta\|_\Sigma \geq \frac{1}{8}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}}\right) \geq \frac{1}{6},$$

as claimed.

For the linear prediction bound, let $x'$ be an independent fresh covariate with the common marginal law $\mathcal{D}_X$. Conditionally on the training samples,

$$\left(\mathbb{E}_{x'}\left[x'^\top(\widehat{\theta} - \theta)\right]^2\right)^{1/2} = \|\widehat{\theta} - \theta\|_\Sigma = \|\Sigma^{1/2}(\widehat{\theta} - \theta)\|_2.$$

Combining this identity with Eq. (55) and Eq. (54) gives

$$\sup_{\theta\in\Theta} P_\theta\left(\left(\mathbb{E}_{x\sim\mathcal{D}_X}\left(x^\top(\widehat{\theta} - \theta)\right)^2\right)^{1/2} \geq \frac{1}{8}\nu^{1/\alpha}\left(\frac{d}{n}\right)^{\frac{\alpha-1}{\alpha}}\right) \geq \frac{1}{6}.$$

We now use the same sign argument to obtain the $L_2(\mathcal{D}_X)$ prediction lower bound for arbitrary predictors. For such a predictor $\widehat{h}$, replace $A_j$ in Eq. (52) by

$$\left\{\widehat{h}(\sqrt{d}\,\Sigma^{1/2}e_j) < \widehat{h}(-\sqrt{d}\,\Sigma^{1/2}e_j)\right\}.$$

With this modified sign test, the same pairing and TV bound as in Eq. (52) gives an average wrong-sign probability of at least $3/8$ for each coordinate. If this sign decision is wrong under parameter $\theta$, then the two target values on the pair $\{\sqrt{d}\,\Sigma^{1/2}e_j, -\sqrt{d}\,\Sigma^{1/2}e_j\}$ are ordered oppositely to the two predicted values, and therefore

$$\left|\widehat{h}(\sqrt{d}\,\Sigma^{1/2}e_j) - (\sqrt{d}\,\Sigma^{1/2}e_j)^\top\theta\right| + \left|\widehat{h}(-\sqrt{d}\,\Sigma^{1/2}e_j) - (-\sqrt{d}\,\Sigma^{1/2}e_j)^\top\theta\right| \geq 2\sqrt{d}\,\Delta.$$

Let $M(\theta)$ again denote the number of wrong signs. The same argument as that leading to Eq. (55) therefore gives

$$\sup_{\theta\in\Theta}\mathbb{E}_{P_\theta}M(\theta) \geq \frac{3d}{8}, \qquad \sup_{\theta\in\Theta} P_\theta\left(M(\theta) \geq \frac{d}{4}\right) \geq \frac{1}{6}.$$

On the event $\{M(\theta) \geq d/4\}$,

$$\mathbb{E}_{x\sim\mathcal{D}_X}\left(\widehat{h}(x) - x^\top\theta\right)^2 \geq \frac{d}{4}\cdot\frac{1}{2d}\cdot 2d\Delta^2 = \frac{d\Delta^2}{4},$$

and it follows that

$$\sup_{\theta\in\Theta} P_\theta\left(\mathbb{E}_{x\sim\mathcal{D}_X}\left(\widehat{h}(x) - x^\top\theta\right)^2 \geq \frac{d\Delta^2}{4}\right) \geq \frac{1}{6}.$$

Finally, substituting the value of $\Delta$ gives

$$\frac{d\Delta^2}{4} = \frac{1}{64}\nu^{2/\alpha}\left(\frac{d}{n}\right)^{\frac{2(\alpha-1)}{\alpha}}.$$

Since the predictor was arbitrary, this establishes the claimed prediction lower bound.

□

## G Further Related Work

Mean estimation of random variables over $\mathbb{R}$ for heavy-tailed distributions has been widely studied, and tight PAC bounds and efficient estimators that achieve such bounds are known. When the only assumption on the distribution is the existence of variance $\sigma^2$, the empirical mean is not sample-efficient as it can only guarantee the Chebyshev bound $\sigma\sqrt{\frac{1}{n\delta}}$ with probability $1-\delta$. Some widely used estimators that achieve the best estimation bounds up to constants are the following.

**Median of means** partitions the data into blocks and computes their mean, and then gives the median of these means as the estimator (Lugosi and Mendelson (2019a)). More formally,

$$\mathrm{MoM}_{\{B_1,\dots,B_k\}\vdash[n]}(X_1,\dots,X_n) = \mathrm{median}\left(\frac{1}{|B_1|}\sum_{i\in B_1} X_i,\dots,\frac{1}{|B_k|}\sum_{i\in B_k} X_i\right),$$

where $|B_i| \geq \lfloor \frac{n}{k} \rfloor$. If $k = \lceil 8\log(2/\delta) \rceil$ then the estimator achieves the tight PAC bound of $\mathcal{O}(\sigma\sqrt{\frac{\log(1/\delta)}{n}})$. If the variance is infinite and one only assumes that the central $\alpha$-moment ($\alpha \in (1,2)$) is bounded by some value $\nu$, then median-of-means gets a tight PAC bound of $\mathcal{O}\big(\nu^{\frac{1}{\alpha}}\big(\frac{\log(1/\delta)}{n}\big)^{\frac{\alpha-1}{\alpha}}\big)$ with the same $k$. Note that while the confidence bound requires knowing $\sigma$ or $(\alpha,\nu)$, the estimator itself only depends on $\delta$.

**Trimmed mean** filters or projects samples outside of a threshold and gives empirical means of the refined samples as the estimator. An order-optimal variant of trimmed mean (Oliveira and Resende (2025)) is achieved when partitioning the samples to two equal halves $((X_{0,i},X_{1,i})_{i=1}^n)$ of size $n$, re-arranging the samples of first half to be nondecreasing, then projecting all samples of second half on interval $[X_{0,\epsilon n}, X_{0,(1-\epsilon)n}]$ where $\epsilon = \frac{16\log(1/\delta)}{3n}$. This instance of trimmed mean is order-optimal in both finite variance and finite $\alpha$ central moment settings, and in addition is robust to certain adversarial contaminations.

**Optimal constants and instance-dependent estimators** While both MoM and trimmed mean estimators achieve the optimal PAC bound, the constants of their guarantees are large, and their PAC bounds are for worst-case distributions and not distribution dependent. For the former issue, Catoni (2012) proposed a novel estimator that generalizes the empirical mean and achieves tight constants; however, it requires knowing the variance $\sigma^2$ contrary to other estimators. Lee and Valiant (2022a) achieved tight constants without requiring prior knowledge of $\sigma^2$. Instance-dependent optimality was recently characterized in Dang et al. (2023), showing that median-of-means achieves instance-optimal rates up to constants.

Mean estimation for random variables over $\mathbb{R}^d$ has also been widely studied, and comes with significant additional challenges including computational efficiency of the estimator. When the covariance matrix $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^\top]$ exists, the tight PAC bound (sub-Gaussian) for $l_2$-distance is

$$\|\widehat{\mu}-\mu\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(1/\delta)}{n}}\right) \qquad \text{w.p. } 1-\delta.$$

There are many extensions of median-of-means to $\mathbb{R}^d$, as the median can be defined in several different ways on $\mathbb{R}^d$ (e.g., see Small (1990)). If we choose the median to be the geometric median (i.e. $\arg\min_m \sum_i \|m-X_i\|_2$), which is a convex objective and can be computed efficiently, then PAC bound of $\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{n}}$ is achieved (Minsker (2015)). However, this bound has a $\sqrt{\log 1/\delta}$ gap to the optimal bound. An optimal but computationally inefficient median-of-means variant was proposed in Lugosi and Mendelson (2019c), where the median is defined as

$$\operatorname*{argmin}_{m\in\mathbb{R}^d} \mathrm{radius}\left(\left\{x\in\mathbb{R}^d : \sum \mathbf{1}\left(\|X_i-x\| \leq \|X_i-m\|\right) \geq \frac{n}{2}\right\}\right).$$

Another formulation of the above (by Hopkins et al. (2020)) is the $(0.01, \left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(1/\delta)}{n}}\right)^2)$-combinatorial center of partition means $(Z_i)_{i=1}^k$, where $m$ is a $(\epsilon,\lambda)$-combinatorial center if for all unit vectors $v\in\mathbb{R}^d$

$$\sum_i \mathbf{1}[(Z_i-m)^\top v \geq \sqrt{\lambda}] \leq \epsilon k$$

Hopkins (2020) proposed a computationally efficient variant of the estimator by semidefinite relaxations of the optimization objective.

For the mean estimation problem, for distributions satisfying a finite directional $\alpha$-moment bound with $1 < \alpha < 2$, Cherapanamjeri et al. (2022) give a computationally efficient mean estimator with matching upper and lower bounds. In contrast, for linear regression, works adopting weak moment assumptions have mainly focused on the finite-variance regime, typically under fixed or random covariate designs. The seminal work of Audibert and Catoni (2011) achieved sub-Gaussian bounds with a $L_2 - L_4$ norm equivalence assumption. Later Lugosi and Mendelson (2019b) proposed a median-of-means based estimator that achieved a sub-Gaussian rate without any extra assumption other than existence of the fourth moment. However, these estimators are computationally inefficient. Cherapanamjeri et al. (2020) proposed an SDP relaxation of the estimator that achieves optimal sub-Gaussian rates with a polynomial-time complexity but requires stronger assumptions on the covariates. Sun et al. (2020) proposed a Huber loss based estimator but with suboptimal rates. Later, Pensia et al. (2025) improved the Huber-loss based estimation to achieve sub-Gaussian rates if the noise variables are assumed to be symmetric, or i.i.d. and independent from the covariates. Moreover, Oliveira and Resende (2025) proposed a trimmed mean estimator:

$$\hat{\theta}_k = \arg\min_{\theta\in\mathbb{R}^d} \max_{\psi\in\mathbb{R}^d} \frac{1}{n}\sum_i \mathrm{trim}_k(\ell(\theta^\top x_i - y_i) - \ell(\psi^\top x_i - y_i)),$$

where trimming is done on the interval of $k+1$ to $n-k$ smallest values. For squared loss $\ell$ and $k = 8\log(4/\delta)$, it achieves the following estimation rate:

$$\mathbb{E}\left[|y - \hat{\theta}^\top x|^2\right]^{1/2} \le \mathcal{O}\left(g(\mathcal{D}_X)\sigma\sqrt{\frac{d+\log(1/\delta)}{n}}\right),$$

where $g(\mathcal{D}_X) = \sup_v \frac{\mathbb{E}\left[|v^\top x|^2\right]^{1/2}}{\mathbb{E}[|v^\top x|]}$. This has a $g(\mathcal{D}_X)$ term with the lower bound achieved by Gaussian noise. El Hanchi et al. (2024) studied $\ell_p$-error for $p \ge 2$ and proposed an algorithm that achieves sub-Gaussian rates. Beyond the bounded variance error regime, to our knowledge the only known result for regression is Sun et al. (2020) which is suboptimal even in the finite variance case.

A generalization of the above estimation problems is stochastic optimization, where given a distribution $\mathcal{F}$ over functions $f : \mathbb{R}^d \to \mathbb{R}$, we want to estimate the minimizer of the expected function, i.e. $\mathrm{argmin}_w \mathbb{E}_{f\sim\mathcal{F}}[f(w)]$. For linear regression with any loss $\ell$, we can formulate it as a stochastic optimization problem where function $f_{(x,y)}(w) := \ell(x^\top w - y)$ is drawn when $(x, y) \sim \mathcal{D}$. Das et al. (2024) proposed clipped SGD, which achieves the best known upper bound of

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n} + \frac{\sqrt{\mathrm{Tr}(\Sigma)\|\Sigma\|_2}\log(\log n/\delta)}{n}}.$$

Achieving a sub-Gaussian rate for stochastic optimization remains open.

## H Numerical Experiment

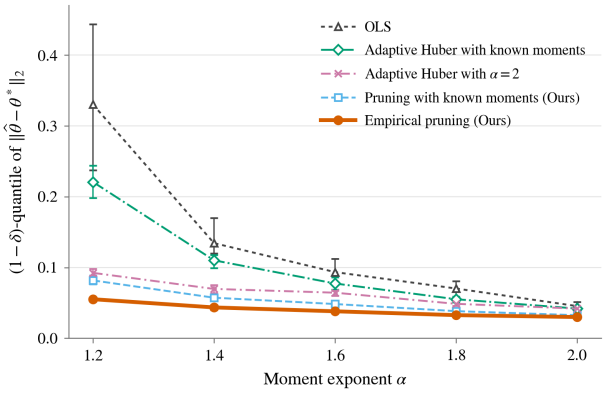


Figure 1: Empirical $(1-\delta)$-quantile of $\|\widehat{\theta} - \theta^*\|_2$, with $\delta = 1/\log n \approx 0.1316$. Error bars show 95% bootstrap confidence intervals for each quantile, computed using 10,000 bootstrap samples of the 200 estimation errors.

In this section, we present a simple proof-of-concept numerical experiment illustrating adaptation to unknown $\alpha$ and $\nu$. We compare the estimation error of empirical pruning with that of pruning based on known moments as the noise becomes heavier-tailed.

**Setting.** We consider linear regression with $n = 2{,}000$, $d = 8$, and $\theta^* = \mathbf{1}$. The covariates have independent standardized Student-$t_5$ coordinates, so $\Sigma = I$ and $K_4 = 9$. The noise is independent of the covariates and follows a centered Pareto distribution with shape $\alpha + 0.1$, rescaled to satisfy $\mathbb{E}|\varepsilon|^\alpha = \nu = 1$. We vary $\alpha \in \{1.2, 1.4, 1.6, 1.8, 2\}$ and set $\delta = 1/\log n$. Each experiment is repeated 200 times, with all methods fitted to the same data in each run. We compare our algorithm with OLS and adaptive Huber regression (Sun et al., 2020), with the latter using either the true $\alpha$ or $\alpha = 2$.

**Results.** In Figure 1, we see that empirical pruning achieves smaller estimation errors than pruning with known moments throughout the tested range of $\alpha$. Both pruning methods have smaller errors than OLS and adaptive Huber regression, with larger differences as $\alpha$ decreases.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The claims are validated by detailed proofs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the results.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides detailed assumptions and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The details of numerical experiment in Appendix H are stated and is reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The minimal numerical experiment is of very small scale and only included as a proof of concept with minimal code, and the paper's focus is the theoretical contributions.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of our numerical experiments are in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1 reports 95% bootstrap confidence intervals for the empirical error quantiles. The caption specifies the number of runs and bootstrap samples.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The numerical experiment is very small scale and can be run on any device with no need of compute resource.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theoretical work. There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a theoretical work. The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.