A New Principle for Tuning-Free Huber Regression

Lili Wang[*], Chao Zheng[†], Wen Zhou[‡] and Wen-Xin Zhou[§]

Abstract

The robustification parameter, which balances bias and robustness, has played a critical role in the construction of sub-Gaussian estimators for heavy-tailed data. Although it can be tuned by cross-validation in traditional practice, in large scale statistical problems such as high dimensional covariance matrix estimation and large scale multiple testing, the number of robustification parameters scales with the size of the problem so that cross-validation can be computationally unaffordable. In this paper, we propose a new data-driven principle to select the robustification parameter for Huber-type sub-Gaussian estimators in three fundamental problems: mean estimation, linear regression and, sparse regression in high dimensions. Our proposal is guided by non-asymptotic deviation analysis, and is conceptually different from cross-validation which relies on the mean squared error to assess the fit. The promising performance of the proposed methods, apart from the theoretical justifications, are further illustrated with extensive numerical experiments and real data analysis.

Keywords: Data adaptive; Heavy tails; Huber loss; M-estimator; Tuning parameters

[*]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China. E-mail: lili-wang@zjgsu.edu.cn.

[†]Department of Mathematics and Statistics, Lancaster University, LA1 4YF, UK. E-mail: c.zheng5@lancaster.ac.uk.

[‡]Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, USA. E-mail: riczw@stat.colostate.edu.

[§]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: wez243@ucsd.edu.

1 Introduction

Data subject to heavy-tailed and/or skewed distributions are frequently observed in many areas, ranging from microarray experiments (Purdom and Holmes, 2005), neuroimaging (Eklund, Nichols and Knutsson, 2016) to finance (Cont, 2001). Rigorously, a random variable X is heavy-tailed if its tail probability $\mathbb{P}(|X| > t)$ decays to zero polynomially in $1/t$ as $t \to \infty$, or equivalently, if X has finite polynomial-order moments. The connection between moment and tail probability is revealed by the well known equality $\mathbb{E}(|X|^k) = k \int_0^\infty t^{k-1} \mathbb{P}(|X| > t)\, dt$ for any $k \geq 1$. When the sampling distribution has only a small number of finite moments, with high chance some observations will deviate wildly from their mean. We refer to such observations as distributional outliers. In contrast, data generated from a Gaussian or sub-Gaussian distribution (Vershynin, 2012) are strongly concentrated around their expected value, and the occurrence of even a single outlier will be rare.

Heavy-tailed data bring new challenges to conventional statistical methods. For linear models, regression estimators based on the least squares loss are suboptimal, both theoretically and empirically, in the presence of heavy-tailed errors. We refer to Catoni (2012) for a deviation analysis, showing that the deviations of the empirical mean can be much worse for non-Gaussian samples than for Gaussian ones. More broadly, this study exemplifies the pitfalls of asymptotic studies in statistics and inspires new thoughts about the notions of optimality commonly used to assess the performance of estimators. In particular, minimax optimality under mean squared error does not quite capture the influence of extreme behaviors of estimators. However, these rare events may have catastrophically negative impacts in practice, leading to wrong conclusions or false discoveries. Since Catoni (2012), non-asymptotic deviation analysis has drawn considerable attention and it is becoming increasingly important to construct sub-Gaussian estimators for heavy-tailed data; see, for example, Brownlees, Joly and Lugosi (2015), Minsker (2015, 2018), Hsu and Sabato (2016), Devroye et al. (2016), Lugosi and Mendelson (2016), Fan, Li and Wang (2017), Lugosi and Mendelson (2019), Lecué and Lerasle (2017) and Zhou et al. (2018), among others.

For linear models, Fan, Li and Wang (2017) and Zhou et al. (2018) proposed Huber-type estimators in both low and high dimensional settings and derived non-asymptotic deviation bounds for the estimation error. To implement either Catoni's or Huber-type method, a tuning parameter τ needs to be specified in advance to ultimately balance between resistance to outliers (robustness) and bias of the estimation. Deviation analysis suggests that this tuning parameter, which we refer to as the robustification parameter, should adapt to the sample

size, dimension, variance of noise and confidence level. Calibration schemes are typically based on cross-validation or Lepski's method, which can be computationally intensive especially for large-scale inference and high dimensional estimation problems where the number of parameters may be exponential in the number of observations. For example, Avella-Medina et al. (2018) proposed adaptive Huber estimators for estimating high dimensional covariance and precision matrices. For a $d \times d$ covariance matrix, although every entry can be robustly estimated by a Huber-type estimator with τ chosen via cross-validation, the overall procedure involves as many as d^2 tuning parameters and therefore the cross-validation method will soon become computationally expensive as d grows. Efficient tuning is important for not only the problem's own interest, but also its applications in a broader context.

This paper develops data-driven Huber-type methods for mean estimation, linear regression, and sparse regression in high dimensions. For each problem, we first provide sub-Gaussian concentration bounds for the Huber-type estimator under minimal moment condition on the errors. These non-asymptotic results are intended primarily to guide the choice of key tuning parameters. Some of them are of independent interest and improve the existing results by weakening the sample size scaling. Secondly, we propose a novel data-driven principle to calibrate the robustification parameter $\tau > 0$ in the Huber loss

$$\ell_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \le \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases} \tag{1}$$

In Huber's original proposal (Huber, 1981), τ is chosen as 1.345 so that the asymptotic efficiency of the estimator is 95% for the normal model. Since then, this has become the default setting and also find its use in high dimensional statistics even though the asymptotic efficiency is no longer well defined; see, for example, Lambert-Lacroix and Zwald (2011), Elsener and van de Geer (2018) and Loh (2017). Guided by non-asymptotic deviation analysis, our proposed τ grows with sample size for bias-robustness trade-off. For linear regression under different regimes, the optimal tuning parameter τ depends on the dimension d differently: $\tau \sim \sigma\sqrt{(n/d)}$ in the low dimensional setting (i.e. d/n is small) and $\tau \sim \sigma\sqrt{n/\log(d)}$ in high dimensions. Thirdly, we provide simple and fast algorithms to implement the data-driven procedure under various scenarios.

The remainder of this paper is organized as follows. In Section 2, we revisit the fundamental mean estimation problem. Motivated by a careful analysis of the truncated sample mean, we introduce a novel data-driven adaptive Huber estimator. We extend this data-driven tuning scheme to robust regression in Section 3 under both low and high dimensional

settings. Extensive numerical experiments are given in Section 4 to demonstrate the finite sample performance of the proposed procedure. All the proofs, together with additional technical details, are relegated to the supplementary files.

2 Robust data-adaptive mean estimation

2.1 Motivation

To motivate our proposed data-driven scheme for Huber-type estimators, we start with revisiting the mean estimation problem. Let $X_1, \ldots, X_n$ ($n \geq 2$) be independent and identically distributed (i.i.d.) random variables from X with mean μ and finite variance $\sigma^2 > 0$. The sample mean, denoted as $\bar{X}_n$, is the most natural estimator for μ. However, it severely suffers from not being robust to heavy-tailed sampling distributions (Catoni, 2012). In order to cancel, or at least dampen, the erratic fluctuations in $\bar{X}_n$ which are more likely to occur if the distribution of X is heavy-tailed, we consider the truncated sample mean

$$m_\tau = \frac{1}{n}\sum_{i=1}^{n} \psi_\tau(X_i) \tag{2}$$

for some $\tau > 0$, where

$$\psi_\tau(x) = \text{sign}(x)\min(|x|, \tau), \quad x \in \mathbb{R} \tag{3}$$

is a truncation function. Here, τ is a tuning parameter that controls the bias and robustness of m_τ. To see this, note that the bias, which is given by Bias $:= \mathbb{E}(m_\tau) - \mu$, satisfies $|\text{Bias}| = |\mathbb{E}\{X - \text{sign}(X)\tau\}I(|X| > \tau)| \leq \tau^{-1}\mathbb{E}(X^2)$. Regarding (distributional) robustness, the following result shows that the truncated sample mean with a properly chosen τ is a sub-Gaussian estimator as long as the second moment is finite.

Proposition 1. Assume that $v_2 := \sqrt{\mathbb{E}(X^2)}$ is finite. For any $z > 0$,

(i) the truncated mean m_τ with $\tau = v\sqrt{n/z}$ for some $v \geq v_2$ satisfies $\mathbb{P}\{|m_\tau - \mu| \geq 2v\sqrt{z/n}\} \leq 2e^{-z}$;

(ii) the truncated mean m_τ with $\tau = cv_2\sqrt{n/z}$ for some $0 < c \leq 1$ satisfies $\mathbb{P}\{|m_\tau - \mu| \geq 2(v_2/c)\sqrt{z/n}\} \leq 2e^{-z/c^2}$.

Proposition 1 shows that how the procedure would perform under various idealized scenarios, as such providing guidance on the choice of τ. Here $z > 0$ is a user-specified parameter that controls the confidence level; see further discussions before Remark 2. Given a properly tuned τ, the sub-Gaussian performance is achieved; conversely, if the resulting estimator performs well, it means that the data are truncated at the right level and therefore can be further exploited. An ideal τ is such that the sample mean of truncated data $\psi_\tau(X_1), \ldots, \psi_\tau(X_n)$ serves as a good estimator of μ. The influence of distributional outliers is weakened due to proper truncation. At the same time, we may expect that the empirical second moment for the same truncated data will provide a reasonable estimate of v_2^2. Motivated by this observation, we propose to choose τ by solving the equation

$$\tau = \left\{ \sum_{i=1}^{n} \psi_\tau^2(X_i) \right\}^{1/2} \sqrt{\frac{n}{z}}, \quad \tau > 0,$$

which is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\psi_\tau^2(X_i)}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{4}$$

We will show that under mild conditions, equation (4) has a unique solution, denoted as $\widehat{\tau}_z$, which gives rise to a data-driven mean estimator

$$m_{\widehat{\tau}_z} = \frac{1}{n} \sum_{i=1}^{n} \min(|X_i|, \widehat{\tau}_z) \operatorname{sign}(X_i). \tag{5}$$

To understand the statistical property of $\widehat{\tau}_z$, consider the population version of (4):

$$\frac{\mathbb{E}\{\psi_\tau^2(X)\}}{\tau^2} = \frac{\mathbb{E}\{\min(X^2, \tau^2)\}}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{6}$$

The following result establishes existence and uniqueness of the solution to (6).

Proposition 2. Assume that $v_2 = \sqrt{\mathbb{E}(X^2)}$ is finite.

(i) Provided $0 < z < n\mathbb{P}(|X| > 0)$, equation (6) has a unique solution, denoted by τ_z, which satisfies

$$[\mathbb{E}\{\min(X^2, q_{z/n}^2)\}]^{1/2} \sqrt{\frac{n}{z}} \le \tau_z \le v_2 \sqrt{\frac{n}{z}},$$

where $q_\alpha := \inf\{t : \mathbb{P}(|X| > t) \le \alpha\}$ is the upper α-quantile of $|X|$.

(ii) Let $z = z_n > 0$ satisfy $z_n \to \infty$ and $z = o(n)$. Then $\tau_z \to \infty$ and $\tau_z \sim v_2 \sqrt{n/z}$ as $n \to \infty$.

We now move to the sample version. As a direct consequence of Proposition 2, the following result ensures existence and uniqueness of the solution to equation (4).

Proposition 3. Provided $0 < z < \sum_{i=1}^n I(|X_i| > 0)$, equation (4) admits a unique solution.

Throughout, denote $\widehat{\tau}_z$ the solution to (4), which is unique and positive whenever $z < \sum_{i=1}^n I(|X_i| > 0)$. For completeness, we set $\widehat{\tau}_z = 0$ on the event $\{z \geq \sum_{i=1}^n I(|X_i| > 0)\}$. If the distribution of X satisfies $\mathbb{P}(X = 0) = 0$, then $\widehat{\tau}_z > 0$ with probability one, provided $0 < z < n$. With both τ_z and $\widehat{\tau}_z$ well defined, we investigate the statistical property of $\widehat{\tau}_z$.

Theorem 1. Assume $\mathbb{E}(X^2) < \infty$ and $\mathbb{P}(X = 0) = 0$. For any $1 \leq z < n$ and $0 < r < 1$, we have

$$\mathbb{P}(|\widehat{\tau}_z/\tau_z - 1| \geq r) \leq e^{-a_1^2 r^2 z^2/(2z + 2a_1 rz/3)} + e^{-a_2^2 r^2 z/2} + 2e^{-(a_1 \wedge a_2)^2 z/8}, \tag{7}$$

where

$$a_1 = a_1(z, r) = \frac{P(\tau_z)}{2Q(\tau_z)} \frac{2 + r}{(1 + r)^2} \quad \text{and} \quad a_2 = a_2(z, r) = \frac{P(\tau_z - \tau_z r)}{2Q(\tau_z)} \frac{2 - r}{1 - r} \tag{8}$$

with $P(t) = \mathbb{E}\{X^2 I(|X| \leq t)\}$ and $Q(t) = \mathbb{E}\{\psi_t^2(X)\}$.

Remark 1. Here we give some direct implications of Theorem 1.

(i) Let $z = z_n \geq 1$ satisfy $z = o(n)$ and $z \to \infty$ as $n \to \infty$. By Proposition 2, $\tau_z \to \infty$ and $\tau_z \sim v_2\sqrt{n/z}$, which further implies $P(\tau_z) \to v_2^2$ and $Q(\tau_z) \to v_2^2$ as $n \to \infty$.

(ii) With $r = 1/2$ and $z = \log^\kappa(n)$ for some $\kappa \geq 1$ in (7), the constants $a_1 = a_1(z, 1/2)$ and $a_2 = a_2(z, 1/2)$ satisfy $a_1 \to 5/9$ and $a_2 \to 3/2$ as $n \to \infty$. The resulting $\widehat{\tau}_z$ satisfies that with probability approaching one, $\tau_z/2 \leq \widehat{\tau}_z \leq 3\tau_z/2$.

We end this section with a uniform deviation bound for m_τ. Uniformity of the rate over a neighborhood of the optimal tuning scale requires an additional $\log(n)$-factor. As a result, we show that the data-driven estimator $m_{\widehat{\tau}_z}$ is tightly concentrated around the mean with high probability.

Theorem 2. For $z \geq 1$, let $\tau_z^* = v_2\sqrt{n/z}$. Then with probability at least $1 - 2ne^{-z}$,

$$\sup_{\tau_z^*/2 \leq \tau \leq 3\tau_z^*/2} |m_\tau - \mu| \leq 4v_2(z/n)^{1/2} + v_2 n^{-1/2}. \tag{9}$$

Let $z = 2\log(n)$ and $\widehat{\tau}_z$ be the solution to (4). The mean estimator $m_{\widehat{\tau}_z}$ given in (5) satisfies $|m_{\widehat{\tau}_z} - \mu| \leq 4v_2\sqrt{2\log(n)/n} + v_2 n^{-1/2}$ with probability at least $1 - c_1 n^{-c_2}$ for all sufficiently large n, where $c_1, c_2 > 0$ are absolute constants.

2.2 Adaptive Huber estimator

For the truncation method, even with the theoretically desirable tuning parameter $\tau = v_2\sqrt{n/z}$, the deviation of the resulting estimator only scales with v_2 rather than the standard deviation σ. The optimal deviation, which is enjoyed by the sample mean with sub-Gaussian data, is of order $\sigma\sqrt{z/n}$. To achieve such an optimal order, Fan, Li and Wang (2017) modified Huber's method to construct an estimator that exhibits fast (sub-Gaussian type) concentration under finite variance condition.

Specified in (1), the Huber loss is continuously differentiable with $\ell_\tau'(x) = \psi_\tau(x)$ where $\psi_\tau(\cdot)$ is given in (3). Given a sample of observations $X_1, \ldots, X_n$ from X with mean μ and finite variance σ^2, Huber's estimator is obtained by solving the optimization problem

$$\widehat{\mu}_\tau = \operatorname*{argmin}_{\theta\in\mathbb{R}} \sum_{i=1}^n \ell_\tau(X_i - \theta), \tag{10}$$

or equivalently, $\widehat{\mu}_\tau$ is the unique solution to

$$0 = \sum_{i=1}^n \psi_\tau(X_i - \theta) = \sum_{i=1}^n \min(|X_i - \theta|, \tau)\,\mathrm{sign}(X_i - \theta). \tag{11}$$

We refer to Catoni (2012) for a general class of robust mean estimators. The following result is Theorem 5 in Fan, Li and Wang (2017), which shows the exponential-type concentration of $\widehat{\mu}_\tau$ when τ is properly calibrated.

Proposition 4. Let $z > 0$ and $v \geq \sigma$. Provided $n \geq 8z$, $\widehat{\mu}_\tau$ with $\tau = v\sqrt{n/z}$ satisfies the bound $|\widehat{\mu}_\tau - \mu| \leq 4v\sqrt{z/n}$ with probability at least $1 - 2e^{-z}$.

Proposition 4 indicates that a theoretically desirable tuning parameter for the Huber estimator is $\tau \sim \sigma\sqrt{n/z}$. Motivated by the data-driven approach proposed in Section 2.1, we consider the following modification of (6):

$$\frac{\mathbb{E}\{\psi_\tau^2(X-\mu)\}}{\tau^2} = \frac{\mathbb{E}[\min\{(X-\mu)^2, \tau^2\}]}{\tau^2} = \frac{z}{n}, \quad \tau > 0. \tag{12}$$

According to Proposition 2, provided $0 < z < n\mathbb{P}(X \neq \mu)$, this equation has a unique solution, denoted by $\tau_{z,\mu}$, which satisfies

$$\sqrt{\mathbb{E}\left[\min\{(X-\mu)^2, \bar{q}_{z/n}\}\right]}\sqrt{\frac{n}{z}} \leq \tau_{z,\mu} \leq \sigma\sqrt{\frac{n}{z}},$$

where $\bar{q}_\alpha = \inf\{t : \mathbb{P}(|X-\mu| > t) \leq \alpha\}$ is the upper α-quantile of $|X-\mu|$. From a large sample perspective, if $z = z_n$ satisfies $z \to \infty$ and $z = o(n)$, then $\tau_{z,\mu} \to \infty$ and $\tau_{z,\mu} \sim \sigma\sqrt{n/z}$ as $n \to \infty$.

In light of (11) and (12), a clearly motivated data-driven estimate of μ can be obtained by solving the following system of equations:

$$\begin{cases} f_1(\theta,\tau) := \sum_{i=1}^n \psi_\tau(X_i-\theta) = 0, \\ f_2(\theta,\tau) := n^{-1}\sum_{i=1}^n \min\{(X_i-\theta)^2,\tau^2\}/\tau^2 - n^{-1}z = 0, \end{cases} \quad \theta\in\mathbb{R}, \tau>0. \tag{13}$$

Observe that for any given $\tau>0$, equation $f_1(\cdot,\tau)=0$ always admits a unique solution, and for any given θ, equation $f_2(\theta,\cdot)=0$ has a unique solution provided $z<\sum_{i=1}^n I(X_i\neq\theta)$. With initial values $\theta^{(0)}=\bar{X}_n$ and $\tau^{(0)}=\widehat{\sigma}_n\sqrt{n/z}$ where $\widehat{\sigma}_n^2$ denotes the sample variance, we solve (13) successively by computing a sequence of solutions $\{(\theta^{(k)},\tau^{(k)})\}_{k\geq 1}$ that fulfill $f_2(\theta^{(k-1)},\tau^{(k)})=0$ and $f_1(\theta^{(k)},\tau^{(k)})=0$ for $k\geq 1$. For a predetermined tolerance level ϵ, we stop the algorithm within the ℓ-th iteration step when $\max\{|\theta^{(\ell)}-\theta^{(\ell-1)}|,|\tau^{(\ell)}-\tau^{(\ell-1)}|\}\leq\epsilon$. We then use $\theta^{(\ell)}$ as our final robust estimator of μ.

In the case of $z=1$, we see that the algorithm stops in the first iteration and delivers the solution $\bar{X}_n$. According to the results in Section 2.1, if $z\geq 1$ is fixed, there is no net improvement in terms of robustness; instead, we should let $z=z_n$ slowly grow with the sample size to gain robustness without introducing extra bias. Specifically, we choose $z=\log(n)$ throughout the numerical experiments carried out in this paper.

Remark 2. Our proposed procedure has some similarities to the estimator considered in Bickel (1975), which is obtained as the solution of $\sum_{i=1}^n \psi_{\widehat{\sigma}}(X_i-\theta)=0$, where $\widehat{\sigma}$ is chosen independently as the normalized interquartile range

$$\widehat{\sigma}^{(1)} = \{X_{(n-[n/4]+1)} - X_{([n/4])}\}/2\Phi^{-1}(3/4)$$

or the symmetrized interquartile range

$$\widehat{\sigma}^{(2)} = \text{median}\{|X_i-m|\}/\Phi^{-1}(3/4),$$

where $X_{(1)}<\cdots<X_{(n)}$ are the order statistics and m is the sample median. Provided that X has a symmetric distribution, the consistency of $\widehat{\sigma}^{(1)}$ or $\widehat{\sigma}^{(2)}$ can be established. Unlike this classical approach, we waive the symmetry requirement by allowing the robustification parameter to diverge to reduce the bias that is induced by the Huber loss when the distribution is asymmetric. Another difference is that Bickel's proposal is a two-step method that estimates the scale and location separately, whereas our procedure estimates μ and calibrates τ simultaneously by solving a system of equations. In fact, as a direct extension of the idea in Section 2.1, we may also tune τ independently from estimation by solving

$$\binom{n}{2}^{-1}\sum_{1\leq i<j\leq n}\frac{\min\{(X_i-X_j)^2/2,\tau^2\}}{\tau^2} = \frac{z}{n}, \quad \tau>0.$$

Let X' be an independent copy of X. Then the population version of this equation is $\mathbb{E}\left[\min\{(X-X')^2/2, \tau^2\}\right]\tau^{-2} = z/n$, whose solution is unique under mild conditions and scales as $\sigma\sqrt{n/z}$.

3 Robust data-adaptive linear regression

In this section, we extend the proposed data-driven method for robust mean estimation to regression problems. Consider the linear regression model

$$Y_i = \beta_0^* + \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where Y_i's represent response variables, $\boldsymbol{X}_i$'s are d-dimensional vector of covariates, β_0^* and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ are the intercept and vector of regression coefficients, respectively, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent regression errors with zero mean and finite variance. For simplicity, we also introduce $\boldsymbol{Z}_i = (1, \boldsymbol{X}_i^\intercal)^\intercal$ for $i = 1, \ldots, n$ and use $\boldsymbol{\theta}^* = (\beta_0^*, \boldsymbol{\beta}^{*\intercal})^\intercal$ to denote the total vector of unknown parameters. The goal is to estimate $\boldsymbol{\theta}^*$ from observed data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$.

3.1 Adaptive Huber regression in low dimensions

We start with the low-dimensional setting where $d \ll n$. In the presence of heavy-tailed errors, finite sample properties of the least squares method are suboptimal both theoretically and empirically. The necessity of finding robust alternatives to the least squares was discussed in Huber (1973) under Huber's ϵ-contamination model. Under different hypotheses that allow for heavy-tailed distributions, we refer to Audibert and Catoni (2011) and Sun, Zhou and Fan (2017) for non-asymptotic analysis of Huber-type robust regression methods; the former focuses on the excess risk bounds and the latter provides deviation bounds for the estimator along with non-asymptotic Bahadur representations.

Given any $\tau > 0$, Huber's M-estimator is defined as

$$\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau^\intercal)^\intercal \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}), \tag{15}$$

where $\ell_\tau(\cdot)$ is given in (1). By the convexity of Huber loss, the solution to (15) is uniquely determined via the first-order condition: $\sum_{i=1}^n \psi_\tau(Y_i - \boldsymbol{Z}_i^\intercal \widehat{\boldsymbol{\theta}}_\tau)\boldsymbol{Z}_i = \boldsymbol{0}$.

Most of the desirable features of Huber's method are established under the assumption that the distribution of errors is symmetric around zero. Due to asymmetry, the bias induced

by the Huber loss is nonnegligible. To make this statement precise, note that $\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau^\intercal)^\intercal$ is a natural M-estimator of

$$\boldsymbol{\theta}_\tau^* = (\beta_{0,\tau}^*, \boldsymbol{\beta}_\tau^{*\intercal})^\intercal = \underset{(\beta_0, \boldsymbol{\beta}^\intercal)^\intercal \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^n \mathbb{E}\{\ell_\tau(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})\}, \tag{16}$$

whereas the true parameters β_0^* and $\boldsymbol{\beta}^*$ are identified as $\operatorname{argmin}_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \mathbb{E}\{(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2\}$. For a fixed $\tau > 0$, although $\widehat{\beta}_{0,\tau}$ and $\widehat{\boldsymbol{\beta}}_\tau$ are robust estimates of $\beta_{0,\tau}^*$ and $\boldsymbol{\beta}_\tau^*$, respectively, in general $(\beta_{0,\tau}, \boldsymbol{\beta}_\tau)$ differs from $(\beta_0^*, \boldsymbol{\beta}^*)$, as unveiled by the following result.

Proposition 5. Assume that ε and $\boldsymbol{X}$ are independent, and that the function $\alpha \mapsto \mathbb{E}\{\ell_\tau(\varepsilon - \alpha)\}$ has a unique minimizer, denoted by $\alpha_\tau = \operatorname{argmin}_{\alpha \in \mathbb{R}} \mathbb{E}\{\ell_\tau(\varepsilon - \alpha)\}$, which satisfies

$$\mathbb{P}(|\varepsilon - \alpha_\tau| \le \tau) > 0. \tag{17}$$

Assume further that $\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$ is positive definite. Then we have

$$\beta_{0,\tau}^* = \beta_0^* + \alpha_\tau \quad \text{and} \quad \boldsymbol{\beta}_\tau^* = \boldsymbol{\beta}^*. \tag{18}$$

Moreover, α_τ with $\tau > \sigma$ satisfies the bound

$$|\alpha_\tau| \le \frac{\sigma^2 - \mathbb{E}\{\psi_\tau^2(\varepsilon)\}}{1 - \tau^{-2}\sigma^2} \frac{1}{2\tau}. \tag{19}$$

Note also that Huber loss minimization is equivalent to the following penalized least squares problem (She and Owen, 2011):

$$(\widehat{\boldsymbol{\mu}}_\tau, \widehat{\boldsymbol{\theta}}_\tau) = \underset{\boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n (Y_i - \mu_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2 + \tau \sum_{i=1}^n |\mu_i| \right\}, \tag{20}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\intercal$ and $\widehat{\boldsymbol{\theta}}_\tau$ here coincides with that in (15). The loss function in (20) can be written as $\sum_{i=1}^n (Y_i - \mu_i - \beta_0 - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2/2 + \tau \sum_{i=1}^n |\mu_i|$. This explains from a different perspective that why the bias arises only at the intercept. The larger the value of τ is, the sparser the $\widehat{\boldsymbol{\mu}}_\tau$ is and therefore the smaller the estimation bias is.

The message delivered by Proposition 5 calls attention to intercept estimation, a problem of independent interest that needs to be treated with greater caution. If the distribution of ε is asymmetric, α_τ is typically non-zero for any $\tau > 0$: the smaller the τ is, the larger the bias becomes and so is the prediction error. To balance bias and robustness, in the following we propose two modifications, one-step and two-step, of Huber's method that are robust against heavy-tailed and asymmetric error distributions and meanwhile maintain high efficiency in the normal case.

3.1.1 One-step method

As pointed out in Zhou et al. (2018), there is an inherent bias-robustness trade-off in the choice of τ, which should adapt to the sample size, dimension and the variance of noise. Theorem 3 below makes this statement precise. To begin with, we impose the following moment conditions.

Condition 1. The covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors from $\boldsymbol{X}$. There exists some constant $A_0 > 0$ such that for any $\boldsymbol{u} \in \mathbb{R}^{d+1}$ and $t \in \mathbb{R}$, $\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{z} \rangle| \geq A_0 \|\boldsymbol{u}\|_2 \cdot t) \leq 2e^{-t^2}$, where $\boldsymbol{z} = \mathbf{S}^{-1/2}\boldsymbol{Z}$ and $\mathbf{S} = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$ is positive definite. The regression errors ε_i are independent and satisfy $\mathbb{E}(\varepsilon_i | \boldsymbol{X}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2 | \boldsymbol{X}_i) \leq \sigma^2$ almost surely.

Theorem 3. Assume Condition 1 holds. For any $z > 0$ and $v \geq \sigma$, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ in (15) with $\tau = v\sqrt{n/(d+z)}$ satisfies the bound

$$\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)\|_2 \leq c_1 v \sqrt{\frac{d+z}{n}}$$

with probability at least $1 - 2e^{-z}$ provided $n \geq c_2(d+z)$, where $c_1, c_2 > 0$ are constants depending only on A_0.

Theorem 3 establishes a sub-Gaussian concentration bound for $\widehat{\boldsymbol{\theta}}_\tau$ under the optimal sampling size scaling, which improves that in Theorem 2.1 in Zhou et al. (2018). To achieve a sub-Gaussian performance under the finite variance condition, the key observation is that the robustification parameter τ should adapt to the sample size, dimension, variance of noise and confidence level for optimal trade-off between bias and robustness. Extending our data-driven proposal for mean estimation, we estimate $\boldsymbol{\theta}^*$ and calibrate τ simultaneously by solving the system of equations

$$\begin{cases} g_1(\boldsymbol{\theta}, \tau) := \sum_{i=1}^n \psi_\tau(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})\boldsymbol{Z}_i = \mathbf{0}, \\ g_2(\boldsymbol{\theta}, \tau) := (n-d)^{-1} \sum_{i=1}^n \min\{(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2, \tau^2\}/\tau^2 - n^{-1}(d+z) = 0, \end{cases} \quad \boldsymbol{\theta} \in \mathbb{R}^{d+1}, \tau > 0. \tag{21}$$

With initial values $\boldsymbol{\theta}^{(0)} := \widehat{\boldsymbol{\theta}}_{\text{ols}} = (\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\intercal)^{-1} \sum_{i=1}^n Y_i \boldsymbol{Z}_i$ and $\tau^{(0)} = \widehat{\sigma}_n \sqrt{n/(d+z)}$ where $\widehat{\sigma}_n^2 = (n-d)^{-1} \sum_{i=1}^n (Y_i - \boldsymbol{Z}_i^\intercal \widehat{\boldsymbol{\theta}}_{\text{ols}})^2$, for $k \geq 1$, compute $\tau^{(k)}$ as the solution to $g_2(\boldsymbol{\theta}^{(k-1)}, \tau^{(k)}) = 0$ and then compute $\boldsymbol{\theta}^{(k)}$ as the solution to $g_1(\boldsymbol{\theta}^{(k)}, \tau^{(k)}) = 0$. Repeat the above steps until convergence or until the maximum number of iterations is reached. Denote by $(\widehat{\boldsymbol{\theta}}, \widehat{\tau})$ the final solution, then set $\widehat{\boldsymbol{\theta}}^{\text{I}} := \widehat{\boldsymbol{\theta}}_{\widehat{\tau}}$ as our one-step estimator.

The main advantage of the proposed adaptive Huber regression over the classical one with $\tau = 1.345\sigma$ is that the estimation bias with respect to intercept is alleviated. When τ is of the order $\sigma\sqrt{(n/d)}$, from (19) in Proposition 5 we see that the order of bias $|\alpha_\tau|$ is roughly $\sigma\sqrt{(d/n)}$. Examining the proof of Proposition 5, we find that the decrease in bias is linear in $1/\tau$ when the second moment is finite, and is quadratic in $1/\tau$ if the third moment is finite. We contrast this decrease with the stochastic term that dominates the error bound, which is given by the quadratic form of the score function evaluated at $\boldsymbol{\theta}^*$. The order of this term is roughly $\sigma\sqrt{(d/n)} + \tau d/n$. Balancing out the two terms yields the proposed choice of τ. However, by letting τ grow with sample size, the bias is reduced at the expense of losing some robustness of the coefficients estimators. When τ scales as a constant, such as $c\sigma$, the corresponding Huber loss is Lipschitz with bounded score function, and since $\boldsymbol{\beta}^*_\tau = \boldsymbol{\beta}^*$ for any $\tau > 0$, there will be no sacrifice in bias. The tuning constant c is typically chosen to ensure a given level of *asymptotic efficiency*. Asymptotic properties of general robust M-estimators have been well studied in the literature; see Avella-Medina and Ronchetti (2015) for a recent selective overview. The next result establishes the deviations of the Huber estimator with a fixed τ from a non-asymptotic viewpoint, representing a useful complement to Theorem 3.

Theorem 4. Suppose Condition 1 and the assumptions in Proposition 5 hold. Assume further that

$$\rho_\tau := \mathbb{P}(|\varepsilon - \alpha_\tau| \le \tau/2) > 0. \tag{22}$$

Then for any $z > 0$, the estimator $\widehat{\boldsymbol{\theta}}_\tau$ in (15) satisfies

$$\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*_\tau)\|_2 \lesssim \rho_\tau^{-1} A_0 \left(\sigma\sqrt{\frac{d+z}{n}} + \tau\frac{d+z}{n}\right) \tag{23}$$

with probability at least $1 - 2e^{-z}$ provided $n \ge c_3(d+z)$, where $c_3 = c_3(\rho_\tau, A_0) > 0$.

3.1.2 Two-step method

Motivated by our bias-robustness analysis and the results of finite sample investigation, we further introduce a two-step procedure that estimates the regression coefficients and intercept successively.

In the first stage, we compute the Huber estimator $\widehat{\boldsymbol{\theta}}_\tau = (\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}^{*\intercal}_\tau)^\intercal$ by solving the optimization problem in (15) with $\tau = c\sigma$ for some constant $c > 1$. We take $c = 1.345$ so that its efficiency at the normal model is 95%. For σ, it can be estimated simultaneously

with $\boldsymbol{\theta}^*$ by solving a system of equations as in Huber's proposal 2 (Huber, 1981), or we can fix σ at an initial robust estimate and then optimize over $\boldsymbol{\theta}$ (Hampel et al., 1986). We follow the former route and consider an iterative simultaneous estimation procedure, starting at iteration 0 with an initial estimate $\boldsymbol{\theta}^{(0)}$. At iteration $k = 0, 1, 2, \ldots$ it applies a simple procedure to obtain $\widehat{\sigma}^{(k)}$, which is then used to update $\boldsymbol{\theta}^{(k)}$, producing $\boldsymbol{\theta}^{(k+1)}$. The procedure involves two steps.

Step 1: Scale estimation. Using the current estimate $\boldsymbol{\theta}^{(k)}$, we compute the vector of residuals $\boldsymbol{r}^{(k)} = (r_1^{(k)}, \ldots, r_n^{(k)})^\intercal$ and the robustification parameter $\tau^{(k)} = 1.345\widehat{\sigma}^{(k)}$, where $\widehat{\sigma}^{(k)}$ denotes the median absolute deviation (MAD) estimator $\text{median}\{|r_i^{(k)} - \text{median}(r_i^{(k)})|\}/\Phi^{-1}(3/4)$.

Step 2: Weighted least squares. Compute the $n \times n$ diagonal matrix $\mathbf{W}^{(k)} = \text{diag}((1 + w_1^{(k)})^{-1}, \ldots, (1 + w_n^{(k)})^{-1})$, where $w_i^{(k)} = |r_i^{(k)}|/\tau^{(k)} - 1$ if $|r_i^{(k)}| > \tau^{(k)}$ and $w_i^{(k)} = 0$ if $|r_i^{(k)}| \leq \tau^{(k)}$. Then we update $\boldsymbol{\theta}^{(k)}$ to produce $\boldsymbol{\theta}^{(k+1)}$ via weighted least squares, that is,

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\text{argmin}} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})^2}{1 + w_i^{(k)}} = (\mathbf{Z}^\intercal \mathbf{W}^{(k)} \mathbf{Z})^{-1} \mathbf{Z}^\intercal \boldsymbol{Y},$$

where $\mathbf{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\intercal \in \mathbb{R}^{n \times (d+1)}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$.

Starting with $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\text{ols}}$, we repeat the above two steps for $s = 0, 1, 2, \ldots$ until convergence. We use $\widehat{\boldsymbol{\beta}}^{\text{II}} \in \mathbb{R}^d$ to denote the vector of coefficients estimates extracted from the final solution produced by the above procedure.

In the second stage, observe that $\beta_0^* = \mathbb{E}(\delta_i)$, where $\delta_i = Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* = \beta_0^* + \varepsilon_i$ are the residuals. To estimate β_0^*, defining fitted residuals $\widehat{\delta}_i = Y_i - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}}^{\text{II}}$, we solve the system of equations

$$\begin{cases} f_1(\beta_0, \tau) := n^{-1} \sum_{i=1}^n \min\{(\widehat{\delta}_i - \beta_0)^2, \tau^2\}/\tau^2 - n^{-1}\log(n) = 0, \\ f_2(\beta_0, \tau) := \sum_{i=1}^n \psi_\tau(\widehat{\delta}_i - \beta_0) = 0, \end{cases} \tag{24}$$

in the same way as for solving (13) to obtain $\widehat{\beta}_0^{\text{II}}$. Stack $\widehat{\beta}_0^{\text{II}}$ on top of $\widehat{\boldsymbol{\beta}}^{\text{II}}$ we obtain the two-step estimator $\widehat{\boldsymbol{\theta}}^{\text{II}} \in \mathbb{R}^{d+1}$ of $\boldsymbol{\theta}^*$.

Both the one-step and two-step methods are computationally efficient. For the former, letting τ diverge with sample size reduces the estimation bias in intercept at the cost of losing some robustness for estimating coefficients; the latter achieves high degree of robustness for estimating both the intercept and regression coefficients, and therefore takes the biggest advantage when the error distributions are heavy-tailed and skewed. While at the normal

model, since $\max_{1\le i\le n}|\varepsilon_i| \sim \sigma\sqrt{2\log(2n)}$ and the order of τ is roughly $\sigma\sqrt{(n/d)}$, the adaptive Huber estimator is almost identical to the least squares estimator. Numerical results in Section 4 provide strong support for the tail-adaptivity of the proposed data-driven Huber regression.

3.2 Adaptive Huber regression in high dimensions

We now move to the high dimensional setting where $d \gg n$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\intercal \in \mathbb{R}^d$ is sparse with $\|\boldsymbol{\beta}^*\|_0 := \sum_{j=1}^d I(\beta_j^* \neq 0) = s \ll n$. Since the invention of the Lasso by Tibshirani (1996), a verity of variable selection methods have been developed for finding a small group of covariates that are associated with the response from a large pool. We refer to Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015) for comprehensive reviews along this line.

Given observations $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, the Lasso is the solution to

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) \in \underset{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^n (Y_i - \beta_0 - \boldsymbol{X}_i^\intercal\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\},$$

where $\lambda > 0$ is a regularization parameter. Thinking of the noise variable as being Gaussian, this can be interpreted as a penalized maximum likelihood estimate, in which the ℓ_1-norm penalizes the fitted coefficients to induce sparsity. However, least squares fitting is sensitive to the tails of error distributions, particularly for ultra-high dimensional covariates as the maximum spurious correlation between the covariates and the realized noise can be large, and therefore is not an ideal choice in the presence of heavy-tailed noise.

Recently, Fan, Li and Wang (2017) modified Huber's procedure (Huber, 1973) to obtain an ℓ_1-regularized robust estimator which fulfills desirable concentration bounds under only finite variance condition on the regression errors. According to the discussions in Section 3.1, the intercept, albeit being often ignored in the literature, plays an important role in the study of robust methods. To take into account the effect of intercept, we consider the regularized Huber estimator of the form

$$\widehat{\boldsymbol{\theta}}_{\text{H}}(\tau,\lambda) \in \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta}^\intercal)^\intercal\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \mathcal{L}_\tau(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{25}$$

where $\mathcal{L}_\tau(\boldsymbol{\theta}) := n^{-1}\sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \ell_\tau(Y_i - \beta_0 - \boldsymbol{X}_i^\intercal\boldsymbol{\beta})$, τ and λ denote the robustification and regularization parameters, respectively.

Provided finite variance of the distribution of ε_i, Theorem 5 below reveals that the ℓ_1-regularized Huber regression with properly tuned (τ, λ) gives rise to statistically consistent estimators with ℓ_1- and ℓ_2-errors scaling as $s\sqrt{\log(d)/n}$ and $\sqrt{s\log(d)/n}$, respectively, under the sample size scaling $n \gtrsim s\log(d)$. These rates are exactly the minimax rates enjoyed by the standard Lasso with sub-Gaussian errors.

Theorem 5. Assume Condition 1 holds and denote by $\underline{\lambda}_{\mathbf{S}}$ the minimal eigenvalue of $\mathbf{S}$. Assume further that the unknown $\boldsymbol{\beta}^*$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0$. Let $\sigma_{jj} = \mathbb{E}(X_j^2)$ for $j = 1, \ldots, d$. Then the estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau, \lambda)$ given in (25) with $\tau = \sigma\sqrt{n/\log(d)}$ and λ scaling with $A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2} \sigma\sqrt{\log(d)/n}$ satisfies

$$\|\widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau, \lambda) - \boldsymbol{\theta}^*\|_2 \lesssim \frac{\lambda\sqrt{s}}{\underline{\lambda}_{\mathbf{S}}} \quad \text{and} \quad \|\widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau, \lambda) - \boldsymbol{\theta}^*\|_1 \lesssim \frac{\lambda s}{\underline{\lambda}_{\mathbf{S}}} \tag{26}$$

with probability at least $1 - 5d^{-1}$ as long as $n \ge c_1 s\log(d)$, where $c_1 > 0$ is a constant depending only on $(A_0, \max_{1\le j\le d}\sigma_{jj}, \underline{\lambda}_{\mathbf{S}})$.

Remark 3. The main purpose of using Huber loss for data fitting is to gain robustness against outliers from either contamination models (Huber, 1973) or heavy-tailed models considered in this paper. For other purposes, different loss functions have been proposed to replace the quadratic loss, such as the nonconvex Tukey and Cauchy losses, the quantile loss and the asymmetric quadratic loss, among others. We refer to Loh (2017), Alquier, Cottet and Lecué (2017) and Fan et al. (2018) for the most recent studies on regularized M-estimators with general loss functions.

In practice, it is computationally demanding to choose the optimal values of tuning parameters τ and λ by a two-dimensional grid search using cross-validation. Combining the data-driven method in Section 3.1 and the refitted cross-validation (RCV) technique (Fan, Guo and Hao, 2012), we consider the following procedure that estimates $\boldsymbol{\theta}^*$ and tunes τ simultaneously. Given a random sample of size n, we first randomly split it to two subsamples $\{(Y_{1i}, \boldsymbol{X}_{1i})\}_{i=1}^{n_1}$ and $\{(Y_{2i}, \boldsymbol{X}_{2i})\}_{i=1}^{n_2}$, where $n_1 = n_2 = n/2$ if n is even or $n_1 = (n+1)/2$, $n_2 = (n-1)/2$ otherwise. To begin with, we take cross-validated Lasso estimators computed separately from the two subsamples as initial values. At the k-th iteration ($k \ge 1$), using the estimates $\widehat{\boldsymbol{\theta}}_1^{(k-1)}$ and $\widehat{\boldsymbol{\theta}}_2^{(k-1)}$ from the last iteration, we compute $\tau^{(k)}$ which is the solution

to

$$\frac{1}{2\{n_1-\widehat{s}_2^{(k-1)}\}}\sum_{i=1}^{n_1}\frac{\min\{(Y_{1i}-\boldsymbol{Z}_{1i}^\intercal\widehat{\boldsymbol{\theta}}_2^{(k-1)})^2,\tau^2\}}{\tau^2} + \frac{1}{2\{n_2-\widehat{s}_1^{(k-1)}\}}\sum_{i=1}^{n_2}\frac{\min\{(Y_{2i}-\boldsymbol{Z}_{2i}^\intercal\widehat{\boldsymbol{\theta}}_1^{(k-1)})^2,\tau^2\}}{\tau^2}=\frac{\log(nd)}{n}, \tag{27}$$

where $\widehat{s}_1^{(k-1)}=\|\widehat{\boldsymbol{\beta}}_1^{(k-1)}\|_0$ and $\widehat{s}_2^{(k-1)}=\|\widehat{\boldsymbol{\beta}}_2^{(k-1)}\|_0$. Next, take $\tau=\tau^{(k)}$ and compute $\widehat{\boldsymbol{\theta}}_1^{(k)}$ and $\widehat{\boldsymbol{\theta}}_2^{(k)}$ as the solutions to the optimization problems

$$\min_{\boldsymbol{\theta}}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}\ell_\tau(Y_{1i}-\boldsymbol{Z}_{1i}^\intercal\boldsymbol{\theta})+\lambda_1\|\boldsymbol{\beta}\|_1\right\}\quad\text{and}\quad\min_{\boldsymbol{\theta}}\left\{\frac{1}{n_2}\sum_{i=1}^{n_2}\ell_\tau(Y_{2i}-\boldsymbol{Z}_{2i}^\intercal\boldsymbol{\theta})+\lambda_2\|\boldsymbol{\beta}\|_1\right\}, \tag{28}$$

respectively, where $\lambda_1,\lambda_2>0$ are chosen via cross-validation. Repeat the above two steps until convergence or until the maximum number of iterations is reached. The resulting τ is denoted by $\widehat{\tau}_{\mathrm{rcv}}$. The final estimator is then defined as the solution to (25) with $\tau=\widehat{\tau}_{\mathrm{rcv}}$ and λ calibrated via cross-validation.

To implement the data-driven Huber regression in high dimensions, again, starting with some initial guess we iteratively solve (27) and (28). For the convex optimization problems in (28), the minimizer satisfies the Karush–Kuhn–Tucker conditions, and therefore can be found by solving the following system of nonsmooth equations:

$$\begin{cases} -n^{-1}\sum_i\psi_\tau(Y_i-\boldsymbol{Z}_i^\intercal\widehat{\boldsymbol{\theta}})=0,\\ -n^{-1}\sum_i\psi_\tau(Y_i-\boldsymbol{Z}_i^\intercal\widehat{\boldsymbol{\theta}})X_{ij}+\lambda\widehat{s}_j=0,\quad j=1,\ldots,d\\ \widehat{\beta}_j-S(\widehat{\beta}_j+\widehat{s}_j)=0,\quad j=1,\ldots,d\end{cases} \tag{29}$$

where $\widehat{\boldsymbol{\theta}}=(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}}^\intercal)^\intercal\in\mathbb{R}^{d+1}$ with $\widehat{\boldsymbol{\beta}}=(\widehat{\beta}_1,\ldots,\widehat{\beta}_d)^\intercal$, $\widehat{s}_j\in\partial|\widehat{\beta}_j|$ and $S(z)=\mathrm{sign}(z)(|z|-1)_+$ is the soft-thresholding operator. Instead of directly applying the Semismooth Newton Algorithm (SNA) to the entire system of equations, we adapt the Semismooth Newton Coordinate Descent (SNCD) algorithm proposed by Yi and Huang (2017), which combines SNA with cyclic coordinate descent in solving (29). More specifically, in SNCD we divide (29) into two parts in order to avoid cumbersome matrix operations as in solving the entire system. In a cyclic fashion we update the intercept only using the first equation and update the coefficients with its subgradients using the last two equations, therefore reducing the computational cost from $O(nd^2)$ to $O(nd)$ at each iteration. The computational scalability and efficiency are gained especially when d is very large. After obtaining a solution path of (28), we employ the cross-validation method to calibrate λ_1 and λ_2 and obtain the associated $\widehat{\boldsymbol{\theta}}_1^{(k)}$ and $\widehat{\boldsymbol{\theta}}_2^{(k)}$.

Remark 4. The above regularized data-adaptive Huber regression method is a direct extension of the one-step method proposed in Section 3.1 to high dimensions. Also, note that Proposition 5 also holds in high dimensions as long as the population Gram matrix $\mathbf{S}$ is positive definite. Therefore, to further reduce the estimation bias of intercept, we suggest a two-step procedure that estimates the regression coefficients using standard regularized Huber regression and then estimates the intercept by applying the adaptive-Huber method to fitted residuals as in (24). Section 4.1.3 provides numerical studies of both the one- and two-step regularized adaptive Huber estimators.

4 Empirical analysis

4.1 Simulated examples

In this section, we examine numerically the finite sample performance of the proposed data-adaptive Huber (DA-Huber) methods for mean estimation and linear regressions. We consider the following four distribution settings to investigate the robustness and efficiency of the proposed method in a wide variety of scenarios.

(1) Normal distribution $\mathcal{N}(0, \sigma^2)$ with mean zero and variance $\sigma^2 > 0$;

(2) Skewed generalized t distribution (Theodossiou, 1998) $\mathsf{sgt}(\mu, \sigma^2, \lambda, p, q)$, where we set mean parameter $\mu = 0$, variance parameter $\sigma^2 = q/(q-2)$ with $q > 2$, shape parameter $p = 2$ and skewness parameter $\lambda = 0.75$;

(3) Lognormal distribution $\mathsf{LN}(\mu, \sigma^2)$ with the log location parameter $\mu = 0$ and log shape parameter $\sigma > 0$;

(4) Pareto distribution $\mathsf{Par}(x_m, \alpha)$ with scale parameter $x_m = 1$ and shape parameter $\alpha > 0$.

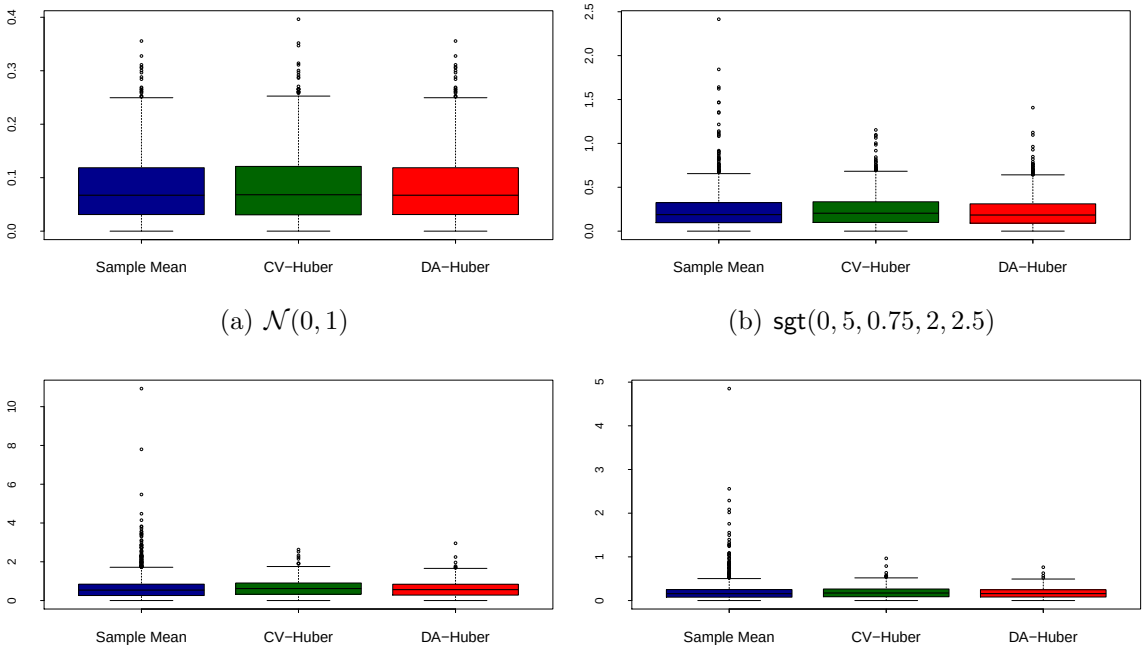
Except the normal distribution, all the other three are skewed and heavy-tailed.

4.1.1 Mean estimation

For each setting, we generate an independent sample of size $n = 100$ and compute three mean estimators: the sample mean, the Huber estimator with τ chosen via five-fold cross-validation (CV-Huber), and the proposed DA-Huber mean estimator. Figure 1 displays the

boxplots of the estimation error based on 2000 simulations, and Figure 2 illustrates the α-quantile of the estimation error with α ranging from 0.5 to 1. The DA-Huber estimator and sample mean perform almost identically for the normal model. For the heavy-tailed skewed distributions, the deviation of the sample mean from the population mean grows rapidly with the confidence level, which is in striking contrast to the CV- and DA-Huber estimators.

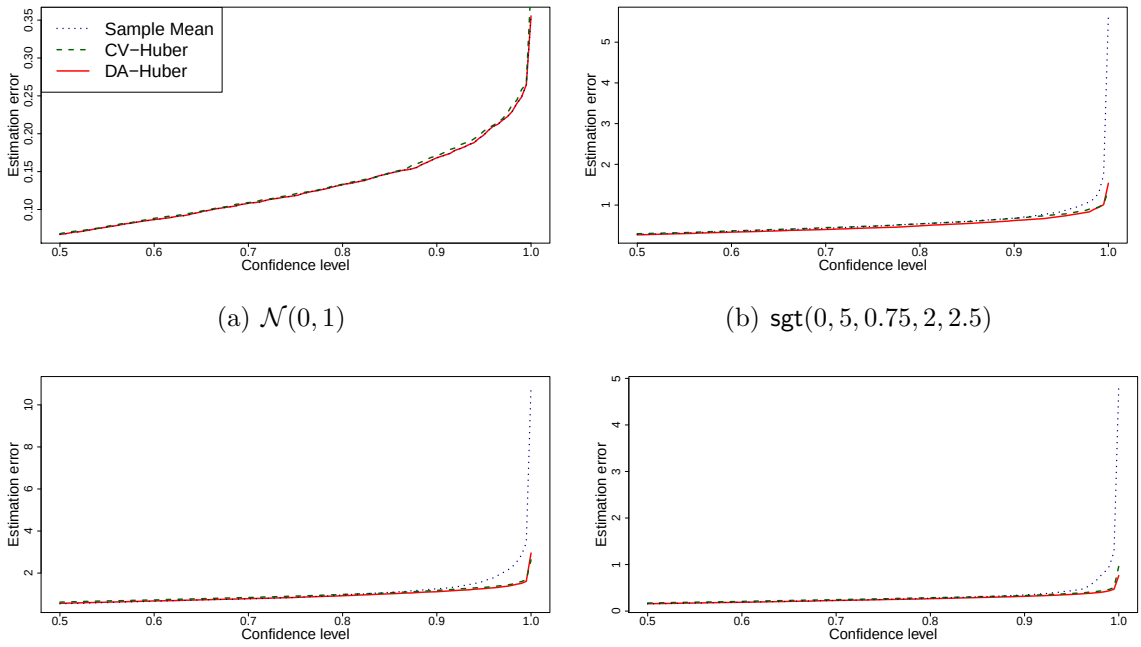
In Figure 3, we examine the 99%-quantile of the estimation error versus a distribution parameter that measures the tail behavior. Namely, for normal distributions we let σ vary between 1 and 4; for skewed generalized t distributions, we increase the shape parameter q from 2.5 to 4; for lognormal and Pareto distributions, the shape parameters σ and α vary from 0.25 to 2 and 1.5 to 3, respectively. The Huber-type estimators show significant improvement in deviations from the population mean as the tails become heavier. In summary, the most attractive feature of our method is its adaptivity: (i) it is as efficient as the sample mean for the normal model and is more robust for asymmetric and/or heavy-tailed data; (ii) it performs as good as the cross-validation but with much less computational cost. The latter is particularly important for large-scale inference with a myriad of parameters to be estimated simultaneously.



(a) $\mathcal{N}(0,1)$ (b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

Figure 1: Estimation errors for the sample mean, CV-Huber and DA-Huber estimators under different settings based on 2000 simulations.



(a) $\mathcal{N}(0,1)$ (b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

Figure 2: Estimation error versus confidence level for the sample mean, CV-Huber and DA-Huber estimators based on 2000 simulations.

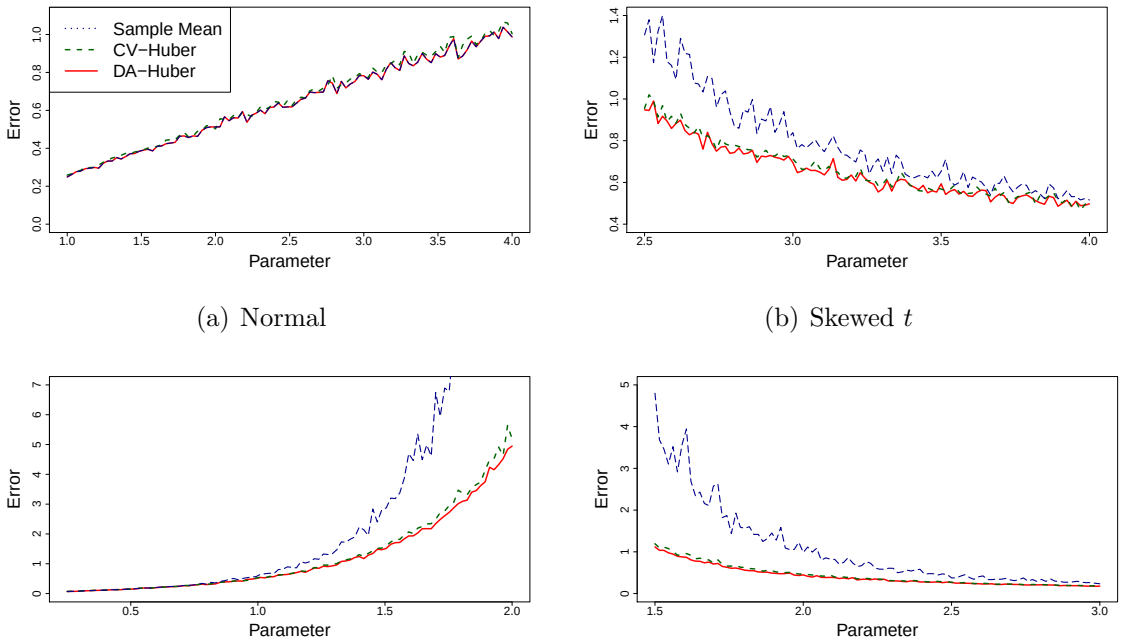
4.1.2 Linear regression

We generate data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from linear model (14) with $n = 500$ and $d = 5$. The intercept and vector of regression coefficients are taken to be $\beta_0 = 5$ and $\boldsymbol{\beta}^* = (1, -1, 1, -1, 1)^\intercal$, respectively. The covariates $\boldsymbol{X}_i$ are i.i.d. random vectors that consist of independent coordinates from a uniform distribution $\mathsf{Unif}(-1.5, 1.5)$.

We compare the DA-Huber regression estimator with the ordinary least squares (OLS) estimator and classical robust M-estimators with Huber loss $\ell_\tau(\cdot)$ as in (1) and Tukey's biweight loss

$$\ell_\tau^{\mathrm{T}}(x) = \begin{cases} 1 - (1 - x^2/\tau^2)^3 & \text{if } |x| \le \tau, \\ 1 & \text{if } |x| > \tau. \end{cases}$$

The tuning parameter τ in $\ell_\tau^{\mathrm{T}}(\cdot)$ and $\ell_\tau(\cdot)$ is taken to be 4.685 and 1.345, respectively, according to the asymptotic 95% efficiency rule. We carry out 1000 Monte Carlo simulations to: (1) evaluate the overall performance of the DA-Huber methods comparing with three competing methods, labeled as OLS, Tukey and Huber; see Figures 4 and 5, and (2) demonstrate the robustness of different methods with varying degrees of heavy-tailedness and skewness; see



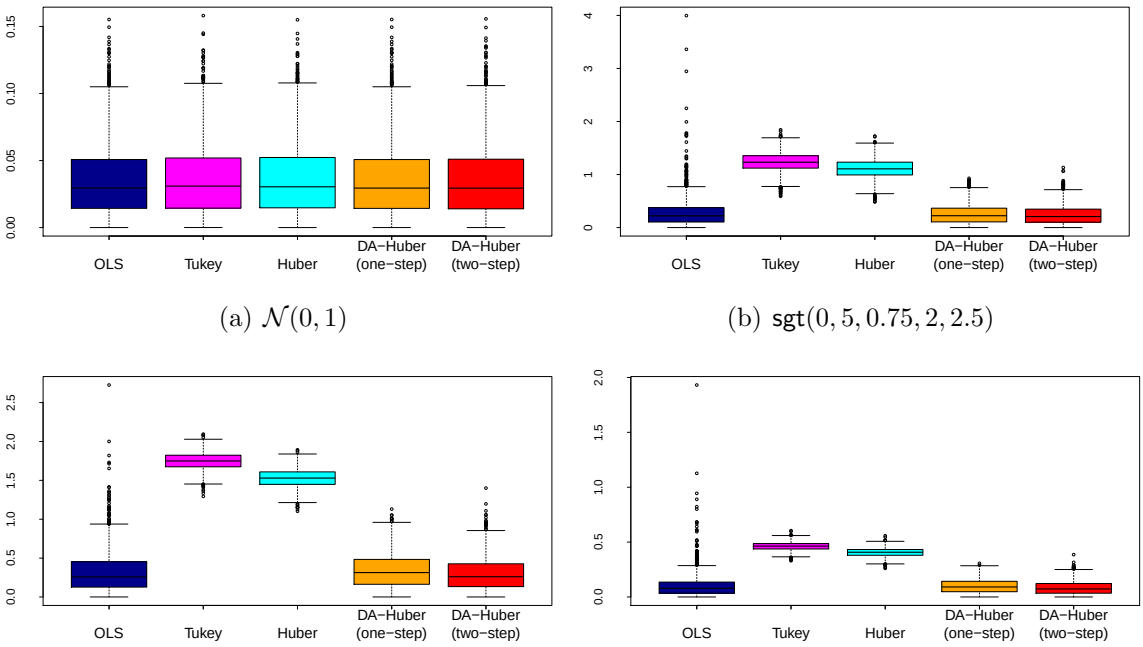
(a) Normal (b) Skewed t

(c) Lognormal (d) Pareto

Figure 3: Empirical 99%-quantile of the estimation error versus distribution parameter (that measures tails) for the sample mean, CV-Huber and DA-Huber estimators.

Figures 6 and 7.

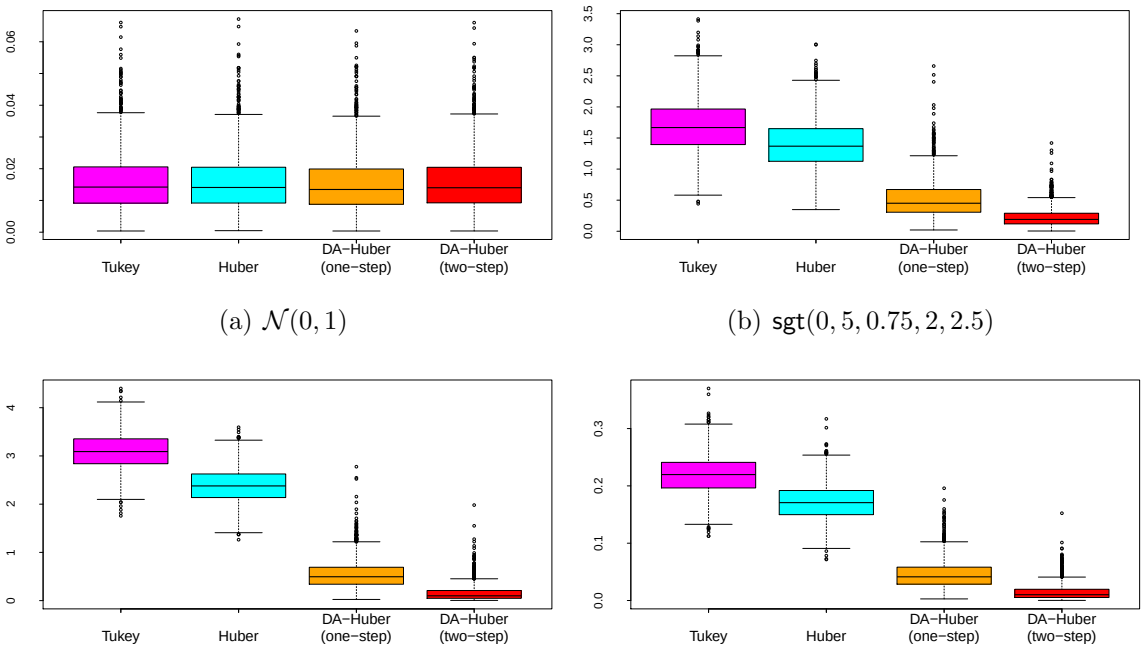
Figures 4 and 5 display the boxplots of the estimation error of intercept $|\widehat{\beta}_0 - \beta_0^*|$ and the total ℓ_2-error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$, respectively, for a fixed distribution parameter as did in Section 4.1.1. Both the one-step and two-step DA-Huber estimators outperform the other methods across all examples. For estimating the intercept, the DA-Huber rectifies the nonnegligible bias in classical robust M-estimators, as predicted by theory. In the normal case, the DA-Huber estimator performs almost identically with the OLS and is therefore highly efficient. The ℓ_2-error of OLS tends to spread out (due to outliers) and thus is not reported in Figure 5. Figures 6 and 7 illustrate, respectively, the average estimation error of intercept and the total ℓ_2-error versus the distribution parameter that controls the shape of tails. In the normal case, the one-step DA-Huber and OLS slightly outperform the others; with heavy-tailed and skewed errors, the DA-Huber methods enjoy notable advantage and the two-step approach is most desirable since it strikes the perfect balance between bias and robustness. Overall, the numerical results indicate that the proposed methods have substantial advantages in the presence of asymmetric and heavy-tailed errors, while maintaining high efficiency for the normal model.



(a) $\mathcal{N}(0,1)$ (b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

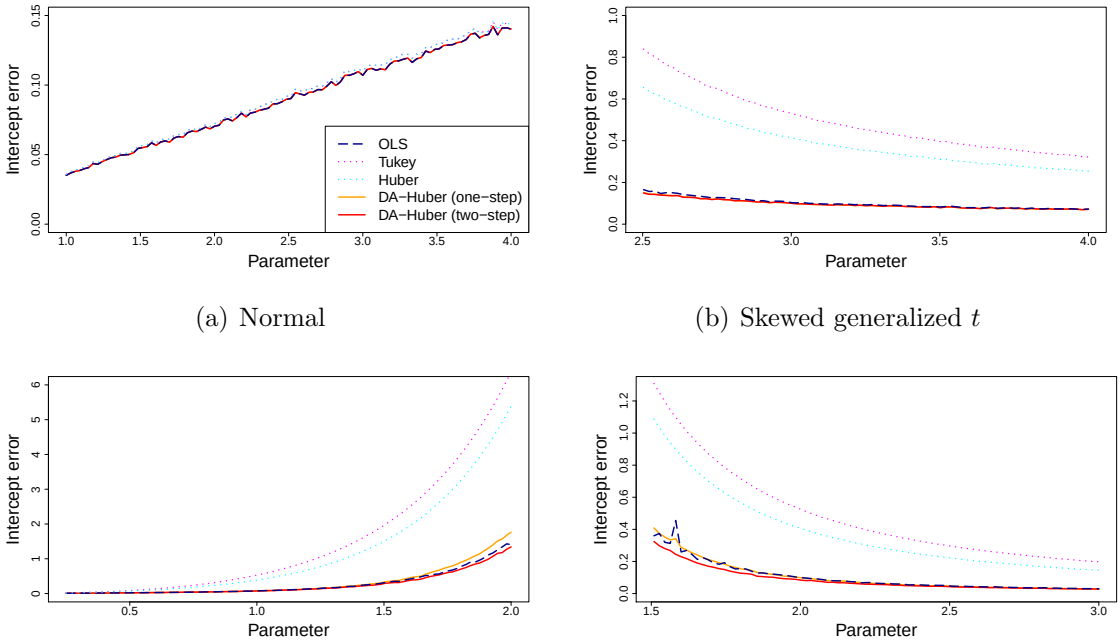
Figure 4: Estimation errors of intercept under different settings.



(a) $\mathcal{N}(0,1)$ (b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

Figure 5: Total ℓ_2-errors under different settings.



(a) Normal

(b) Skewed generalized t

(c) Lognormal

(d) Pareto

Figure 6: Average estimation error of intercept versus distribution parameters controlling tails for the OLS estimator, standard Tukey's and Huber's estimators, and data-adaptive Huber estimators (one-step and two-step).

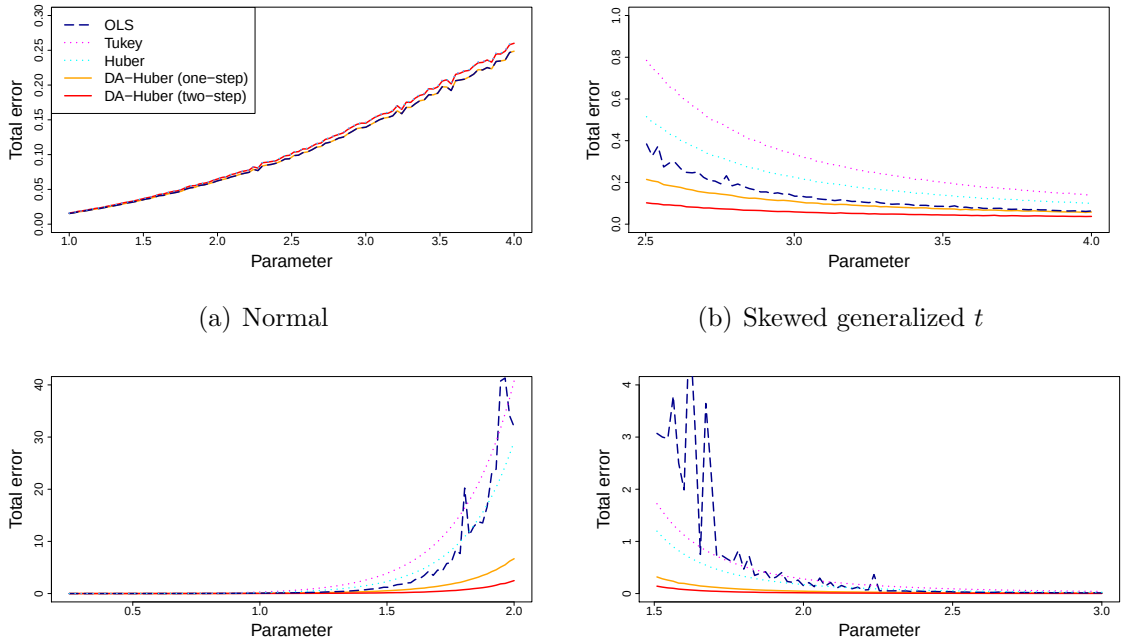
4.1.3 Sparse linear regression

Now we consider the sparse linear regression model

$$Y_i = \beta_0^* + \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0 \ll n$ and $d \gg n$. In the examples below, we take $n = 250$, $d = 1000$ and $s = 20$. We set $\beta_0^* = 3$ and $\boldsymbol{\beta}^* = (3, \ldots, 3, 0, \ldots, 0)^\intercal$, where the first $s = 20$ elements of $\boldsymbol{\beta}^*$ all equal 3 and the rest are zero. As before, the covariates $\boldsymbol{X}_i$ are i.i.d. random vectors that consist of independent coordinates from $\mathsf{Unif}(-1.5, 1.5)$, and the regression errors are generated from one of the four distributions: normal, skewed generalized t, lognormal and Pareto.

To implement the iterative procedure proposed in Section 3.2, at the k-th iteration, we use the five-fold cross-validation to choose regularization parameters $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ in the optimization programs in (28), producing $\widehat{\boldsymbol{\theta}}_1^{(k)}$ and $\widehat{\boldsymbol{\theta}}_2^{(k)}$. We evaluate the proposed regularized DA-Huber estimators by the following measurements.



(a) Normal (b) Skewed generalized t

(c) Lognormal (d) Pareto

Figure 7: Average ℓ_2-errors versus distribution parameters controlling tails for the OLS estimator, standard Tukey's and Huber's estimators, and data-adaptive Huber estimators (one-step and two-step).

(1) RG, the relative gain of the DA-Huber estimator with respect to the Lasso in terms of ℓ_1- and ℓ_2-errors:

$$\mathrm{RG}_1 = \frac{\|\widehat{\boldsymbol{\theta}}_{\mathrm{H}} - \boldsymbol{\theta}\|_1}{\|\widehat{\boldsymbol{\theta}}_{\mathrm{lasso}} - \boldsymbol{\theta}\|_1} \quad \text{and} \quad \mathrm{RG}_2 = \frac{\|\widehat{\boldsymbol{\theta}}_{\mathrm{H}} - \boldsymbol{\theta}\|_2}{\|\widehat{\boldsymbol{\theta}}_{\mathrm{lasso}} - \boldsymbol{\theta}\|_2};$$

(2) FP, the number of false positives (the number of noise covariates that are selected);

(3) FN, the number of false negatives (the number of signal covariates that are missing).

Table 1 summaries the relative gains of the DA-Huber estimators under ℓ_1- and ℓ_2-errors and the numbers of false positive and false negative discoveries. Across all the four models, both one- and two-step DA-Huber estimators outperform the Lasso with smaller ℓ_1-errors and fewer false positive discoveries, therefore are less greedy in model selection. For the normal model, the proposed robust methods and Lasso perform equally well; while in the presence of heavy-tailed skewed errors, the DA-Huber methods lead to remarkably better outputs in regard of both estimation and model selection. Similar phenomenon can also be

observed from Figure 8, which displays the empirical distributions of the ℓ_2-errors of the Lasso and DA-Huber estimators.

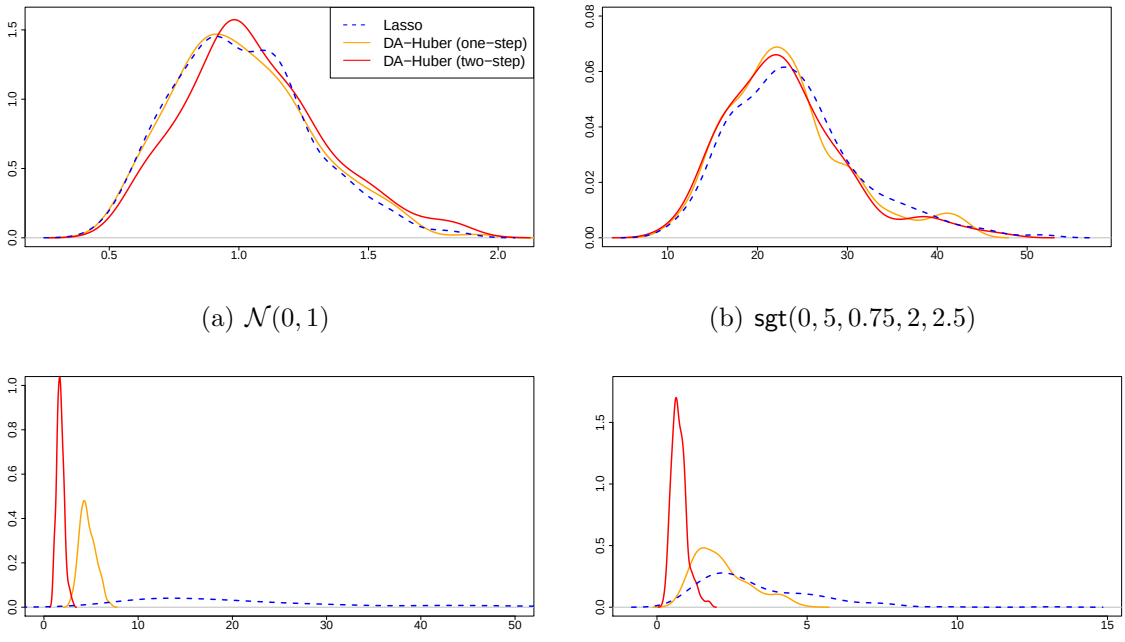
		Lasso	DA-Huber (one-step)	DA-Huber (two-step)
	RG_1	1	0.934	0.914
Normal	RG_2	1	1.003	1.027
$\mathcal{N}(0,1)$	FP	87.9	77.6	73.5
	FN	0	0	0
	RG_1	1	0.875	0.862
Skewed generalized t	RG_2	1	0.983	0.981
$\mathsf{sgt}(0,5,0.75,2,2.5)$	FP	86.1	63.1	60.7
	FN	0	0	0
	RG_1	1	0.347	0.227
Lognormal	RG_2	1	0.495	0.305
$\mathsf{LN}(0,1.5)$	FP	80.8	21.9	26.6
	FN	0.26	0	0
	RG_1	1	0.653	0.417
Pareto	RG_2	1	0.845	0.512
$\mathsf{Par}(1,2)$	FP	85.1	34.5	44.2
	FN	0	0	0

Table 1: RG, FP and FN of the Lasso and DA-Huber estimators under different models. The results are based on 200 simulations.

4.2 Real data examples

In this section, using three real data sets, we demonstrate the desirable performance of the proposed DA-Huber methods in terms of prediction accuracy.

Liu and Rubin (1995) reported a data collected from a clinical trial on endogenous creatinine clearance of 34 male patients where 28 samples are free from missing data. For the four recorded variables, it is known that the level of serum creatinine is closely related to the endogenous creatinine clearance with the body weight and age properly adjusted. Linear



(a) $\mathcal{N}(0,1)$ (b) $\mathsf{sgt}(0,5,0.75,2,2.5)$

(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

Figure 8: Distritbuions of the ℓ_2-errors of the Lasso and DA-Huber estimators.

model (14) is a natural preliminary fit to the data. In addition, we observe that the empirical kurtosis of the level of serum creatinine is 19.66, which hints potential heavy-tailedness in the data. The second example is the hedonic housing crime data (Harrison and Rubinfeld, 1978), which was originally used to study the association between housing market and local air quality. Interestingly, this data also provides some insights on how crime rates vary with respect to house-economics features, such as the proportion of residential land zoned for lots greater than $25,000$ square feet, the proportion of non-retail business within a town, proportion of owner units built prior to 1940, proportion of adults without high school education, median value of owner-occupied homes, average number of rooms in owner units, and distance to five employment centers in Boston region. This data set contains 506 locations and the empirical kurtosis of the crime rate is 39.75. The last data set is the well-known G-Econ data reported by Nordhaus et al. (2006), which was used to show the dependence of gross cell product (GCP) on geographical variables measured on a spatial scale of one degree. The original data contains $27,445$ terrestrial grid cells and 47 predictors, and varies abruptly across different latitude and longitude. For example, the sizes of grid cell may change substantially from the equator to the poles. Similar to Nordhaus et al. (2006), we focus on regions from 35 to 50 latitudes (parallel north) that contain a large number of

major economic centers, such as Tokyo, New York, Paris and London. Excluding cells with empty inputs, 808 observations remain for studying the relationship between the GCP (in USD) in 1990 and ten explanatory variables as discussed in Nordhaus et al. (2006), including distance to coast, distance to major navigable lakes, distance to major navigable rivers, distance to ice-free ocean, elevation, standard deviation of elevations, elevation from shuttle radar topography mission data, latitude, average precipitation, and average temperature. The empirical kurtosis of the GCP is 256.58, suggesting strong heavy-tailedness.

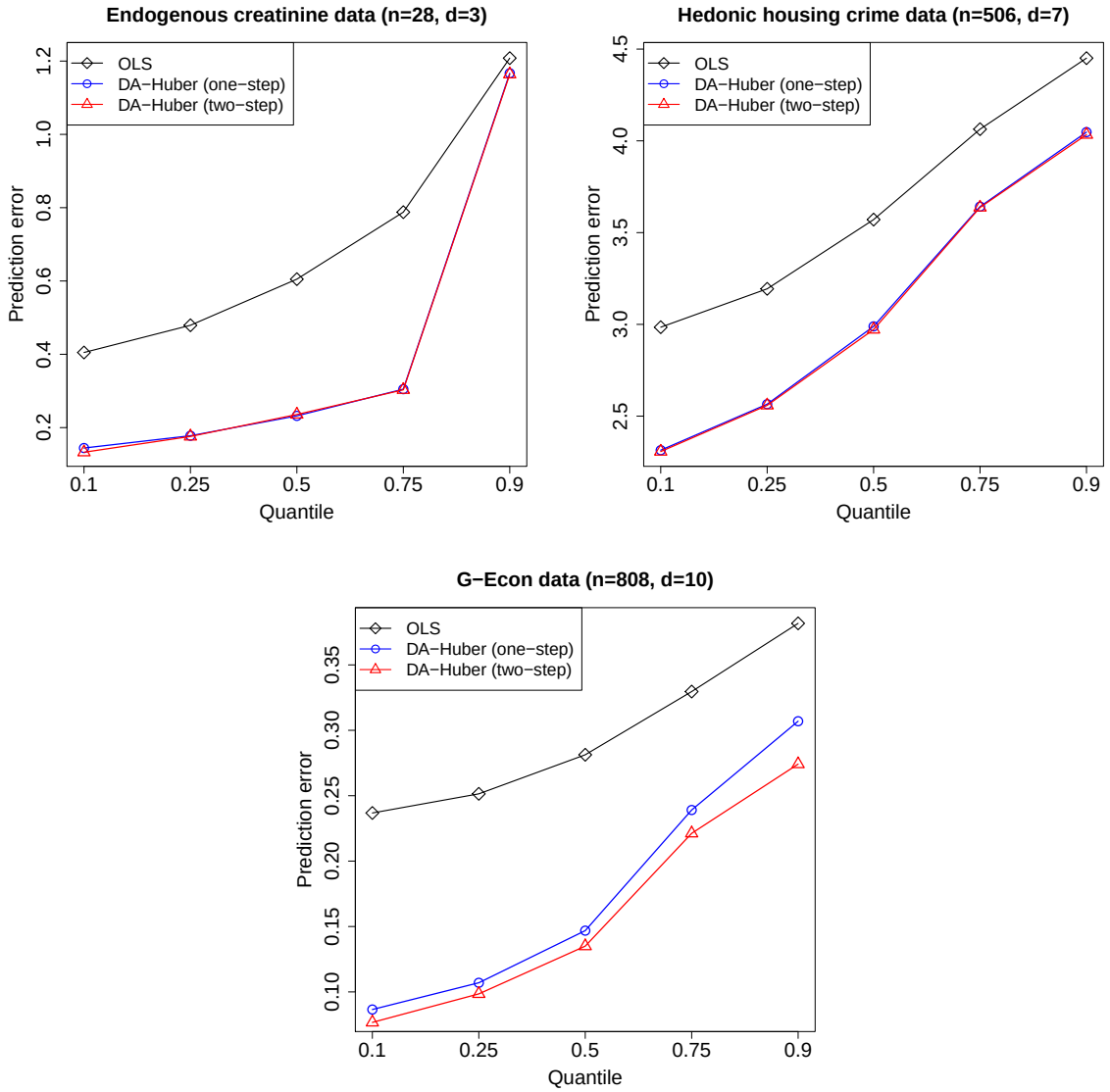


Figure 9: Comparison of the quantiles of mean absolute prediction errors for the OLS (black diamonds), one-step DA-Huber (blue circles), and two-step DA-Huber (red triangles). The results are based on 100 random splittings.

From the simulation studies in Section 4.1 we see that both the one-step and two-step DA-Huber estimators outperform the OLS in terms of estimation accuracy. For the real data, we focus on the prediction accuracy by investigating the mean absolute prediction errors. Specifically, upon splitting the data into $K = 7$ groups randomly, we predict the responses of one group using the regression coefficients estimated from the other $K - 1$ groups. Various quantile levels of the K mean absolute prediction errors were computed for different estimators. We repeat the random splitting 100 times. Figure 9 displays the empirical medians of α-quantiles of the mean absolute prediction errors for the three data sets, where α ranges from 0.1 to 0.9. The two data-adaptive Huber estimators substantially outperform the OLS with smaller prediction errors. When heavy-tailedness prevails and the intercept is nonnegligible, such as the GCP in G-Econ data, the two-step estimator displays the best performance. In general, the one- and two-step methods perform comparably well. For the endogenous creatinine data, the 0.9-quantiles of the mean absolute prediction errors of the three methods are comparable, which is possibly due to the small sample size ($n = 28$). To sum up, the data-adaptive Huber methods provide notably better predictions than the least squares for these three real data examples.

5 Summary

Balancing bias and robustness, the robustification parameter plays the central role in recent development on robust estimation and inference for heavy-tailed data. In this paper, we have proposed a new principle to choose the robustification parameter adaptively from data for a variety of fundamental statistical problems, including mean estimations, linear regression and the sparse regression in high dimensions. Inspired by the censored moment equation approach, the proposed principle is genuinely tuning-free and data-adaptive. It is conceptually different from the traditional practice on selecting the robustification parameter based on cross-validation, which is not only computationally demanding but also lacks of the underpinning mathematical guarantees. The proposed principle is guided by non-asymptotic deviation analysis and paves a unified pathway for choosing robustification parameter for robust estimation and inference in general, particularly for M-estimations. In light of numerical evidences from both synthetic and real data, our proposal outperforms those widely known robust estimation procedures in terms of estimation, variable selection, and prediction.

References

Alquier, P., Cottet, V. and Lecué, G. (2017). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *arXiv preprint arXiv:1702.01402.*

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

Avella-Medina, M., Battey, H. S., Fan, J. and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105, 271–284.

Avella-Medina, M. and Ronchetti, E. (2015). Robust statistics: A selective overview and new directions. *WIREs Computational Statistics,* 7, 372–393.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association,* 70, 428–434.

Brownlees, C., Joly, E. and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43, 2507–2536.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 48, 1148–1185.

Cont. R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1, 223–236.

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

Eklund, A., Nichols, T. and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7900–7905.

Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *The Annals of Statistics,* 46, 3481–3509.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society, Series B,* 74, 37–65.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society, Series B,* 79, 247–265.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 96, 1348–1360.

Hampel F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81–102.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton.

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

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics,* 35, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics,* 1, 799–821.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics,* 5, 1015–1053.

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306.*

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5, 19–39.

Loh, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics,* 45, 866–896.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16, 559–616.

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757.*

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

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21, 2308–2335.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics,* 46, 2871–2903.

Nordhaus, W., Azam, Q., Corderi, D., Hood, K., Victor, N. M., Mohammed, M., Miltner, A. and Weiss, J. (2006). The G-Econ database on gridded output: Methods and data. Yale University, New Haven. Available at https://gecon.yale.edu.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4, 16.

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

Su, W., Bogdan, M. and Candés, E. (2017). False discoveries occur early on the Lasso path. *The Annals of Statistics,* 45, 2133–2150.

Sun, Q., Zhou, W.-X. and Fan, J. (2017). Adaptive Huber regression. *arXiv preprint arXiv:1706.06991.*

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science,* 44, 1650–1661.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B,* 58, 267–288.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 210–268. Cambridge Univ. Press, Cambridge.

Yi, C. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics,* 26, 547–557.

Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics,* 46. 1904–1931.

Appendix

A Proofs of Results in Section 2

A.1 Preliminaries

We first introduce some useful notions of the distribution of a random variable. Let X be a non-degenerate real-valued random variable with finite variance. For $t \geq 0$, we define the tail probability of $|X|$, the second moments of truncated and censored versions of X by

$$G(t) = \mathbb{P}(|X| > t), \;\; P(t) = \mathbb{E}\{X^2 I(|X| \leq t)\} \;\; \text{and} \;\; Q(t) = \mathbb{E}\{\psi_t(X)\}^2, \tag{30}$$

respectively, where $\psi_t(x) = (|x| \wedge t)\,\mathrm{sign}(x)$ for $x \in \mathbb{R}$. Moreover, for $t > 0$, we define

$$p(t) = t^{-2}P(t) \;\; \text{and} \;\; q(t) = t^{-2}Q(t). \tag{31}$$

By definition, it is straightforward that $Q(t) = P(t) + t^2 G(t)$ and $q(t) = p(t) + G(t)$. The following result provides some useful connections among these functions. See (2.3) and (2.4) in Hahn, Kuelbs and Weiner (1990). We reproduce them here for the sake of readability.

Lemma 1. Let functions G, Q, p and q be given in (30) and (31).

(i) For any $t > 0$, we have

$$Q(t) = 2\int_0^t yG(y)\,dy, \quad q'(t) = -2t^{-1}p(t), \tag{32}$$

and

$$q(t) = \mathbb{P}(X \neq 0) - 2\int_0^t y^{-1}p(y)dy. \tag{33}$$

In addition, function $Q : [0, \infty) \to \mathbb{R}$ is non-decreasing with $\lim_{t\to\infty} Q(t) = \mathbb{E}(X^2)$.

(ii) Function $q : (0, \infty) \to \mathbb{R}$ is non-increasing and positive everywhere with $q(0+) := \lim_{s\downarrow 0} q(s) = \mathbb{P}(X \neq 0)$. Moreover,

$$q(s) = \mathbb{P}(X \neq 0) \text{ for all } 0 \leq s \leq \Delta := \inf\{y > 0 : G(y) < \mathbb{P}(X \neq 0)\}, \tag{34}$$

$q(s)$ decreases strictly and continuously on (Δ, ∞), and $\lim_{t\to\infty} q(t) = 0$.

Proof of Lemma 1. Notice $Q(t) = \mathbb{E}\{(|X| \wedge t)^2\}$ and it holds almost surely that

$$\begin{aligned}(|X| \wedge t)^2 &= 2\int_0^t I(|X| > t)y\,dy + 2\int_0^{|X|} I(|X| \leq t)y\,dy \\ &= 2\int_0^t I(|X| > t)y\,dy + 2\int_0^t I(|X| > y)I(|X| \leq t)y\,dy \\ &= 2\int_0^t I(|X| > y)y\,dy.\end{aligned}$$

Taking expectations on both sides implies $Q(t) = \mathbb{E}\{(|X| \wedge t)^2\} = 2\int_0^t \mathbb{P}(|X| > y)y\,dy = 2\int_0^t yG(y)dy$, as stated. Hence, $Q'(t) = 2tG(t)$. In (31), taking derivatives with respect to t on both sides gives $2tq(t) + t^2q'(t) = 2tG(t) = 2t\{q(t) - p(t)\}$. The second equation in (32) therefore follows. To prove (33), note that, for any $0 < s < t$, $q(t) = q(s) - 2\int_s^t p(y)y^{-1}dy$. On event $\{|X| > 0\}$, it holds almost surely that

$$0 < \frac{(|X| \wedge s)^2}{s^2} \leq 1, \quad \text{and} \quad \frac{(|X| \wedge s)^2}{s^2} \to 1 \quad \text{as } s \to 0.$$

By the dominated convergence theorem,

$$q(s) = \mathbb{E}\{s^{-2}(|X| \wedge s)^2\} = \mathbb{E}\{s^{-2}(|X| \wedge s)^2 I(|X| > 0)\} \to \mathbb{P}(|X| > 0) \quad \text{as } s \to 0.$$

Then, in $q(t) = q(s) - 2\int_s^t p(y)y^{-1}dy$ for all $0 < s < t$, letting s tend to zero yields (33). The monotonicity of Q follows directly from (32) and the limit of $Q(t)$ derives from the monotone convergence theorem. These complete the part (i) of Lemma 1.

We now show the remaining properties of function q in the part (ii). By the definition of Δ in (34), we have $\mathbb{P}(0 < |X| \le y) = 0$ and thus $p(y) = 0$ for all $0 < y < \Delta$. This, together with (33), implies $q(s) = \mathbb{P}(X \neq 0) > 0$ for all $0 \le s \le \Delta$. It is easy to see that $p(y) > 0$ for any $y > \Delta$, and therefore $q(\cdot)$ is strictly decreasing on (Δ, ∞). Finally, note that

$$0 < \frac{(|X| \wedge s)^2}{s^2} \le 1, \quad \text{and} \quad \frac{(|X| \wedge s)^2}{s^2} \to 0 \quad \text{as } s \to \infty$$

almost surely. The dominated convergence theorem leads to $\lim_{t\to\infty} q(t) = 0$. □

A.2 Proof of Proposition 1

Note that the truncated mean m_τ can be written as $m_\tau = \tau n^{-1} \sum_{i=1}^n \psi_1(X_i/\tau)$, where it can be easily verified that $-\log(1 - u + u^2) \le \psi_1(u) \le \log(1 + u + u^2)$ for all $u \in \mathbb{R}$. For any $y > 0$, it follows that

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^n \{\tau\psi_1(X_i/\tau) - \mu\} \ge y\right] &\le \exp\{-(y + n\mu)/\tau\}\mathbb{E}\left[\exp\left\{\sum_{i=1}^n \psi_1(X_i/\tau)\right\}\right] \\
&= \exp\{-(y + n\mu)/\tau\} \prod_{i=1}^n \mathbb{E}\exp\{\psi_1(X_i/\tau)\} \\
&\le \exp\{-(y + n\mu)/\tau\} \prod_{i=1}^n \mathbb{E}\exp\{\log(1 + X_i/\tau + X_i^2/\tau^2)\} \\
&= \exp\{-(y + n\mu)/\tau\} \prod_{i=1}^n \mathbb{E}(1 + X_i/\tau + X_i^2/\tau^2) \\
&\le \exp\{-(y + n\mu)/\tau\} \prod_{i=1}^n \exp\{\mu/\tau + \mathbb{E}(X_i^2)/\tau^2\} \\
&\le \exp(-y/\tau + nv^2/\tau^2) \\
&= \exp\left\{nv^2\left(\frac{1}{\tau} - \frac{y}{2nv^2}\right)^2 - \frac{y^2}{4nv^2}\right\}.
\end{aligned}$$

Similarly,

$$\mathbb{P}\left[\sum_{i=1}^n \{\tau\psi_1(X_i/\tau) - \mu\} \le -y\right] \le \exp\left\{nv^2\left(\frac{1}{\tau} - \frac{y}{2nv^2}\right)^2 - \frac{y^2}{4nv^2}\right\}.$$

In particular, taking $\tau = 2v^2 n/y$ gives

$$\mathbb{P}\left[\left|\sum_{i=1}^n \{\tau\psi_1(X_i/\tau) - \mu\}\right| \ge y\right] \le 2\exp\left(-\frac{y^2}{4nv^2}\right).$$

This proves Part (i) by taking $y = 2v(nz)^{1/2}$.

Part (ii) can be proved similarly. We therefore omit the details. □

A.3 Proof of Proposition 2

Proof of (i). Using the notation in Section A.1, equation (6) can be written as $q(\tau) = z/n$. By Lemma 1, the function q satisfies $\max_{t\geq 0} q(t) = \lim_{t\to 0} q(t) = \mathbb{P}(|X| > 0)$, $\lim_{t\to\infty} q(t) = 0$ and is strictly decreasing on (Δ, ∞). Provided $z/n < \mathbb{P}(|X| > 0)$, equation (6) has a unique solution that lies in (Δ, ∞).

By definition, this unique solution τ_z satisfies

$$\tau_z^2 = \mathbb{E}(X^2 \wedge \tau_z^2)\frac{n}{z} \leq \mathbb{E}(X^2)\frac{n}{z}. \tag{35}$$

On the other hand, note that $\mathbb{E}(X^2 \wedge \tau^2) \geq \tau^2\mathbb{P}(|X| > \tau)$ for any $\tau > 0$. It follows that $\mathbb{P}(|X| > \tau_z) \leq z/n$, which implies $\tau_z \geq q_{z/n}$. Substituting this into (35) gives $\tau_z^2 \geq \mathbb{E}(X^2 \wedge q_{z/n}^2)(n/z)$.

Proof of (ii). Recall that $q(\tau_z) = z/n$. Since $z/n \to 0$ and $q(t)$ strictly decreases to zero as $t \to \infty$, we have $\tau_z \to \infty$ and therefore $\mathbb{E}(X^2 \wedge \tau_z^2) \to \mathbb{E}(X^2)$ as $n \to \infty$. The stated results follow immediately. □

A.4 Proof of Proposition 3

Define

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(|X_i| > t), \quad q_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^2 \wedge t^2}{t^2}, \quad t > 0,$$

and $\Delta_n = \inf\{y > 0 : G_n(y) < G_n(0)\}$, which are the sample versions of $G(t)$, $q(t)$ and Δ given in (30), (31) and (34), respectively. A sample version of Lemma 1 prevails, implying that $q_n(t) = G_n(0)$ for $0 \leq t \leq \Delta_n$ and $q_n(\cdot)$ strictly decreases to zero on (Δ_n, ∞). Therefore, equation (4) has a unique solution on (Δ_n, ∞) if and only if $z/n < G_n(0)$. □

A.5 Proof of Theorem 1

Keep the notation used in the proof of Proposition 3. Recall that $\widehat{\tau}_z$ is uniquely determined and positive on the event $\{z < G_n(0)\}$. Under the condition $\mathbb{P}(X = 0) = 0$, it follows that $\mathbb{P}\{G_n(0) < 1\} = 0$ and therefore $\widehat{\tau}_z$ is positive with probability one. We divide the rest of the proof into four steps.

Step 1 (Preliminaries). Define the function

$$p_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^2 I(|X_i| \le t)}{t^2} \quad \text{for } t > 0.$$

Applying Lemma 1 to p_n and q_n, we see that $q_n'(t) = -2t^{-1}p_n(t)$. It follows that

$$q_n(\tau_z) - q_n(\widehat{\tau}_z) = 2\int_{\tau_z}^{\widehat{\tau}_z} \frac{p_n(t)}{t}dt = 2\int_0^{(\widehat{\tau}_z-\tau_z)/\tau_z} \frac{p_n(\tau_z+\tau_z u)}{1+u}du$$

by change of variables $u = (t-\tau_z)/\tau_z$. By definition, $q_n(\widehat{\tau}_z) = z/n = q(\tau_z)$. This, together with the last display, delivers

$$q_n(\tau_z) - q(\tau_z) = 2\int_0^{(\widehat{\tau}_z-\tau_z)/\tau_z} \frac{p_n(\tau_z+\tau_z u)}{1+u}du.$$

For any $r \in (0,1)$, it holds on the event $\{(\widehat{\tau}_z - \tau_z)/\tau_z \ge r\}$ that

$$\begin{aligned} q_n(\tau_z) - q(\tau_z) &\ge 2\int_0^r \frac{p_n(\tau_z+\tau_z u)}{1+u}du \\ &= 2\int_0^r \frac{p_n(\tau_z+\tau_z u) - p(\tau_z+\tau_z u)}{1+u}du + 2\int_0^r \frac{p(\tau_z+\tau_z u)}{1+u}du \\ &= 2\int_0^r \frac{p_n(\tau_z+\tau_z u) - p(\tau_z+\tau_z u)}{1+u}du + \{q(\tau_z) - q(\tau_z+\tau_z r)\} \\ &=: R_1 + D_1. \end{aligned}$$

Similarly, on the event $\{(\widehat{\tau}_z - \tau_z)/\tau_z \le -r\}$, it holds

$$\begin{aligned} & q_n(\tau_z) - q(\tau_z) \\ \le & -\{q(\tau_z - \tau_z r) - q(\tau_z)\} - 2\int_{-r}^0 \frac{p_n(\tau_z+\tau_z u) - p(\tau_z+\tau_z u)}{1+u}du \\ =: & -D_2 + R_2. \end{aligned}$$

Putting the above calculations together, we arrive at

$$\mathbb{P}(|\widehat{\tau}_z/\tau_z - 1| \ge r) \le \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \ge D_1 + R_1\} + \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \le -D_2 + R_2\}. \quad (36)$$

Set $\zeta_i = (X_i^2 \wedge \tau_z^2)/\tau_z^2$ for $i = 1, \ldots, n$ such that $q_n(\tau_z) - q(\tau_z) = n^{-1}\sum_{i=1}^n \{\zeta_i - \mathbb{E}(\zeta_i)\}$. Note that ζ_i's are bounded random variables satisfying $0 \le \zeta_i \le \min\{1, (|X_i| \wedge \tau_z)/\tau_z\}$ and $\mathbb{E}(\zeta_i^2) \le \mathbb{E}(X_i^2 \wedge \tau_z^2)/\tau_z^2 = z/n$. By Bernstein's inequality, for any $u > 0$ it holds

$$\mathbb{P}\{q_n(\tau_z) - q(\tau_z) \ge u/n\} \le \exp\{-u^2/(2z + 2u/3)\}. \quad (37)$$

On the other hand, applying Theorem 2.19 in de la Peña, Lai and Shao (2009) with $X_i = \zeta_i/n$ therein gives that, for any $0 < u < z$,

$$\mathbb{P}\{q_n(\tau_z) - q(\tau_z) \le -u/n\} \le \exp\{-u^2/(2z)\}. \tag{38}$$

Step 2 (Controlling R_1 and R_2). Note that R_1 and R_2 can be written, respectively, as $R_1 = 2n^{-1}\sum_{i=1}^n\{\xi_i - \mathbb{E}(\xi_i)\}$ and $R_2 = -2n^{-1}\sum_{i=1}^n\{\eta_i - \mathbb{E}(\eta_i)\}$, where

$$\xi_i = \int_0^r \frac{X_i^2 I\{|X_i| \le \tau_z(1+u)\}}{\tau_z^2(1+u)^3}du \quad \text{and} \quad \eta_i = \int_{-r}^0 \frac{X_i^2 I\{|X_i| \le \tau_z(1+u)\}}{\tau_z^2(1+u)^3}du$$

are bounded, nonnegative random variables satisfying

$$\xi_i \le \int_0^r \frac{du}{1+u} \le r, \quad \eta_i \le \int_{-r}^0 \frac{du}{1+u} \le \frac{r}{1-r}.$$

In addition,

$$\mathbb{E}(\xi_i^2) \le \frac{\mathbb{E}[X_i^2 I\{|X_i| \le \tau_z(1+r)\}]}{\tau_z^2}\left\{\int_0^r \frac{du}{(1+u)^2}\right\}^2 \le q(\tau_z + \tau_z r)r^2 \le q(\tau_z)r^2,$$

and

$$\mathbb{E}(\eta_i^2) \le \frac{\mathbb{E}\{X_i^2 I(|X_i| \le \tau_z)\}}{\tau_z^2}\left\{\int_{-r}^0 \frac{du}{(1+u)^2}\right\}^2 \le \frac{q(\tau_z)r^2}{(1-r)^2}.$$

Again it follows from Theorem 2.19 in de la Peña, Lai and Shao (2009) that, for any $v > 0$,

$$\mathbb{P}(R_1 \le -2rv/n) \le \exp\{-v^2/(2z)\} \tag{39}$$

$$\text{and} \quad \mathbb{P}\{R_2 \ge 2rv/(1-r)n\} \le \exp\{-v^2/(2z)\}. \tag{40}$$

Step 3 (Bounding D_1 and D_2). By Lemma 1 we have

$$D_1 = q(\tau_z) - q(\tau_z + \tau_z r) = 2\int_{\tau_z}^{\tau_z(1+r)} \frac{P(u)}{u^3}du \ge 2P(\tau_z)\int_{\tau_z}^{\tau_z(1+r)} \frac{1}{u^3}du = \frac{r^2+2r}{(1+r)^2}\frac{P(\tau_z)}{\tau_z^2}. \tag{41}$$

Similarly,

$$D_2 = q(\tau_z - \tau_z r) - q(\tau_z) = 2\int_{\tau_z(1-r)}^{\tau_z} \frac{P(u)}{u^3}du \ge \frac{2r - r^2}{(1-r)^2}\frac{P(\tau_z - \tau_z r)}{\tau_z^2}. \tag{42}$$

Step 4. Together, (36) and (39)–(42) imply that, for any $0 < r < 1$ and $v > 0$,

$$\begin{aligned}
&\mathbb{P}(|\widehat{\tau}_z/\tau_z - 1| \ge r) \\
&\le \mathbb{P}\left\{q_n(\tau_z) - q(\tau_z) \ge \frac{r^2+2r}{(1+r)^2}\frac{P(\tau_z)}{\tau_z^2} - \frac{2rv}{n}\right\} \\
&\quad + \mathbb{P}\left\{q_n(\tau_z) - q(\tau_z) \le -\frac{2r-r^2}{(1-r)^2}\frac{P(\tau_z - \tau_z r)}{\tau_z^2} + \frac{2rv}{(1-r)n}\right\} + 2\exp\{-v^2/(2z)\}.
\end{aligned} \tag{43}$$

Note that

$$\frac{r^2+2r}{(1+r)^2}\frac{P(\tau_z)}{\tau_z^2}-\frac{2rv}{n}=\left\{\frac{P(\tau_z)}{Q(\tau_z)}\frac{2+r}{(1+r)^2}z-2v\right\}\frac{r}{n}$$

and

$$\frac{2r-r^2}{(1-r)^2}\frac{P(\tau_z-\tau_z r)}{\tau_z^2}-\frac{2rv}{(1-r)n}=\left\{\frac{P(\tau_z-\tau_z r)}{Q(\tau_z)}\frac{2-r}{1-r}z-2v\right\}\frac{r}{(1-r)n}.$$

Taking $v=(a_1\wedge a_2)z/2$ for a_1 and a_2 as in (8), the right-hand side of (43) can further be bounded by

$$\mathbb{P}\left\{q_n(\tau_z)-q(\tau_z)\geq\frac{a_1rz}{n}\right\}+\mathbb{P}\left\{q_n(\tau_z)-q(\tau_z)\leq-\frac{a_2rz}{n}\right\}+2\exp\{-v^2/(2z)\}.$$

Combining this with (37), (38) and (43) proves the stated result. □

A.6 Proof of Theorem 2

We start with making a finite approximation of the interval $[1/2,3/2]$ using a sequence $\{c_k\}_{k=1}^n$ of equidistant points $c_k=1/2+k/n$. Then for any $\tau_z^*/2\leq\tau\leq 3\tau_z^*/2$ with $\tau_z^*=v_2\sqrt{n/z}$, there exists some $1\leq k\leq n$ such that $|\tau-\tau_{z,k}^*|\leq v_2(nz)^{-1/2}$, where $\tau_{z,k}^*:=c_kv_2\sqrt{n/z}$. It follows that

$$\sup_{\tau_z^*/2\leq\tau\leq 3\tau_z^*/2}|m_\tau-\mu|\leq\max_{1\leq k\leq n}|m_{\tau_{z,k}^*}-\mu|+\frac{v_2}{\sqrt{nz}}. \tag{44}$$

For $1\leq k<n/2$ so that $1/2\leq c_k<1$, by Proposition 1–(ii) we have $|m_{\tau_{z,k}^*}-\mu|\leq 2(v_2/c_k)\sqrt{z/n}$ with probability at least $1-2e^{-z/c_k^2}$; for $n/2\leq k\leq n$ so that $1\leq c_k\leq 3/2$, from Proposition 1–(i) it follows that $|m_{\tau_{z,k}^*}-\mu|\leq 2c_kv_2\sqrt{z/n}$ with probability at least $1-2e^{-z}$. Apply the union bound over $1\leq k\leq n$ to see that

$$\max_{1\leq k\leq n}|m_{\tau_{z,k}^*}-\mu|\leq 4v_2\sqrt{\frac{z}{n}} \tag{45}$$

with probability at least $1-2ne^{-z}$. Together, (44) and (45) prove (9).

Taking $z=2\log n$ in Proposition 2, Theorem 1 and Remark 1, we find that $\tau_z^*/2\leq\widehat{\tau}_z\leq 3\tau_z^*/2$ with probability at least $1-4n^{-c}$ for all sufficiently large n. The desired result then follows from (9). □

B Proofs of Results in Section 3.1

B.1 Proof of Proposition 5

Define functions $G(\boldsymbol{\theta}) = G(\beta_0, \boldsymbol{\beta}) = \mathbb{E}\{\ell_\tau(Y - \boldsymbol{Z}^\intercal\boldsymbol{\theta})\} = \mathbb{E}\{\ell_\tau(Y - \beta_0 - \boldsymbol{X}^\intercal\boldsymbol{\beta})\}$ and $h(\alpha) = \mathbb{E}\{\ell_\tau(\varepsilon - \alpha)\}$. By the definition and uniqueness of α_τ, for any $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\intercal)^\intercal$ we have

$$\begin{aligned} G(\boldsymbol{\theta}) &= \mathbb{E}\{\ell_\tau(\varepsilon - (\beta_0 - \beta_0^*) - \langle \boldsymbol{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle)\} \\ &= \mathbb{E}[\mathbb{E}\{\ell_\tau(\varepsilon_i - (\beta_0 - \beta_0^*) - \langle \boldsymbol{X}, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) | \boldsymbol{X}\}] \\ &\geq \mathbb{E}\{\ell_\tau(\varepsilon - \alpha_\tau)\} = G(\widetilde{\boldsymbol{\theta}}_\tau^*), \end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_\tau^* := (\beta_0^* + \alpha_\tau, \boldsymbol{\beta}^{*\intercal})^\intercal \in \mathbb{R}^{d+1}$. This implies that $G(\beta_0^* + \alpha_\tau, \boldsymbol{\beta}^*) = \min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} G(\boldsymbol{\theta})$. Moreover, consider the Hessian matrix $\nabla^2 G(\boldsymbol{\theta}) = \mathbb{E}\{I(|Y - \boldsymbol{Z}^\intercal\boldsymbol{\theta}| \leq \tau)\boldsymbol{Z}\boldsymbol{Z}^\intercal\}$, $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$. By (17), $\nabla^2 G(\widetilde{\boldsymbol{\theta}}_\tau^*) = \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau)\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$ is positive definite, such that $\widetilde{\boldsymbol{\theta}}_\tau^*$ is the unique minimizer of the function $\boldsymbol{\theta} \mapsto G(\boldsymbol{\theta})$. This enforces $\beta_{0,\tau}^* = \beta_0^* + \alpha_\tau$ and $\boldsymbol{\beta}_\tau^* = \boldsymbol{\beta}^*$.

Next we prove (19). By the optimality of α_τ and the mean value theorem, we have $h'(\alpha_\tau) = dh(\alpha)/d\alpha|_{\alpha=\alpha_\tau} = 0$ and

$$h''(\widetilde{\alpha}_\tau)\alpha_\tau = h'(\alpha_\tau) - h'(0) = -h'(0) = \mathbb{E}\{\ell_\tau'(\varepsilon)\}. \tag{46}$$

where $\widetilde{\alpha}_\tau = \lambda 0 + (1 - \lambda)\alpha_\tau$ for some $0 \leq \lambda \leq 1$. Note that

$$h''(\widetilde{\alpha}_\tau) = 1 - \mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau). \tag{47}$$

By the convexity of h, $h(\widetilde{\alpha}_\tau) \leq \lambda h(0) + (1 - \lambda)h(\alpha_\tau) \leq h(0) \leq \sigma^2/2$. On the other hand,

$$h(\alpha) \geq \mathbb{E}(\tau|\varepsilon - \alpha| - \tau^2/2)I(|\varepsilon - \alpha| > \tau) \quad \text{for all } \alpha \in \mathbb{R}.$$

Together, the upper and lower bounds on $h(\widetilde{\alpha}_\tau)$ yield

$$\tau\mathbb{E}|\varepsilon - \widetilde{\alpha}_\tau| I(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) \leq \frac{\tau^2}{2}\mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) + \frac{\sigma^2}{2}.$$

Further, by Markov's inequality,

$$\mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) \leq \tau^{-1}\mathbb{E}|\varepsilon - \widetilde{\alpha}_\tau| I(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) \leq \frac{1}{2}\mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) + \frac{\sigma^2}{2\tau^2},$$

implying $\mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) \leq \tau^{-2}\sigma^2$. Substituting this into (47) to reach

$$h''(\widetilde{\alpha}_\tau) \geq 1 - \tau^{-2}\sigma^2. \tag{48}$$

For the right-hand side of (46), we have

$$|\mathbb{E}\{\ell_\tau'(\varepsilon)\}| \leq \mathbb{E}(|\varepsilon| - \tau)I(|\varepsilon| > \tau) \leq \frac{1}{2\tau}\mathbb{E}(\varepsilon^2 - \tau^2)I(|\varepsilon| > \tau) = \frac{\sigma^2}{2\tau} - \frac{\mathbb{E}\{\psi_\tau^2(\varepsilon)\}}{2\tau}$$

where $\psi_\tau(x) = \ell_\tau'(x)$. Combined with (46) and (48), this proves (19) as long as $\tau > \sigma$. □

B.2 Proof of Theorem 3

Without loss of generality, we assume $2e^{-z} \le 1$ throughout the proof. For some $r > 0$ to be determined, define the local neighborhood

$$\boldsymbol{\Theta}_r = \big\{\boldsymbol{\theta} \in \mathbb{R}^{d+1} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2} \le r\big\}, \tag{49}$$

where $\|\cdot\|_{\mathbf{S},2}$ denotes the rescaled ℓ_2-norm $\|\boldsymbol{u}\|_{\mathbf{S},2} = \|\mathbf{S}^{1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u} \in \mathbb{R}^{d+1}$. If $\widehat{\boldsymbol{\theta}}_\tau \notin \boldsymbol{\Theta}_r$, there exists some $\eta \in (0,1)$ such that $\widehat{\boldsymbol{\theta}}_{\tau,\eta} = \boldsymbol{\theta}^* + \eta(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*) \in \boldsymbol{\Theta}_r$; otherwise if $\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_r$, we can simply take $\eta = 1$. By the optimality of $\widehat{\boldsymbol{\theta}}_\tau$, we have $\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{0}$. Applying Lemma A.1 in Sun, Zhou and Fan (2017) to $\mathcal{L}_\tau(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta})$ gives

$$\begin{aligned}
\big\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_{\tau,\eta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\big\rangle &\le \eta\big\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\big\rangle \\
&= \eta\big\langle -\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\big\rangle \\
&\le \|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 \times \|\widehat{\boldsymbol{\theta}}_{\tau,\eta}\|_{\mathbf{S},2}.
\end{aligned} \tag{50}$$

In what follows, we bound the left-hand and right-hand sides of (50) separately, staring with the former. Proposition 6 below shows that $\mathcal{L}_\tau$ is strictly convex on $\boldsymbol{\Theta}_r$ with high probability.

Proposition 6. Assume that $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{Z}\rangle^4 \le \kappa^4\langle \boldsymbol{u}, \mathbf{S}\boldsymbol{u}\rangle^2$ for all $\boldsymbol{u} \in \mathbb{R}^{d+1}$ and some $\kappa > 0$. Let $\tau, r > 0$ satisfy

$$\tau \ge \max(4\sigma, 8\kappa^2 r) \quad \text{and} \quad n \ge c_0(\tau/r)^2(d+z), \tag{51}$$

where $c_0 > 0$ is an absolute constant. Then with probability at least $1 - e^{-z}$,

$$\langle \nabla\mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \ge \frac{1}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2 \quad \text{uniformly over} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r. \tag{52}$$

Since $\widehat{\boldsymbol{\theta}}_{\tau,\eta} \in \boldsymbol{\Theta}_r$ by construction, it holds under the scaling (51) that

$$\big\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_{\tau,\eta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\big\rangle \ge \frac{1}{4}\|\widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2 \tag{53}$$

with probability at least $1 - e^{-z}$.

Next we bound the quadratic form $\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2$, which is bounded by

$$\bigg\|\frac{1}{n}\sum_{i=1}^n \big\{\xi_i\boldsymbol{z}_i - \mathbb{E}(\xi_i\boldsymbol{z}_i)\big\}\bigg\|_2 + \bigg\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\xi_i\boldsymbol{z}_i)\bigg\|_2, \tag{54}$$

where $\xi_i = \ell_\tau'(\varepsilon_i)$ and $\boldsymbol{z}_i = \mathbf{S}^{-1/2}\boldsymbol{Z}_i$. For the first term in (54), define $\boldsymbol{\gamma} = n^{-1}\sum_{i=1}^n\{\xi_i\boldsymbol{z}_i - \mathbb{E}(\xi_i\boldsymbol{z}_i)\}$. To bound $\|\boldsymbol{\gamma}\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^d}\boldsymbol{u}^\intercal\boldsymbol{\gamma}$, by a standard covering argument, we can find a

1/2-net $\mathcal{N}_{1/2}$ of $\mathbb{S}^d$ with $|\mathcal{N}_{1/2}| \le 5^{d+1}$ such that $\|\boldsymbol{\gamma}\|_2 \le 2\max_{\boldsymbol{u}\in\mathcal{N}_{1/2}} |\boldsymbol{u}^\intercal\boldsymbol{\gamma}|$. For every $\boldsymbol{u} \in \mathbb{S}^d$, note that $|\boldsymbol{u}^\intercal\boldsymbol{\gamma}| = |\sum_{i=1}^n \{\xi_i \boldsymbol{u}^\intercal \boldsymbol{z}_i - \mathbb{E}\xi_i \boldsymbol{u}^\intercal \boldsymbol{z}_i\}|$. Under Condition 1, it holds

$$\mathbb{E}|\boldsymbol{u}^\intercal \boldsymbol{z}_i|^k \le A_0^k \, k\Gamma(k/2) \quad \text{for all } k \ge 1. \tag{55}$$

If $k = 2\ell$ for some $\ell \ge 1$, $\mathbb{E}|\boldsymbol{u}^\intercal \boldsymbol{z}_i|^k \le 2A_0^k (k/2)!$; otherwise if $k = 2\ell + 1$ for some $\ell \ge 1$,

$$\mathbb{E}|\boldsymbol{u}^\intercal \boldsymbol{z}_i|^k \le k A_0^k \, \Gamma(\ell + 1/2) = k\sqrt{\pi} A_0^k \frac{(2\ell)!}{4^\ell \ell!} = 2\sqrt{\pi} A_0^k \frac{k!}{2^k \ell!}.$$

Putting the above calculations together to reach

$$\begin{aligned} \mathbb{E}(\xi_i \boldsymbol{u}^\intercal \boldsymbol{z}_i)^2 &= \mathbb{E}\{\mathbb{E}(\xi_i^2 | \boldsymbol{z}_i)(\boldsymbol{u}^\intercal \boldsymbol{z}_i)^2\} \le \sigma^2 2A_0^2, \\ \text{and} \quad \mathbb{E}|\xi_i \boldsymbol{u}^\intercal \boldsymbol{z}_i|^k &\le \frac{k!}{2} \sigma^2 2A_0^2 (A_0\tau/2)^{k-2} \quad \text{for all } k \ge 3. \end{aligned}$$

By Bernstein's inequality,

$$\mathbb{P}\left(|\boldsymbol{u}^\intercal \boldsymbol{\gamma}| \ge 2A_0\sigma \sqrt{\frac{x}{n}} + \frac{A_0}{2}\frac{\tau x}{n} \right) \le 2e^{-x} \quad \text{for any } x > 0.$$

Taking the union bound over all vectors $\boldsymbol{u} \in \mathcal{N}_{1/2}$, we obtain that with probability at least $5^{d+1} \cdot 2e^{-x}$, $\|\boldsymbol{\gamma}\|_2 \le 2\max_{\boldsymbol{u}\in\mathcal{N}_{1/2}} |\boldsymbol{u}^\intercal\boldsymbol{\gamma}| \le 4\sigma A_0 \sqrt{x/n} + A_0\tau x/n$. Taking $x = 2(d + 1 + z) \ge \log(5^{d+1}) + 2z$, we arrive at

$$\mathbb{P}\left\{ \|\boldsymbol{\gamma}\|_2 \ge 4\sqrt{2} A_0 \sigma \sqrt{\frac{d+1+z}{n}} + 2A_0\tau \frac{d+1+z}{n} \right\} \le 2e^{-2z} \le e^{-z}.$$

For the deterministic part $\|\mathbb{E}(\xi_i \boldsymbol{z}_i)\|_2$ in (54), it holds

$$\|\mathbb{E}(\xi_i \boldsymbol{z}_i)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^d} \mathbb{E}(\xi_i \boldsymbol{u}^\intercal \boldsymbol{z}_i) \le \sigma^2 \tau^{-1}.$$

Putting the above calculations together yields that with probability greater than $1 - e^{-z}$,

$$\|\mathbf{S}^{-1/2} \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 \le r_0 := (1 + 4\sqrt{2}A_0)\sigma \sqrt{\frac{d+1+z}{n}} + 2A_0\tau \frac{d+1+z}{n}. \tag{56}$$

Finally, in view of (53) and (56), we take $r = \tau/(8\kappa^2)$. It then follows that with probability greater than $1 - 2e^{-z}$, $\|\widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2} \le 4\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 \le 4r_0$ under the assumed scaling (51). Provided $n \gtrsim A_0\kappa^2(d + z)$ so that $4r_0 < r$, the intermediate estimator $\widehat{\boldsymbol{\theta}}_{\tau,\eta}$ will lie in the interior of $\boldsymbol{\Theta}_r$, which enforces $\eta = 1$ and $\widehat{\boldsymbol{\theta}}_{\tau,\eta} = \widehat{\boldsymbol{\theta}}_\tau$ (otherwise $\widehat{\boldsymbol{\theta}}_{\tau,\eta}$ will lie on the boundary). Putting together the pieces, we arrive at the desired result. □

B.3 Proof of Theorem 4

In view of the proof of Theorem 3, lying in the heart of the arguments is the restricted strong convexity (52) and the deviation bound (56) for a random quadratic form. In the following, we will establish similar results to (52) and (56) when τ is set as a constant rather than a function of (n, d). Since the target parameter now is $\boldsymbol{\theta}_\tau^*$, we slightly change the notation and set

$$\widetilde{\boldsymbol{\theta}}_{\tau,\eta} = \boldsymbol{\theta}_\tau^* + \eta(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau^*) \quad \text{and} \quad \widetilde{\boldsymbol{\Theta}}_r = \{\boldsymbol{\theta} \in \mathbb{R}^{d+1} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2} \le r\}$$

to be the intermediate estimator and the parameter set, respectively.

We start with the deviation bound. Recalling $\boldsymbol{\theta}_\tau^* = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}} \sum_{i=1}^n \mathbb{E}\ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta})$, it follows from Proposition 5 that

$$\mathbf{0} = \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*) = \mathbb{E}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*) = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\{\ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{Z}_i\},$$

where $\mathcal{L}_\tau(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta})$. Let $\boldsymbol{z}_i = \mathbf{S}^{-1/2}\boldsymbol{Z}_i$ such that

$$\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i\right\|_2. \tag{57}$$

Since $\mathbb{E}\{\ell_\tau'(\varepsilon_i - \alpha_\tau)\} = 0$ and by the optimality of α_τ,

$$\begin{aligned}
&\operatorname{var}(\ell_\tau'(\varepsilon_i - \alpha_\tau)) = \mathbb{E}\{\ell_\tau'(\varepsilon_i - \alpha_\tau)\}^2 \\
&= \mathbb{E}(\varepsilon_i - \alpha_\tau)^2 I(|\varepsilon_i - \alpha_\tau| \le \tau) \le 2\mathbb{E}\ell_\tau(\varepsilon_i - \alpha_\tau) \le 2\mathbb{E}\ell_\tau(\varepsilon) \le \sigma^2.
\end{aligned}$$

Following the same argument as in the proof of Theorem 3, it can be shown that with probability at least $1 - 2\cdot 5^{d+1}e^{-x}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i\right\|_2 \le 4A_0\sigma\sqrt{\frac{x}{n}} + A_0\tau\frac{x}{n}.$$

Taking $x = 2(d + 1 + z)$ in the last display, we obtain from (57) that

$$\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 \le r_1 := 4\sqrt{2}A_0\sigma\sqrt{\frac{d+1+z}{n}} + 2A_0\tau\frac{d+1+z}{n} \tag{58}$$

with probability at least $1 - e^{-z}$.

The next proposition provides the restricted strong convexity around $\boldsymbol{\theta}_\tau^*$ when τ is treated as a constant.

Proposition 7. Assume that $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{Z}\rangle^4 \le \kappa^4 \langle \boldsymbol{u}, \mathbf{S}\boldsymbol{u}\rangle^2$ for all $\boldsymbol{u} \in \mathbb{R}^{d+1}$ and some $\kappa > 0$. Let $r > 0$ satisfy

$$r \le \frac{1}{8}\rho_\tau^{1/2}\kappa^{-2}\tau \quad \text{and} \quad n \ge c_0 \rho_\tau^{-2}(\tau/r)^2(d+z), \tag{59}$$

where $c_0 > 0$ is an absolute constant. Then with probability at least $1 - e^{-z}$,

$$\langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*), \boldsymbol{\theta} - \boldsymbol{\theta}_\tau^* \rangle \ge \frac{\rho_\tau}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2}^2 \ \text{ uniformly over } \ \boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}_r. \tag{60}$$

According to (58) and (60), we take $r = \rho_\tau^{1/2}\kappa^{-2}\tau/8$ so that with probability at least $1 - 2e^{-z}$, $\|\widetilde{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2} \le 4\rho_\tau^{-1}\|\mathbf{S}^{-1/2}\nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 \le 4\rho_\tau^{-1} r_1$ under the scaling (59). Provided $n \gtrsim \rho_\tau^{-3}(A_0\kappa^2)^2(d+z)$ so that $4\rho_\tau^{-1}r_1 < r = \rho_\tau^{1/2}\kappa^{-2}\tau/8$, the intermediate estimator $\widetilde{\boldsymbol{\theta}}_{\tau,\eta}$ will lie in the interior of $\widetilde{\boldsymbol{\Theta}}_r$, which enforces $\eta = 1$ and $\widetilde{\boldsymbol{\theta}}_{\tau,\eta} = \widehat{\boldsymbol{\theta}}_\tau$, as desired. □

B.4 Proof of Proposition 6

Since the Huber loss is convex and differentiable, we have

$$\begin{aligned}
\mathcal{T}(\boldsymbol{\theta}) &:= \langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \\
&= \frac{1}{n}\sum_{i=1}^n \big\{\ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}^*) - \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})\big\}\boldsymbol{Z}_i^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\ge \frac{1}{n}\sum_{i=1}^n \big\{\ell_\tau'(\varepsilon_i) - \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta})\big\}\boldsymbol{Z}_i^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^*) I_{\mathcal{E}_i},
\end{aligned} \tag{61}$$

where $I_{\mathcal{E}_i}$ is the indicator function of the event

$$\mathcal{E}_i := \big\{|\varepsilon_i| \le \tau/2\big\} \cap \left\{\frac{|\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}} \le \frac{\tau}{2r}\right\}, \tag{62}$$

on which $|Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}| \le \tau$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_r$. Also, recall that $\ell_\tau''(u) = 1$ for $|u| \le \tau$. For any $R > 0$, define functions

$$\varphi_R(u) = \begin{cases} u^2 & \text{if } |u| \le \frac{R}{2}, \\ (u-R)^2 & \text{if } \frac{R}{2} \le u \le R, \\ (u+R)^2 & \text{if } -R \le u \le -\frac{R}{2}, \\ 0 & \text{if } |u| > R, \end{cases} \quad \text{and} \quad \psi_R(u) = I(|u| \le R).$$

In particular, φ_R is R-Lipschitz and satisfies

$$u^2 I(|u| \le R/2) \le \varphi_R(u) \le u^2 I(|u| \le R). \tag{63}$$

It then follows that

$$\mathcal{T}(\boldsymbol{\theta}) \geq g(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2}/(2r)}(\langle \boldsymbol{Z}_i, \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle)\psi_{\tau/2}(\varepsilon_i). \tag{64}$$

To bound the right-hand side of (64), consider the supremum of a random process indexed by $\boldsymbol{\Theta}_r$:

$$\Delta_r := \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \frac{|g(\boldsymbol{\theta}) - \mathbb{E}g(\boldsymbol{\theta})|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2}^2}. \tag{65}$$

For any $\boldsymbol{\theta}$ fixed, write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. By (63),

$$\begin{aligned}
\mathbb{E}g(\boldsymbol{\theta}) &\geq \mathbb{E}\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle^2 - \mathbb{E}\left\{\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle^2 I\left(|\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle| \geq \frac{\tau}{4r}\|\boldsymbol{\delta}\|_{\mathbf{S},2}\right)\right\} - \mathbb{E}\big\{\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle^2 I(|\varepsilon_i| > \tau/2)\big\} \\
&\geq \|\boldsymbol{\delta}\|_{\mathbf{S},2}^2 - \frac{4}{\tau^2}\left(\frac{4r^2}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}^2}\mathbb{E}\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle^4 + \mathbb{E}\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle^2\varepsilon_i^2\right).
\end{aligned} \tag{66}$$

Recall that $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{Z}_i\rangle^4 \leq \kappa^4\|\boldsymbol{\delta}\|_{\mathbf{S},2}^4$ for all $\boldsymbol{u} \in \mathbb{R}^{d+1}$. Joint with (66), this implies

$$\mathbb{E}g(\boldsymbol{\theta}) \geq \|\boldsymbol{\delta}\|_{\mathbf{S},2}^2 - \|\boldsymbol{\delta}\|_{\mathbf{S},2}^2\big(\sigma^2 + 4\kappa^4 r^2\big)\frac{4}{\tau^2} \geq \frac{1}{2}\|\boldsymbol{\delta}\|_{\mathbf{S},2}^2 \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}_r, \tag{67}$$

where the last inequality holds if $\tau \geq \max(4\sigma, 8\kappa^2 r)$. By (64), (65) and (67),

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2}^2} \geq \frac{1}{2} - \Delta_r \quad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r. \tag{68}$$

The following lemma provides an upper bound on the stochastic term Δ_r.

Lemma 2. For any $x > 0$,

$$\Delta_r \leq \mathbb{E}\Delta_r + \left\{(\mathbb{E}\Delta_r)^{1/2}\frac{\tau}{2r} + \sqrt{2}\kappa^2\right\}\sqrt{\frac{x}{n}} + \frac{\tau^2}{16r^2}\frac{x}{3n}$$

with probability at least $1 - e^{-x}$. Moreover,

$$\mathbb{E}\Delta_r \leq \sqrt{\frac{2\pi}{n}}\left(\frac{2\tau}{r}\sqrt{d+1} + 1\right).$$

Substituting this into Lemma 2, we obtain that with probability at least $1 - e^{-z}$,

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2}^2} \geq \frac{1}{4} \quad \text{uniformly over } \boldsymbol{\theta} \in \boldsymbol{\Theta}_r$$

for all sufficiently large n that scales as $(\tau/r)^2(d+z)$ up to an absolute constant. This proves (52). □

B.5 Proof of Proposition 7

Following the proof of Proposition 6, now we have

$$\begin{aligned}&\langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*), \boldsymbol{\theta} - \boldsymbol{\theta}_\tau^* \rangle \\ &\geq \frac{1}{n}\sum_{i=1}^n \big\{ \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}_\tau^*) - \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}) \big\} \boldsymbol{Z}_i^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*) I_{\mathcal{E}_{i,\tau}},\end{aligned}$$

where

$$\mathcal{E}_{i,\tau} = \big\{ |\varepsilon_i - \alpha_\tau| \leq \tau/2 \big\} \cap \left\{ \frac{|\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_\tau^* \rangle|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2}} \leq \frac{\tau}{2r} \right\}$$

On $\mathcal{E}_{i,\tau}$, $|Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}_\tau^*| = |\varepsilon_i + \beta_0^* - \beta_{0,\tau}^*| = |\varepsilon_i - \alpha_\tau| \leq \tau$ and $|Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}| \leq \tau$ for all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}_r$. Moreover, let $g(\boldsymbol{\theta})$ be as in (64) except with $\boldsymbol{\theta}^*$ replaced by $\boldsymbol{\theta}_\tau^*$. By (22) and Markov's inequality, we obtain that for every $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}_r$,

$$\mathbb{E} g(\boldsymbol{\theta}) \geq \big(\rho_\tau - 16\kappa^4 r^2 \tau^{-2}\big) \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2}^2 \geq \frac{3}{4}\rho_\tau \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S},2}^2,$$

provided $r \leq \rho_\tau^{1/2} \kappa^{-2} \tau/8$. Keep all other statements the same, we then get the desired result. □

B.6 Proof of Lemma 2

For $g(\boldsymbol{\theta})$ given in (64), we write $g(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n g_i(\boldsymbol{\theta})$. Observing that $0 \leq \varphi_R(u) \leq R^2/4$ and $0 \leq \psi_R(u) \leq 1$ for all $u \in \mathbb{R}$, we have

$$0 \leq g_i(\boldsymbol{\theta}) \leq \frac{\tau^2}{16 r^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2.$$

It then follows from Theorem 7.3 in Bousquet (2003), a variant of Talagrand's inequality, that for any $x > 0$,

$$\Delta_r \leq \mathbb{E}\Delta_r + (\mathbb{E}\Delta_r)^{1/2} \frac{\tau}{2r}\sqrt{\frac{x}{n}} + \sigma_n \sqrt{\frac{2x}{n}} + \frac{\tau^2}{16 r^2} \times \frac{x}{3n} \tag{69}$$

with probability at least $1 - e^{-x}$, where $\sigma_n^2 := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_r} \mathbb{E} g_i^2(\boldsymbol{\theta}) / \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^4$. By (63), $\mathbb{E} g_i^2(\boldsymbol{\theta}) \leq \mathbb{E}\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle^4 \leq \kappa^4 (\boldsymbol{\delta}^\intercal \mathbf{S} \boldsymbol{\delta})^2$ with $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$, which implies $\sigma_n^2 \leq \kappa^4$.

It remains to bound the expectation $\mathbb{E}\Delta_r$. Applying the symmetrization inequality for empirical processes, and by the connection between Gaussian complexity and Rademacher complexity, we obtain $\mathbb{E}\Delta_r \leq 2\sqrt{\pi/2}\, \mathbb{E}(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_r} |\mathbb{G}_{\boldsymbol{\theta}}|)$, where

$$\mathbb{G}_{\boldsymbol{\theta}} := \frac{1}{n}\sum_{i=1}^n \frac{g_i}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2} \varphi_{\tau \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}/(2r)}(\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle) \psi_{\tau/2}(\varepsilon_i),$$

and g_i are i.i.d. standard normal random variables that are independent of the observed data. For any $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_r$, it holds

$$\mathbb{E}^*\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}|\mathbb{G}_{\boldsymbol{\theta}}|\right) \le \mathbb{E}^*|\mathbb{G}_{\boldsymbol{\theta}_0}| + 2\mathbb{E}^*\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\mathbb{G}_{\boldsymbol{\theta}}\right), \tag{70}$$

where $\mathbb{E}^*$ denotes the conditional expectation given $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Taking the expectation with respect to $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ on both sides, we see that (70) remains valid with $\mathbb{E}^*$ replaced by $\mathbb{E}$. For $\mathbb{E}|\mathbb{G}_{\boldsymbol{\theta}_0}|$, we take $\boldsymbol{\theta}_0 = (\beta_0^* + r, \boldsymbol{\beta}^{*\intercal})^\intercal \in \boldsymbol{\Theta}_r$ such that $\langle \boldsymbol{Z}_i, \boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\rangle = r$ and

$$\mathbb{G}_{\boldsymbol{\theta}_0} = \frac{\varphi_{\tau/2}(r)}{r^2 n}\sum_{i=1}^n g_i \psi_{\tau/2}(\varepsilon_i).$$

Then it follows from (63) that $\mathbb{E}|\mathbb{G}_{\boldsymbol{\theta}_0}| \le n^{-1/2}$. As in the proof of Lemma 11 in Loh and Wainwright (2015), we use the Gaussian comparison theorem to bound the expectation of the (conditional) Gaussian supremum $\mathbb{E}^*(\sup_{\boldsymbol{\Theta}_r}\mathbb{G}_{\boldsymbol{\theta}})$.

Let var* be the conditional variance given $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. For $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}_r$, write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and $\boldsymbol{\delta}' = \boldsymbol{\theta}' - \boldsymbol{\theta}^*$. Then

$$\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \le \frac{1}{n^2}\sum_{i=1}^n \psi_{\tau/2}^2(\varepsilon_i)\left\{\frac{\varphi_{\tau\|\boldsymbol{\delta}\|_{\mathbf{S},2}/(2r)}(\boldsymbol{Z}_i^\intercal\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}^2} - \frac{\varphi_{\tau\|\boldsymbol{\delta}'\|_{\mathbf{S},2}/(2r)}(\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}')}{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}^2}\right\}^2.$$

Note that $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c > 0$. In particular, taking $R = \tau\|\boldsymbol{\delta}'\|_{\mathbf{S},2}/(2r)$ and $c = \|\boldsymbol{\delta}\|_{\mathbf{S},2}/\|\boldsymbol{\delta}'\|_{\mathbf{S},2}$ delivers

$$\varphi_{\tau\|\boldsymbol{\delta}'\|_{\mathbf{S},2}/(2r)}(\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}') = \frac{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}^2}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}^2}\varphi_{\tau\|\boldsymbol{\delta}\|_{\mathbf{S},2}/(2r)}\left(\frac{\|\boldsymbol{\delta}\|_{\mathbf{S},2}}{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}}\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}'\right).$$

Putting the above calculations together, we obtain

$$\begin{aligned}
&\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \\
&\le \frac{1}{n^2}\sum_{i=1}^n \frac{1}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}^4}\left\{\varphi_{\tau\|\boldsymbol{\delta}\|_{\mathbf{S},2}/(2r)}(\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}) - \varphi_{\tau\|\boldsymbol{\delta}\|_{\mathbf{S},2}/(2r)}\left(\frac{\|\boldsymbol{\delta}\|_{\mathbf{S},2}}{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}}\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}'\right)\right\}^2 \\
&\le \frac{1}{n^2}\sum_{i=1}^n \frac{1}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}^4}\frac{\tau^2\|\boldsymbol{\delta}\|_{\mathbf{S},2}^2}{4r^2}\left(\boldsymbol{Z}_i^\intercal\boldsymbol{\delta} - \frac{\|\boldsymbol{\delta}\|_{\mathbf{S},2}}{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}}\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}'\right)^2 \\
&\le \frac{1}{n^2}\sum_{i=1}^n \frac{\tau^2}{4r^2}\left(\frac{\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_{\mathbf{S},2}} - \frac{\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}'}{\|\boldsymbol{\delta}'\|_{\mathbf{S},2}}\right)^2.
\end{aligned} \tag{71}$$

Next, define another (conditional) Gaussian process indexed by $\boldsymbol{\theta}$:

$$\mathbb{Z}_{\boldsymbol{\theta}} := \frac{\tau}{2rn}\sum_{i=1}^n g_i' \frac{\boldsymbol{Z}_i^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}},$$

where g_i' are i.i.d. standard normal random variables that are independent of all other random variables. By (71), $\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \leq \mathrm{var}^*(\mathbb{Z}_{\boldsymbol{\theta}} - \mathbb{Z}_{\boldsymbol{\theta}'})$. Using the Gaussian comparison inequality (Ledoux and Talagrand, 1991) yields

$$\mathbb{E}^*\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \mathbb{G}_{\boldsymbol{\theta}}\right) \leq 2\mathbb{E}^*\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r} \mathbb{Z}_{\boldsymbol{\theta}}\right) \leq \frac{\tau}{r}\mathbb{E}^*\left\|\frac{1}{n}\sum_{i=1}^n g_i\, \mathbf{S}^{-1/2}\boldsymbol{Z}_i\right\|_2.$$

Combining this with the unconditional version of (70), we obtain

$$\mathbb{E}\Delta_r \leq \sqrt{2\pi}\left(\frac{2\tau}{r}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i \boldsymbol{z}_i\right\|_2 + \frac{1}{\sqrt{n}}\right) \leq \sqrt{\frac{2\pi}{n}}\left(\frac{2\tau}{r}\sqrt{d+1}+1\right),$$

which, together with (69), proves the stated results. □

C Proof of Theorem 5

This proof is based on an argument similar to that used in the proof of Theorem 3. For simplicity, we write $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^\intercal)^\intercal = \widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau,\lambda) \in \mathbb{R}^{d+1}$. For some $r > 0$ to be specified, we use $\boldsymbol{\Theta}_r$ and $\|\cdot\|_{\mathbf{S},2}$ to denote the local neighborhood and rescaled ℓ_2-norm as in (49). As before, let $\widehat{\boldsymbol{\theta}}_\eta$ $(0 < \eta \leq 1)$ be an intermediate estimator satisfying (i) $\widehat{\boldsymbol{\theta}}_\eta \in \boldsymbol{\Theta}_r$, (ii) $\widehat{\boldsymbol{\theta}}_\eta$ lies on the boundary of $\boldsymbol{\Theta}_r$ with $\eta \in (0,1)$ if $\widehat{\boldsymbol{\theta}} \notin \boldsymbol{\Theta}_r$, and (iii) $\widehat{\boldsymbol{\theta}}_1 = \widehat{\boldsymbol{\theta}}$. Moreover, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_\eta$ fulfill

$$\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^*\rangle \leq \eta\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\rangle. \tag{72}$$

Write $\widehat{\boldsymbol{\delta}} = (\widehat{v}_0, \widehat{\boldsymbol{v}}^\intercal)^\intercal = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ and denote by $\mathcal{S} \subseteq \{1,\ldots,d\}$ the the support of $\boldsymbol{\beta}^*$. Define the cone $\mathcal{C} \subseteq \mathbb{R}^{d+1}$ as

$$\mathcal{C} = \left\{\boldsymbol{\theta} \in \mathbb{R}^{d+1} : \|\boldsymbol{v}_{\mathcal{S}^{\mathrm{c}}}\|_1 \leq 3\|\boldsymbol{v}_{\mathcal{S}}\|_1 + |v_0| \ \text{ for } (v_0, \boldsymbol{v}^\intercal)^\intercal = \boldsymbol{\theta} - \boldsymbol{\theta}^*\right\}.$$

It can be shown that the optimal solution $\widehat{\boldsymbol{\theta}}$ to program (25) satisfies

$$\widehat{\boldsymbol{\theta}} \in \mathcal{C} \ \text{ on the event } \ \left\{\lambda \geq 2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty\right\}, \tag{73}$$

from which it follows

$$\|\widehat{\boldsymbol{\delta}}\|_1 = |\widehat{v}_0| + \|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 + \|\widehat{\boldsymbol{v}}_{\mathcal{S}^{\mathrm{c}}}\|_1 \leq 2|\widehat{v}_0| + 4\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 \leq 4\sqrt{s+1}\,\|\widehat{\boldsymbol{\delta}}\|_2. \tag{74}$$

By necessary conditions of extrema in the convex optimization problem (25),

$$\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) + \lambda\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\rangle \leq 0,$$

where $\widehat{\boldsymbol{z}} = (0, \widehat{\boldsymbol{u}}^\intercal)^\intercal$ with $\widehat{\boldsymbol{u}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$ satisfies $\langle \widehat{\boldsymbol{z}}, \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}} \rangle \le \|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1$. Under the scaling $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$, it holds

$$\begin{aligned}
&\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle \le \lambda\big(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \\
&\le \lambda\big(\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \le \frac{\lambda}{2}\big(3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1\big) + \frac{\lambda}{2}|\widehat{v}_0|.
\end{aligned}$$

Together with (72), this implies

$$\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^* \rangle \le \frac{1}{2}\lambda\eta\big(3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1\big) + \frac{1}{2}\lambda\eta|\widehat{v}_0|. \tag{75}$$

Moreover, we introduce $\widehat{\boldsymbol{\delta}}_\eta = \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^*$ and note that $\widehat{\boldsymbol{\delta}}_\eta = \eta\widehat{\boldsymbol{\delta}}$. By (73), we also have $\widehat{\boldsymbol{\theta}}_\eta \in \mathcal{C}$ under the assumed scaling.

To bound the left-hand side of (75), the following proposition reveals that under proper scaling, the Huber loss satisfies the restricted strong convexity condition over $\boldsymbol{\Theta}_r \cap \mathcal{C}$ with high probability. It is a straightforward extension of Proposition 6. We leave the proof to Section C.2.

Proposition 8. Assume that $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{Z} \rangle^4 \le \kappa \langle \boldsymbol{u}, \boldsymbol{\mathrm{S}}\boldsymbol{u} \rangle^2$ for all $\boldsymbol{u} \in \mathbb{R}^{d+1}$ and some $\kappa > 0$. Let $\tau, r > 0$ satisfy

$$\tau \ge \max(4\sigma, 8\kappa^2 r) \quad \text{and} \quad n \ge c_0 \underline{\lambda}_{\boldsymbol{\mathrm{S}}}^{-1} A_0^2 \, (\tau/r)^2 \max_{1 \le j \le d} \sigma_{jj} \, s \log d, \tag{76}$$

where $c_0 > 0$ is an absolute constant. Then with probability at least $1 - d^{-1}$,

$$\langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \ge \frac{1}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\mathrm{S}},2}^2 \quad \text{uniformly over} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}. \tag{77}$$

Let Ω_r be the event on which (77) holds. Then $\mathbb{P}(\Omega_r^{\mathrm{c}}) \le d^{-1}$ under the scaling (76) and it holds on $\Omega_r \cap \{\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty\}$ that

$$\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^* \rangle \ge \frac{1}{4}\|\widehat{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\mathrm{S}},2}^2 \ge \frac{1}{4}\underline{\lambda}_{\boldsymbol{\mathrm{S}}}^{1/2}\|\widehat{\boldsymbol{\delta}}_\eta\|_2 \|\widehat{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\mathrm{S}},2}.$$

Substituting this lower bound into (75) yields

$$\frac{1}{4}\underline{\lambda}_{\boldsymbol{\mathrm{S}}}^{1/2}\|\widehat{\boldsymbol{\delta}}_\eta\|_2 \|\widehat{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\mathrm{S}},2} \le \frac{1}{2}\lambda\eta\big(|\widehat{v}_0| + 3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1\big) \le \frac{3}{2}\lambda\sqrt{s+1}\,\|\eta\widehat{\boldsymbol{\delta}}\|_2 = \frac{3}{2}\lambda\sqrt{s+1}\,\|\widehat{\boldsymbol{\delta}}_\eta\|_2.$$

Canceling $\|\widehat{\boldsymbol{\delta}}_\eta\|_2$ on both sides delivers

$$\|\widehat{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\mathrm{S}},2} \le \frac{6\lambda}{\underline{\lambda}_{\boldsymbol{\mathrm{S}}}^{1/2}}\sqrt{s+1} \quad \text{and} \quad \|\widehat{\boldsymbol{\delta}}_\eta\|_1 \le \frac{24\lambda}{\underline{\lambda}_{\boldsymbol{\mathrm{S}}}}(s+1), \tag{78}$$

under the assumed scaling $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$ and (76) .

It remains to tune the parameters τ, λ and r. The following result provides a concentration inequality for $\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$.

Proposition 9. Assume Condition 1 holds and let $\tau = \sigma\sqrt{n/t}$ for some $t > 0$. Then with probability at least $1 - 2(d^{-1} + e^{-t})$,

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty \le \max_{1\le j\le d} \sigma_{jj}^{1/2}\sigma\left(2\sqrt{2}A_0\sqrt{\frac{\log d}{n}} + A_0\frac{\log d}{\sqrt{nt}} + \sqrt{\frac{t}{n}}\right) \bigvee 2\sigma\sqrt{\frac{t}{n}}. \tag{79}$$

Applying Proposition 9 with $t = \log d$, we see that

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty \le c_1 A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2}\sigma\sqrt{\frac{\log d}{n}}$$

with probability at least $1 - 4d^{-1}$, where $c_1 > 0$ is an absolute constant. We therefore choose $\lambda = c_2 A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2}\sigma\sqrt{\log(d)/n}$ for some constant $c_2 \ge 2c_1$, such that $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$ with high probability. According to (76), we take $r = \tau/(8\kappa^2)$. Putting the above calculations together, we conclude that

$$\|\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^*\|_{\mathbf{S},2} \le 6c_2\underline{\lambda}_{\mathbf{S}}^{-1/2} A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2}\sigma\sqrt{\frac{(s+1)\log d}{n}} < r$$

with probability at least $1 - 5d^{-1}$, assuming the scaling $n \gtrsim A_0^2\kappa^4\underline{\lambda}_{\mathbf{S}}^{-1}\max_{1\le j\le d}\sigma_{jj}\, s\log d$. By the construction of $\widehat{\boldsymbol{\theta}}_\eta$, with the same probability we must have $\eta = 1$ and therefore $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\eta$. The stated result (26) then follows from (78). □

C.1 Proof of (73)

From the optimality of $\widehat{\boldsymbol{\theta}}$ we see that

$$\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \mathcal{L}_\tau(\boldsymbol{\theta}^*) \le \lambda\big(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1\big).$$

By direct calculation, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1 &\ge \|\boldsymbol{\beta}^*_{\mathcal{S}} + \widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1 - \|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \big(\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + \|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1\big)\\ &\ge \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1.\end{aligned}$$

By the convexity of $\mathcal{L}_\tau$ and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \mathcal{L}_\tau(\boldsymbol{\theta}^*) &\ge \langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\delta}}\rangle \ge -\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty\|\widehat{\boldsymbol{\delta}}\|_1\\ &\ge -\frac{\lambda}{2}\big(|\widehat{v}_0| + \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1 + \|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1\big).\end{aligned}$$

Putting the above calculations together delivers

$$0 \le \frac{\lambda}{2}\big(|\widehat{v}_0| + 3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1\big),$$

from which the conclusion follows. □

C.2 Proof of Proposition 8

The proof is almost identical to that of Proposition 6. With slight abuse of notation, define the supremum of a random process indexed by $\boldsymbol{\Theta}_r \cap \mathcal{C}$:

$$\Delta_r := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}} \frac{|g(\boldsymbol{\theta}) - \mathbb{E}g(\boldsymbol{\theta})|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2}.$$

Provided $\tau \geq \max(4\sigma, 8\kappa^2 r)$, it can be shown that

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2} \geq \frac{1}{2} - \Delta_r \quad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}. \tag{80}$$

The following lemma is a slight modification of Lemma 2.

Lemma 3. Let $\boldsymbol{G}_n = n^{-1} \sum_{i=1}^n g_i \boldsymbol{Z}_i$, where $g_1, \ldots, g_n$ are i.i.d. standard normal random variables that are independent of $\{(\boldsymbol{X}_i, \varepsilon_i)\}_{i=1}^n$. Then for any $x > 0$,

$$\Delta_r \leq \mathbb{E}\Delta_r + \left\{(\mathbb{E}\Delta_r)^{1/2}\tau/(2r) + \sqrt{2}\kappa^2\right\}\sqrt{\frac{x}{n}} + \frac{\tau^2}{48r^2}\frac{x}{n}$$

with probability at least $1 - e^{-x}$, and

$$\mathbb{E}\Delta_r \leq \sqrt{2\pi} \cdot \left\{8\underline{\lambda}_{\mathbf{S}}^{-1/2}(\tau/r)(s+1)^{1/2}\mathbb{E}\|\boldsymbol{G}_n\|_\infty + n^{-1/2}\right\}.$$

Write $\boldsymbol{Z}_i = (Z_{i0}, Z_{i1}, \ldots, Z_{id})^\intercal$ such that $\|\frac{1}{n}\sum_{i=1}^n g_i \boldsymbol{Z}_i\|_\infty = \max_{0 \leq j \leq d} |\frac{1}{n}\sum_{i=1}^n g_i Z_{ij}|$. By the sub-Gaussianity of $\boldsymbol{Z}$, Lemma 5.5 in Vershynin (2012) and the Legendre duplication formula, i.e. $\Gamma(s)\Gamma(s+1/2) = 2^{1-2s}\sqrt{\pi}\,\Gamma(2s)$, we calculate that for $j = 0, 1, \ldots, d$, $\mathbb{E}(g_i Z_{ij})^2 = \sigma_{jj}$ with $\sigma_{00} = 1$, and for $k \geq 3$,

$$\begin{aligned}
\mathbb{E}|g_i Z_{ij}|^k &\leq 2^{k/2}\frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi}} \cdot A_0^k \sigma_{jj}^{k/2} k\Gamma(k/2) \\
&= 2A_0^k \sigma_{jj}^{k/2}\frac{(k-1)!}{2^{k/2}} \leq \frac{m!}{2}A_0^2\sigma_{jj}\left(A_0\sqrt{\frac{\sigma_{jj}}{2}}\right)^{k-2}.
\end{aligned}$$

Then it follows from Lemma 14.12 in Bühlmann and van de Geer (2011) that

$$\mathbb{E}\left(\max_{0 \leq j \leq d}\left|\frac{1}{n}\sum_{i=1}^n g_i Z_{ij}\right|\right) \leq A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2}\left\{\sqrt{\frac{2\log(d+2)}{n}} + \frac{\log(d+2)}{n}\right\}.$$

Substituting this into Lemma 2 and taking $x = \log d$, we obtain that with probability at least $1 - d^{-1}$,

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S},2}^2} \geq \frac{1}{4} \quad \text{uniformly over } \boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}$$

for all sufficiently large n that scales as $A_0^2 \underline{\lambda}_{\mathbf{S}}^{-1}(\tau/r)^2 \max_{1 \leq j \leq d} \sigma_{jj}\, s \log d$ up to an absolute constant. This proves (77). □

C.3 Proof of Proposition 9

To begin with, write $\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*) = n^{-1}\sum_{i=1}^n \xi_i \boldsymbol{Z}_i$ such that

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty = \max_{1\le j\le d}\left|\frac{1}{n}\sum_{i=1}^n \xi_i X_{ij}\right| \bigvee \left|\frac{1}{n}\sum_{i=1}^n \xi_i\right|,$$

where $\xi_i = \ell_\tau'(\varepsilon_i)$ are i.i.d. bounded random variables with $|\xi_i| \le \min(\tau, |\varepsilon_i|)$. To bound $|n^{-1}\sum_{i=1}^n \xi_i|$, it follows immediately from Proposition 1 that $|n^{-1}\sum_{i=1}^n \xi_i| \le 2\sigma\sqrt{t/n}$ with probability at least $1-2e^{-t}$.

Next we use the union bound and Bernstein's inequality to bound the maximum. For every $1 \le j \le d$,

$$|\mathbb{E}(\xi_i X_{ij})| = |\mathbb{E}\{\mathbb{E}(\xi_i|X_{ij})X_{ij}\}| \le \mathbb{E}|X_{ij}|\sigma^2\tau^{-1} \le \sigma_{jj}^{1/2}\sigma^2\tau^{-1}$$
$$\text{and} \quad \mathbb{E}(\xi_i X_{ij})^2 = \mathbb{E}\{(\xi_i^2|X_{ij})X_{ij}^2\} \le \sigma_{jj}\sigma^2.$$

By the sub-Gaussianity of $\boldsymbol{X}_i$ and Lemma 5.5 in Vershynin (2012), we find that

$$\mathbb{E}|X_{ij}|^k \le C_j^k k\Gamma(k/2) \quad \text{for all } k \ge 2,$$

where $C_j = A_0\sigma_{jj}^{1/2}$. Using the same argument that leads to (56), it can be derived that $\mathbb{E}|\xi_i X_{ij}|^k \le k!\sigma^2 C_j^2(C_j\tau/2)^{k-2}$ for all $k \ge 3$. Then it follows from Bernstein' inequality that, for any $x > 0$,

$$\left|\frac{1}{n}\sum_{i=1}^n(\xi_i X_{ij} - \mathbb{E}\xi_i X_{ij})\right| \le 2\sigma C_j\sqrt{\frac{x}{n}} + C_j\frac{\tau x}{2n}$$

with probability at least $1-2e^{-x}$. Putting together the pieces and taking $x = 2\log d$, we arrive at the stated result. □

C.4 Proof of Lemma 3

Reviewing the proof of Lemma 2, we only need to bound the expectation $\mathbb{E}\Delta_r$. To this end, it suffices to focus on the (conditional) Gaussian process

$$\mathbb{Z}_{\boldsymbol{\theta}} = \frac{\tau}{2rn}\sum_{i=1}^n g_i'\frac{\boldsymbol{Z}_i^\intercal(\boldsymbol{\theta}-\boldsymbol{\theta}^*)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2}}, \quad \boldsymbol{\theta} \in \Theta_r \cap \mathcal{C},$$

where g_i' are i.i.d. standard normal random variables that are independent of all other random variables. For every $\boldsymbol{\theta} \in \Theta_r \cap \mathcal{C}$, similarly to (74) it holds

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_1 \le 4\sqrt{s+1}\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le 4\underline{\lambda}_{\mathbf{S}}^{-1/2}\sqrt{s+1}\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathbf{S},2},$$

implying

$$\sup_{\boldsymbol{\theta}\in\Theta_r\cap\mathcal{C}} \mathbb{Z}_{\boldsymbol{\theta}} \le 2\underline{\lambda}_{\mathbf{S}}^{-1/2}\sqrt{s+1}\,\frac{\tau}{r}\left\|\frac{1}{n}\sum_{i=1}^{n} g_i \boldsymbol{Z}_i\right\|_\infty.$$

Keep all other statements the same, we obtain

$$\mathbb{E}\Delta_r \le \sqrt{2\pi}\left(8\underline{\lambda}_{\mathbf{S}}^{-1/2}\sqrt{s+1}\,\frac{\tau}{r}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} g_i \boldsymbol{Z}_i\right\|_\infty + \frac{1}{\sqrt{n}}\right),$$

and then get the desired result. □

References

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *In Stochastic Inequalities and Applications. Progress in Probability,* 56, 213–247. Birkhäuser, Basel.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer, Berlin.

Hahn, M. G., Kuelbs, J. and Weiner, D. C. (1990). The asymptotic joint distribution of self-normalized censored sums and sums of squares. *The Annals of Probability,* 18, 1284–1341.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16, 559–616.

Sun, Q., Zhou, W.-X. and Fan, J. (2017). Adaptive Huber regression. *arXiv preprint arXiv:1706.06991.*

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 210–268. Cambridge Univ. Press, Cambridge.