A New Principle for Tuning-Free Huber Regression

Lili Wang[♭], Chao Zheng[♯], Wen Zhou[†], and Wen-Xin Zhou[¶]

[♭]*Zhejiang Gongshang University,* [♯]*University of Southampton*

[†]*Colorado State University,* [¶]*University of California San Diego*

Supplementary Material

This online Supplementary Material contains proofs of all theoretical results in the main text, as well as additional empirical studies. Results from Section 2 in the main text are justified and further discussed in Section S1. In Section S2, we prove the results from Section 3.1. Section S3 presents the proof for Theorem 5. Additional numerical studies are provided in Section S4. Section S5 reports real-data examples to further demonstrate the prediction performance of the DA-Huber methods. Finally, in Section S6, we discuss a general class of robust loss functions, to which the obtained results for Huber loss apply.

S1 Proofs of Results in Section 2

S1.1 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

$$G(t) = \mathbb{P}(|X| > t),\ P(t) = \mathbb{E}\{X^2 I(|X| \le t)\} \ \text{ and } \ Q(t) = \mathbb{E}\{\psi_t(X)\}^2, \tag{S1.1}$$

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

$$p(t) = t^{-2}P(t) \ \text{ and } \ q(t) = t^{-2}Q(t). \tag{S1.2}$$

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 S1.1. Let functions G, Q, p and q be given in (S1.1) and (S1.2).

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

$$Q(t) = 2\int_0^t yG(y)\,\mathrm{d}y, \quad q'(t) = -2t^{-1}p(t), \tag{S1.3}$$

and

$$q(t) = \mathbb{P}(X \neq 0) - 2\int_0^t y^{-1}p(y)\mathrm{d}y. \tag{S1.4}$$

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,

$$q(s) = \mathbb{P}(X \neq 0) \ \text{ for all } 0 \le s \le \Delta := \inf\{y > 0 : G(y) < \mathbb{P}(X \neq 0)\}, \tag{S1.5}$$

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

Proof of Lemma S1.1. Notice $Q(t) = \mathbb{E}\{(|X| \wedge t)^2\}$ and it holds almost surely that

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

Taking expectations on both sides implies $Q(t) = \mathbb{E}\{(|X|\wedge t)^2\} = 2\int_0^t \mathbb{P}(|X| > y)y \,\mathrm{d}y = 2\int_0^t yG(y)\mathrm{d}y$, as stated. Hence, $Q'(t) = 2tG(t)$. In (S1.2), 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 (S1.3) therefore follows. To prove (S1.4), note that, for any $0 < s < t$, $q(t) = q(s) - 2\int_s^t p(y)y^{-1}\mathrm{d}y$. On event $\{|X| > 0\}$, it holds almost surely that

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

By the dominated convergence theorem,

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

Then, in $q(t) = q(s) - 2\int_s^t p(y)y^{-1}\mathrm{d}y$ for all $0 < s < t$, letting s tend to zero yields (S1.4). The monotonicity of Q follows directly from (S1.3) and the limit of $Q(t)$ derives from the monotone convergence theorem. These complete the part (i) of Lemma S1.1.

We now show the remaining properties of function q in the part (ii). By the definition of Δ in (S1.5), we have $\mathbb{P}(0 < |X| \le y) = 0$ and thus $p(y) = 0$ for all

$0 < y < \Delta$. This, together with (S1.4), implies $q(s) = \mathbb{P}(X \neq 0) > 0$ for all $0 \leq s \leq \Delta$. 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

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

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

S1.2 Proof of Proposition 1

To be self-contained, we first formally define the subGaussian estimator as follows. Let X be a real-valued random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{var}(X) > 0$, and assume that $X_1, \ldots, X_n$ are independent and identically distributed $(i.i.d.)$ from X. Given $\alpha \in (0, 1)$, we say $\widehat{\mu}$ (which possibly depends on α) is a subGaussian estimator of μ if it satisfies the bound

$$|\widehat{\mu} - \mu| \leq C\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - \alpha$, where $C > 0$ is an absolute constant. For the sparse linear regression model $Y_i = \boldsymbol{X}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i$ under Condition 3.1 in the main paper, that is, $\mathbb{E}(\varepsilon_i^2 | \boldsymbol{X}_i) = \sigma^2$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\intercal \in \mathbb{R}^d$ is sparse with $\|\boldsymbol{\beta}^*\|_0 := \sum_{j=1}^d I(\beta_j^* \neq 0) = s \ll n$, a subGaussian estimator $\widehat{\boldsymbol{\beta}}$ admits, for $\alpha \in (0, 1)$,

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \lesssim \sigma\sqrt{\frac{s\{\log d + \log(1/\alpha)\}}{n}}$$

with probability $1 - \alpha$.

Proof. Note that the truncated mean m_τ can be written as $m_\tau = (\tau/n)\sum_{i=1}^n \psi_1(X_i/\tau)$, 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

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

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 $\tau = 2v^2n/y$ 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. The above proof essentially follows a similar argument that used in the proof of Propositions 2.1-2.2 in Catoni (2012). □

S1.3 Proof of Proposition 2

Proof of (i). Using the notation in Section S1.1, equation (2.4) can be written as $q(\tau) = z/n$. By Lemma S1.1, 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 (2.4) has a unique solution that lies in (Δ, ∞).

By definition, this unique solution τ_z satisfies

$$\tau_z^2 = \mathbb{E}(X^2 \wedge \tau_z^2)\frac{n}{z} \leq \mathbb{E}(X^2)\frac{n}{z}. \tag{S1.6}$$

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 (S1.6) 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. □

S1.4 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 (S1.1), (S1.2) and (S1.5), respectively. A sample version of Lemma S1.1 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 (2.2) has a unique solution on (Δ_n, ∞) if and only if $z/n < G_n(0)$. □

S1.5 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 S1.1 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

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

$$= 2\int_0^r \frac{p_n(\tau_z+\tau_z u)-p(\tau_z+\tau_z u)}{1+u}\mathrm{d}u + \{q(\tau_z)-q(\tau_z+\tau_z r)\}$$
$$=: R_1 + D_1.$$

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

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

Putting the above calculations together, we arrive at

$$\mathbb{P}(|\widehat{\tau}_z/\tau_z-1|\ge r) \le \mathbb{P}\{q_n(\tau_z)-q(\tau_z)\ge D_1+R_1\} + \mathbb{P}\{q_n(\tau_z)-q(\tau_z)\le -D_2+R_2\}. \tag{S1.7}$$

Set $\zeta_i = (X_i^2\wedge\tau_z^2)/\tau_z^2$ for $i=1,\ldots,n$ such that $q_n(\tau_z)-q(\tau_z) = (1/n)\sum_{i=1}^n\{\zeta_i-\mathbb{E}(\zeta_i)\}$. 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

$$\mathbb{P}\{q_n(\tau_z)-q(\tau_z)\ge u/n\} \le \exp\{-u^2/(2z+2u/3)\}. \tag{S1.8}$$

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 $0<u<z$,

$$\mathbb{P}\{q_n(\tau_z)-q(\tau_z)\le -u/n\} \le \exp\{-u^2/(2z)\}. \tag{S1.9}$$

Step 2 *(Controlling R_1 and R_2).* Note that R_1 and R_2 can be written, respectively,

as $R_1 = (2/n)\sum_{i=1}^n\{\xi_i - \mathbb{E}(\xi_i)\}$ and $R_2 = -(2/n)\sum_{i=1}^n\{\eta_i - \mathbb{E}(\eta_i)\}$, where

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

are bounded, nonnegative random variables satisfying

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

In addition,

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

and

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

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

$$\mathbb{P}(R_1 \le -2rv/n) \le \exp\{-v^2/(2z)\} \tag{S1.10}$$

$$\text{and} \quad \mathbb{P}\{R_2 \ge 2rv/(1-r)n\} \le \exp\{-v^2/(2z)\}. \tag{S1.11}$$

Step 3 *(Bounding D_1 and D_2).* By Lemma S1.1 we have

$$D_1 = q(\tau_z) - q(\tau_z + \tau_z r) = 2\int_{\tau_z}^{\tau_z(1+r)} \frac{P(u)}{u^3}\mathrm{d}u \ge 2P(\tau_z)\int_{\tau_z}^{\tau_z(1+r)} \frac{\mathrm{d}u}{u^3} = \frac{r^2+2r}{(1+r)^2}\frac{P(\tau_z)}{\tau_z^2}. \tag{S1.12}$$

Similarly,

$$D_2 = q(\tau_z - \tau_z r) - q(\tau_z) = 2\int_{\tau_z(1-r)}^{\tau_z} \frac{P(u)}{u^3}\mathrm{d}u \ge \frac{2r-r^2}{(1-r)^2}\frac{P(\tau_z - \tau_z r)}{\tau_z^2}. \tag{S1.13}$$

STEP 4. Together, (S1.7) and (S1.10)-(S1.13) imply that, for any $0 < r < 1$ and $v > 0$,

$$\begin{aligned}
&\mathbb{P}(|\widehat{\tau}_z/\tau_z - 1| \geq r) \\
&\leq \mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \geq \frac{r^2 + 2r}{(1+r)^2}\frac{P(\tau_z)}{\tau_z^2} - \frac{2rv}{n} \right\} \\
&\quad + \mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \leq -\frac{2r - r^2}{(1-r)^2}\frac{P(\tau_z - \tau_z r)}{\tau_z^2} + \frac{2rv}{(1-r)n} \right\} + 2\exp\{-v^2/(2z)\}.
\end{aligned} \tag{S1.14}$$

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 (2.6), the right-hand side of (S1.14) can further be bounded by

$$\mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \geq \frac{a_1 rz}{n} \right\} + \mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \leq -\frac{a_2 rz}{n} \right\} + 2\exp\{-v^2/(2z)\}.$$

Combining this with (S1.8), (S1.9) and (S1.14) proves the stated result. □

S1.6 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 $\tau_{z,k}^* := c_k v_2\sqrt{n/z}$. It follows that

$$\sup_{\tau_z^*/2 \leq \tau \leq 3\tau_z^*/2} |m_\tau - \mu| \leq \max_{1 \leq k \leq n} |m_{\tau_{z,k}^*} - \mu| + \frac{v_2}{\sqrt{nz}}. \tag{S1.15}$$

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

$$\max_{1\le k\le n} |m_{\tau^*_{z,k}} - \mu| \le 4v_2\sqrt{\frac{z}{n}} \tag{S1.16}$$

with probability at least $1 - 2ne^{-z}$. Together, (S1.15) and (S1.16) prove (2.7).

Taking $z = 2\log n$ in Proposition 2, Theorem 1 and Remark 1, we find that $\tau^*_z/2 \le \widehat{\tau}_z \le 3\tau^*_z/2$ with probability at least $1 - 4n^{-c}$ for all sufficiently large n. The desired result then follows from (2.7). □

S1.7 Location and Scale Equivariant of Our Estimator

As noticed in Remark 2 in the main text, our proposed estimator in (2.11) is similar to the estimator discussed by Bickel (1975). It is known that coupling an M-estimation of location with an estimate of scale can lead to scale invariant (Huber and Ronchetti, 2009). In our setting, τ mimics a scale parameter and our proposed estimator is expected to also enjoy the desirable location and scale equivariance. As a matter of fact, for real constants $a \neq 0$ and b, (2.11) is

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \operatorname{sign}(aX_i + b - \theta)\min\{|aX_i + b - \theta|, \tau\} = 0 \\ \displaystyle\frac{1}{n}\sum_{i=1}^{n} \min\{(aX_i + b - \theta)^2, \tau^2\}/\tau^2 - \frac{z}{n} = 0, \end{cases}$$

which is equivalent to

$$\begin{cases} \sum_{i=1}^{n} \text{sign}(X_i - \widetilde{\theta}) \min\{|X_i - \widetilde{\theta}|, \widetilde{\tau}\} = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \min\{(X_i - \widetilde{\theta})^2, \widetilde{\tau}^2\}/\widetilde{\tau}^2 - \frac{z}{n} = 0 \end{cases}$$

by letting $\widetilde{\theta} = a^{-1}(\theta - b)$ and $\widetilde{\tau} = |a|^{-1}\tau$. Recall an estimator T_n is location and scale equivariant if $T_n(aX_1+b, \ldots, aX_n+b) = aT_n(X_1, \ldots, X_n)+b$. Therefore, our estimator for θ is indeed location and scale equivariant.

S1.8 The Influence Function of Our Estimator

Let $\psi(u) = \text{sign}(u)\min(|u|, 1)$ and $\chi(u) = \min(u^2, 1)/2 - z/(2n)$. As discussed in Remark 2 in the main paper and Section S1.7 above, for fixed z, n and $\tau_n > 0$, our estimator defined in (2.11) can be rewritten as the solution to

$$\begin{cases} \sum_{i=1}^{n} \psi\left(\frac{X_i - \theta_n}{\tau_n}\right) = 0 \\ \sum_{i=1}^{n} \chi\left(\frac{X_i - \theta_n}{\tau_n}\right) = 0 \end{cases} \tag{S1.17}$$

which mimics equation (1.6) in Bickel (1975). The estimator in (2.11) can also be viewed as a variant of that characterized by equations (6.28) and (6.29) in Huber and Ronchetti (2009) for joint location and scale estimation.

For $X \sim F$, denote F_n the corresponding empirical cumulative distribution function based on n *i.i.d.* observations. Write estimators $\theta_n = \theta(F_n)$ and $\tau_n = \tau(F_n)$ in terms of statistical functionals $\theta(F)$ and $\tau(F)$, which are defined by $\mathbb{E}_F[\psi(\{X-\theta(F)\}/\tau(F))] = 0$

and (2.9) in the main paper, *i.e.* $\mathbb{E}_F[\chi(\{X-\theta(F)\}/\tau(F))]=0$. Following Hampel et al. (1986) and Huber and Ronchetti (2009), the influence functions $\mathrm{IC}(x;F,\theta)$ and $\mathrm{IC}(x;F,\tau)$ of our estimator (θ_n,τ_n) can be computed by inserting $F_t=(1-t)F+t\delta_x$ for $X_i\sim F$ into the population version of (S1.17) and then taking derivative with respect to t at $t=0$. Specifically, it is easy to see that $\psi(u)$ is odd and $\chi(u)$ is even. Then, assume that F is symmetric for simplicity, we can derive from the results in Huber and Ronchetti (2009) that

$$\mathrm{IC}(x;F,\theta)=\frac{\psi(x/\tau(F))\tau(F)}{\mathbb{E}_F\{\psi'(X_i/\tau(F))\}}=\frac{\mathrm{sign}(x)\min\{|x|,\tau(F)\}}{\mathbb{P}_F\{X_i\le\tau(F)\}}$$

and

$$\mathrm{IC}(x;F,\tau)=\frac{\chi(x/\tau(F))\tau(F)}{\mathbb{E}_F\{\chi'(X_i/\tau(F))X_i/\tau(F)\}}=\frac{\min\{x^2/\tau(F),\tau(F)\}/2-z\tau(F)/(2n)}{\mathbb{E}_F[X_i^2/\tau^2(F)\mathbb{I}\{|X_i|\le\tau(F)\}]}.$$

The gross error sensitive of our estimator $\mathrm{GES}(\tau;F)=\sup_x|\mathrm{IC}(x;F,\tau)|$ is bounded for each fixed n. This reflects the relative influence of individual outlier toward our estimator for finite sample. On the other hand,

$$\mathrm{GES}(\tau;F)\ge\tau(F)\sup_x\frac{|\min\{x^2,\tau^2(F)\}/2-z\tau^2(F)/(2n)|}{\mathbb{E}_F(X_i^2)}\gtrsim\tau^3(F)\left(\frac{1}{2}-\frac{z}{2n}\right).$$

Results from Section 2.1 in the main paper suggest that $\tau(F)\asymp\sqrt{n/z}$ for the adaptive Huber estimator for each $z>0$. Hence, as $n\to\infty$, $\mathrm{GES}(\tau;F)$ diverges uniformly. This unboundedness of GES is due to the divergence of $\tau(F)$ in n, which is designed for our focus on the tail robustness against distributional outliers other than against a minority of data points generated from different distribution that is not the generating process

of interest. The later scenario is typically modeled by the Huber's ϵ-contamination model in the traditional robust statistics. In fact, for the fixed τ such as $\tau = 1.345\sigma$, it guarantees a bounded GES$(\tau; F)$ and a high finite sample breakdown point (Hampel et al., 1986; Maronna et al., 2018).

For data from a heavy-tailed and/or highly-skewed distribution, it is intuitive that the distribution is better represented when more observations are accumulated. Therefore, the divergent $\tau(F)$ will be desirable in our proposal (or (S1.17) above) to reduce the bias, see Proposition 5 for example. This unboundedness also reflect a critical trade-off in designing the robust M-estimators as observed by Loh (2017). In addition, for the adversarially contaminated errors, different scaling of the robustification parameter such as $\sqrt{\log n}$ (Dalalyan and Thompson, 2019) or n^{β} for some small $\beta > 0$ (Yohai and Zamar 1997; (1.38) in Huber and Ronchetti 2009) might provide better results either theoretically or empirically. We leave the study of Huber's M-estimator along with its data-driven tuning for the adversarial contamination as future work.

S2 Proofs of Results in Section 3.1

S2.1 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 $\widetilde{\boldsymbol{\theta}}_\tau^* := (\beta_0^* + \alpha_\tau, \boldsymbol{\beta}^{*\intercal})^\intercal \in \mathbb{R}^{\bar{d}}$. This implies that $G(\beta_0^* + \alpha_\tau, \boldsymbol{\beta}^*) = \min_{\boldsymbol{\theta} \in \mathbb{R}^{\bar{d}}} G(\boldsymbol{\theta})$. Moreover, consider the Hessian matrix $\nabla^2 G(\boldsymbol{\theta}) = \mathbb{E}\{I(|Y - \boldsymbol{Z}^\intercal \boldsymbol{\theta}| \leq \tau) \boldsymbol{Z}\boldsymbol{Z}^\intercal\}$, $\boldsymbol{\theta} \in \mathbb{R}^{\bar{d}}$. By (3.4), $\nabla^2 G(\widetilde{\boldsymbol{\theta}}_\tau^*) = \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau)\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$ 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 (3.5). By the optimality of α_τ and the mean value theorem, we have $h'(\alpha_\tau) = \frac{\mathrm{d}h(\alpha)}{\mathrm{d}\alpha}|_{\alpha=\alpha_\tau} = 0$ and

$$h''(\widetilde{\alpha}_\tau)\alpha_\tau = h'(\alpha_\tau) - h'(0) = -h'(0) = \mathbb{E}\{\ell_\tau'(\varepsilon)\}. \tag{S2.1}$$

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

$$h''(\widetilde{\alpha}_\tau) = 1 - \mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau). \tag{S2.2}$$

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,

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

implying $\mathbb{P}(|\varepsilon - \widetilde{\alpha}_\tau| > \tau) \le \tau^{-2}\sigma^2$. Substituting this into (S2.2) to reach

$$h''(\widetilde{\alpha}_\tau) \ge 1 - \tau^{-2}\sigma^2. \tag{S2.3}$$

For the right-hand side of (S2.1), we have

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

where $\psi_\tau(x) = \ell'_\tau(x)$. Combined with (S2.1) and (S2.3), this proves (3.5) as long as $\tau > \sigma$. □

S2.2 Proof of Theorem 3

The proof is based on a similar argument to that used in the proof of Theorem 2.1 in Zhou et al. (2018). The main improvement comes from Proposition S2.1 below, which provides a form of the restricted strong convexity (RSC) for the empirical loss function. By exploiting the strong convexity of Huber loss and empirical process theory, we establish the RSC property for sub-exponential design under the scaling $n \gtrsim d$.

Write $\bar{d} = d + 1$ throughout the proof. For some $r > 0$ to be determined, define the local neighborhood

$$\mathbf{\Theta}_r = \big\{\boldsymbol{\theta} \in \mathbb{R}^{\bar{d}} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}} \le r\big\}, \tag{S2.4}$$

where $\|\cdot\|_{\mathbf{S}}$ denotes the rescaled ℓ_2-norm $\|\boldsymbol{u}\|_{\mathbf{S}} = \|\mathbf{S}^{1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u} \in \mathbb{R}^{\bar{d}}$. If $\widehat{\boldsymbol{\theta}}_\tau \notin \boldsymbol{\Theta}_r$, there exists some $\eta \in (0,1)$ such that $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \eta(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*) \in \boldsymbol{\Theta}_r$; 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 2 in Sun, Zhou, and Fan (2019) to $\mathcal{L}_\tau(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta})$ gives

$$\begin{aligned}\big\langle \nabla\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \big\rangle &\le \eta\big\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* \big\rangle \\ &= \eta\big\langle -\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* \big\rangle \\ &\le \|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 \times \|\widetilde{\boldsymbol{\theta}}\|_{\mathbf{S}}. \end{aligned} \tag{S2.5}$$

In what follows, we bound the left-hand and right-hand sides of (S2.5) separately, staring with the former. Proposition S2.1 below shows that $\mathcal{L}_\tau$ is strictly convex on $\boldsymbol{\Theta}_r$ with high probability.

Proposition S2.1. Let $m_4 = \sup_{\boldsymbol{u}\in\mathbb{S}^d} \mathbb{E}\langle \mathbf{S}^{-1/2}\boldsymbol{Z}, \boldsymbol{u}\rangle^4$ with $\mathbf{S} = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\intercal)$. Let $\tau, r > 0$ satisfy

$$\tau \ge \max(4\sigma, 8m_4^{1/2}r) \quad \text{and} \quad n \ge c_0(\tau/r)^2(d+z), \tag{S2.6}$$

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

$$\langle \nabla\mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \ge \frac{1}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2 \quad \text{uniformly over} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r. \tag{S2.7}$$

Proof of Theorem 3. Since $\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_r$ by construction, it holds under the scaling (S2.6) that

$$\big\langle \nabla\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \big\rangle \ge \frac{1}{4}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2 \tag{S2.8}$$

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

$$\Bigg\| \underbrace{\frac{1}{n}\sum_{i=1}^n \big\{\xi_i \boldsymbol{z}_i - \mathbb{E}(\xi_i \boldsymbol{z}_i)\big\}}_{:=\boldsymbol{\gamma}} \Bigg\|_2 + \Bigg\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\xi_i \boldsymbol{z}_i) \Bigg\|_2, \tag{S2.9}$$

where $\xi_i = \ell'_\tau(\varepsilon_i)$ and $\boldsymbol{z}_i = \mathbf{S}^{-1/2}\boldsymbol{Z}_i$. 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 $|\mathcal{N}_{1/2}| \le 5^{\bar{d}}$ 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 3.1, we have

$$\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{z}_i|^k = A_0^k k \int_0^\infty t^{k-1}\mathbb{P}(|\boldsymbol{u}^\intercal\boldsymbol{z}_i| \ge A_0 t)\,\mathrm{d}t \le A_0^k k \int_0^\infty t^{k-1}e^{-t}\,\mathrm{d}t = A_0^k k!, \quad k \ge 1,$$

from which it follows that $\mathbb{E}(\xi_i \boldsymbol{u}^\intercal\boldsymbol{z}_i)^2 = \mathbb{E}\{\mathbb{E}(\xi_i^2|\boldsymbol{z}_i)(\boldsymbol{u}^\intercal\boldsymbol{z}_i)^2\} \le \sigma^2$ and

$$\mathbb{E}|\xi_i \boldsymbol{u}^\intercal\boldsymbol{z}_i|^k \le \frac{k!}{2}\, 2A_0^2\sigma^2(A_0\tau)^{k-2} \quad \text{for all } k = 3, 4, \ldots.$$

By Bernstein's inequality,

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

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

$$\mathbb{P}\bigg(\|\boldsymbol{\gamma}\|_2 \ge 4A_0\sigma\sqrt{\frac{2\bar{d}+z}{n}} + 2A_0\tau\frac{2\bar{d}+z}{n}\bigg) \le e^{-z}.$$

For the deterministic part $\|\mathbb{E}(\xi_i \boldsymbol{z}_i)\|_2$ in (S2.9), it holds $\|\mathbb{E}(\xi_i \boldsymbol{z}_i)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^d} \mathbb{E}(\xi_i \boldsymbol{u}^\top \boldsymbol{z}_i) \leq \sigma^2\tau^{-1}$. Putting the above calculations together yields that with probability greater than $1-e^{-z}$,

$$\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 < r_0 := (1+4A_0)\sigma\sqrt{\frac{2\bar{d}+z}{n}} + 2A_0\tau\frac{2\bar{d}+z}{n}. \tag{S2.10}$$

Finally, in view of (S2.8) and (S2.10), we take $r = \tau/(8m_4^{1/2})$. It then follows that with probability greater than $1-2e^{-z}$, $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_{\mathbf{S}} \leq 4\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 < 4r_0$ under the assumed scaling (S2.6). Provided $n \gtrsim d+z$ so that $4r_0 \leq r$, the intermediate estimator $\widetilde{\boldsymbol{\theta}}$ will lie in the interior of $\boldsymbol{\Theta}_r$, which enforces $\eta = 1$ and $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\tau$ (otherwise $\widetilde{\boldsymbol{\theta}}$ will lie on the boundary). Putting together the pieces, we arrive at the desired result. □

S2.3 Proof of Theorem 4

In view of the proof of Theorem 3, lying in the heart of the arguments is the restricted strong convexity (S2.7) and the deviation bound (S2.10) for a random quadratic form. In the following, we will establish similar results to (S2.7) and (S2.10) 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

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

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

We start with the deviation bound. Recalling $\boldsymbol{\theta}^*_\tau = \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^{\bar{d}}} \sum_{i=1}^n \mathbb{E}\ell_\tau(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\theta})$,

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

$$\|\mathbf{S}^{-1/2}\nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i\right\|_2. \tag{S2.11}$$

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

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

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

$$\left\|\frac{1}{n}\sum_{i=1}^n \ell_\tau'(\varepsilon_i - \alpha_\tau)\boldsymbol{z}_i\right\|_2 < 4A_0\sigma\sqrt{\frac{x}{n}} + 2A_0\tau\frac{x}{n}.$$

Taking $x = 2\bar{d} + z$ in the last display, we obtain from (S2.11) that

$$\|\mathbf{S}^{-1/2}\nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 < r_1 := 4A_0\sigma\sqrt{\frac{2\bar{d}+z}{n}} + 2A_0\tau\frac{2\bar{d}+z}{n} \tag{S2.12}$$

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 S2.2. Let $m_4 = \sup_{\boldsymbol{u}\in\mathbb{S}^d} \mathbb{E}\langle \mathbf{S}^{-1/2}\boldsymbol{Z}, \boldsymbol{u}\rangle^4$. Let $\tau, r > 0$ satisfy

$$\tau \ge 8m_4^{1/2}r \quad \text{and} \quad n \ge c_0\rho_\tau^{-2}(\tau/r)^2(d+z), \tag{S2.13}$$

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

$$\langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*), \boldsymbol{\theta} - \boldsymbol{\theta}_\tau^* \rangle \geq \frac{\rho_\tau}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S}}^2 \quad \text{uniformly over} \quad \boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}_r. \tag{S2.14}$$

According to (S2.12) and (S2.14), we take $r = \tau/(8m_4^{1/2})$ so that with probability at least $1 - 2e^{-z}$, $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_\tau^*\|_{\mathbf{S}} \leq 2\rho_\tau^{-1} \|\mathbf{S}^{-1/2} \nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*)\|_2 < 2\rho_\tau^{-1} r_1$ under the scaling (S2.13). Provided $n \gtrsim d + z$ so that $2\rho_\tau^{-1} r_1 \leq r = \tau/(8m_4^{1/2})$, the intermediate estimator $\widetilde{\boldsymbol{\theta}}$ will lie in the interior of $\widetilde{\boldsymbol{\Theta}}_r$, which enforces $\eta = 1$ and $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\tau$, as desired. □

S2.4 Proof of Proposition S2.1

Since the Huber loss is convex and differentiable, we have

$$\begin{aligned}
\mathcal{T}(\boldsymbol{\theta}) &:= \langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \\
&= \frac{1}{n} \sum_{i=1}^n \big\{ \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}^*) - \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}) \big\} \boldsymbol{Z}_i^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\geq \frac{1}{n} \sum_{i=1}^n \big\{ \ell_\tau'(\varepsilon_i) - \ell_\tau'(Y_i - \boldsymbol{Z}_i^\intercal \boldsymbol{\theta}) \big\} \boldsymbol{Z}_i^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}^*) I_{\mathcal{E}_i},
\end{aligned} \tag{S2.15}$$

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

$$\mathcal{E}_i := \big\{ |\varepsilon_i| \leq \tau/2 \big\} \cap \left\{ \frac{|\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}} \leq \frac{\tau}{2r} \right\}, \tag{S2.16}$$

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

$$u^2 I(|u| \le R/2) \le \varphi_R(u) \le u^2 I(|u| \le R). \tag{S2.17}$$

It then follows that

$$\mathcal{T}(\boldsymbol{\theta}) \ge g(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \varphi_{\tau \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}/(2r)}(\langle \boldsymbol{Z}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle) \psi_{\tau/2}(\varepsilon_i). \tag{S2.18}$$

To bound the right-hand side of (S2.18), consider the supremum of a random process indexed by $\boldsymbol{\Theta}_r$:

$$\Delta_r := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_r} \frac{-g(\boldsymbol{\theta}) + \mathbb{E} g(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2}. \tag{S2.19}$$

Write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_r$. By (S2.17),

$$\begin{aligned} \mathbb{E} g(\boldsymbol{\theta}) &\ge \mathbb{E} \langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle^2 - \mathbb{E} \left\{ \langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle^2 I\left(|\langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle| \ge \frac{\tau}{4r} \|\boldsymbol{\delta}\|_{\mathbf{S}} \right) \right\} - \mathbb{E} \left\{ \langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle^2 I(|\varepsilon_i| > \tau/2) \right\} \\ &\ge \|\boldsymbol{\delta}\|_{\mathbf{S}}^2 - \frac{4}{\tau^2} \left(\frac{4r^2}{\|\boldsymbol{\delta}\|_{\mathbf{S}}^2} \mathbb{E} \langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle^4 + \mathbb{E} \langle \boldsymbol{Z}_i, \boldsymbol{\delta} \rangle^2 \varepsilon_i^2 \right). \end{aligned} \tag{S2.20}$$

Recall that $\mathbb{E}\langle \boldsymbol{\delta}, \boldsymbol{Z}_i\rangle^4 \le m_4 \|\boldsymbol{\delta}\|_{\mathbf{S}}^4$ for all $\boldsymbol{u} \in \mathbb{R}^{\bar{d}}$ ($\bar{d} = d+1$). In conjunction with (S2.20), this implies

$$\mathbb{E}g(\boldsymbol{\theta}) \ge \|\boldsymbol{\delta}\|_{\mathbf{S}}^2 - \|\boldsymbol{\delta}\|_{\mathbf{S}}^2\big(\sigma^2 + 4m_4 r^2\big)\frac{4}{\tau^2} \ge \frac{1}{2}\|\boldsymbol{\delta}\|_{\mathbf{S}}^2 \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}_r, \tag{S2.21}$$

where the last inequality holds if $\tau \ge \max(4\sigma, 8m_4^{1/2} r)$. By (S2.18), (S2.19) and (S2.21),

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2} \ge \frac{1}{2} - \Delta_r \quad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r. \tag{S2.22}$$

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

Lemma S2.1. For any $x > 0$,

$$\Delta_r \le 1.25\frac{\tau}{r}\sqrt{\frac{d+1}{n}} + (2m_4)^{1/2}\sqrt{\frac{x}{n}} + \frac{\tau^2}{r^2}\frac{x}{3n}$$

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

Substituting this into Lemma S2.1, we obtain that with probability at least $1 - e^{-z}$,

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

for all sufficiently large n that scales as $(\tau/r)^2(d+z)$ up to an absolute constant. This proves (S2.7). □

S2.5 Proof of Proposition S2.2

Following the proof of Proposition S2.1, now we have

$$\langle \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}_\tau^*), \boldsymbol{\theta} - \boldsymbol{\theta}_\tau^* \rangle$$

$$\geq \frac{1}{n}\sum_{i=1}^{n}\left\{\ell'_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta}^*_\tau) - \ell'_\tau(Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta})\right\}\boldsymbol{Z}_i^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^*_\tau)I_{\mathcal{E}_{\tau,i}},$$

where

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

On $\mathcal{E}_{\tau,i}$, $|Y_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\theta}^*_\tau| = |\varepsilon_i + \beta_0^* - \beta^*_{0,\tau}| = |\varepsilon_i - \alpha_\tau| \leq \tau$ 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 (S2.18) except with $\boldsymbol{\theta}^*$ replaced by $\boldsymbol{\theta}^*_\tau$. By assumption $\rho_\tau := \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau/2) > 0$ and Markov's inequality, we obtain that for every $\boldsymbol{\theta} \in \widetilde{\boldsymbol{\Theta}}_r$,

$$\mathbb{E}g(\boldsymbol{\theta}) \geq \rho_\tau\big(1 - 16m_4 r^2\tau^{-2}\big)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_\tau\|^2_{\mathbf{S}} \geq \frac{3}{4}\rho_\tau\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_\tau\|^2_{\mathbf{S}},$$

provided $0 < r \leq \tau/(8m_4^{1/2})$. Keep all other statements the same, we then get the desired result. □

S2.6 Proof of Lemma S2.1

For $g(\boldsymbol{\theta})$ given in (S2.18), we write $g(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n g_i(\boldsymbol{\theta})$. 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 $0 \leq g_i(\boldsymbol{\theta}) \leq (\tau/4r)^2\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2_{\mathbf{S}}$. It then follows from Theorem 7.3 in Bousquet (2003), a variant of Talagrand's inequality, that for any $x > 0$,

$$\begin{aligned}\Delta_r &\leq \mathbb{E}\Delta_r + (\mathbb{E}\Delta_r)^{1/2}\frac{\tau}{2r}\sqrt{\frac{x}{n}} + (2m_4)^{1/2}\sqrt{\frac{x}{n}} + \frac{\tau^2 x}{48r^2 n} \\ &\leq 1.25\mathbb{E}\Delta_r + (2m_4)^{1/2}\sqrt{\frac{x}{n}} + \frac{\tau^2 x}{3r^2 n} \qquad \text{(S2.23)}\end{aligned}$$

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

It remains to bound the expectation $\mathbb{E}\Delta_r$. By the equality $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for $c, R > 0$ and $u \in \mathbb{R}$, we have

$$\mathcal{E}_i(\boldsymbol{\delta}) := \frac{1}{\|\boldsymbol{\delta}\|_{\mathbf{S}}^2}\varphi_{\tau\|\boldsymbol{\delta}\|_{\mathbf{S}}/(2r)}(\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle)\psi_{\tau/2}(\varepsilon_i) = \varphi_{\tau/(2r)}(\langle \boldsymbol{Z}_i, \boldsymbol{\delta}\rangle)\psi_{\tau/2}(\varepsilon_i), \quad \boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*.$$

Applying the symmetrization inequality for empirical processes yields

$$\mathbb{E}\Delta_r \le 2\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\frac{1}{n}\sum_{i=1}^n e_i\mathcal{E}_i(\boldsymbol{\delta})\right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since φ_R is R-Lipschitz, $\mathcal{E}_i(\boldsymbol{\delta})$ is a $(\tau/2r)$-Lipschitz function in $\langle \boldsymbol{Z}_i, \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\mathbf{S}}\rangle$. By Talagrand's contraction principle (see, e.g. Theorem 4.4, Theorem 4.12 and (4.20) in Ledoux and Talagrand (1991)),

$$\mathbb{E}\Delta_r \le \frac{\tau}{r}\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_r}\frac{1}{n}\sum_{i=1}^n e_i\langle \boldsymbol{Z}_i, \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\mathbf{S}}\rangle\right\} \le \frac{\tau}{r}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n e_i\,\mathbf{S}^{-1/2}\boldsymbol{Z}_i\right\|_2 \le \frac{\tau}{r}\sqrt{\frac{d+1}{n}}.$$

This, together with (S2.23), proves the stated result. □

S3 Proof of Theorem 5

This proof is based on an argument similar to that used in the proof of Theorem 3. Note that Theorem 3 in Fan, Li, and Wang (2017) does not cover the sparse case where $q = 0$. The main reason is that the sparsity property of $\boldsymbol{\beta}^*$ is not inherited by $\boldsymbol{\beta}_\tau^*$. In this proof, we follow a different route to directly establish the convergence around $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_\tau^*$. For simplicity, we write $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^\intercal)^\intercal = \widehat{\boldsymbol{\theta}}_{\mathrm{H}}(\tau, \lambda) \in \mathbb{R}^{\bar{d}}$ with $\bar{d} = d + 1$. For some $r > 0$ to be specified, we use $\boldsymbol{\Theta}_r$ and $\|\cdot\|_{\mathbf{S}}$ to denote the local neighborhood

and rescaled ℓ_2-norm as in (S2.4). 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

$$\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^* \rangle \leq \eta \langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle. \tag{S3.1}$$

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 ℓ_1-cone $\mathcal{C} \subseteq \mathbb{R}^{\bar{d}}$ as

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

It can be shown that the optimal solution $\widehat{\boldsymbol{\theta}}$ to program (3.8) satisfies

$$\widehat{\boldsymbol{\theta}} \in \mathcal{C} \ \text{ on the event } \ \big\{ \lambda \geq 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty \big\}, \tag{S3.2}$$

from which it follows

$$\|\widehat{\boldsymbol{\delta}}\|_1 = |\widehat{v}_0| + \|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 + \|\widehat{\boldsymbol{v}}_{\mathcal{S}^c}\|_1 \leq 2|\widehat{v}_0| + 4\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 \leq 4(s+1)^{1/2}\|\widehat{\boldsymbol{\delta}}\|_2. \tag{S3.3}$$

By necessary conditions of the optima for convex problem (3.8),

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

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

$$\begin{aligned}
&\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle \\
&\leq \lambda\big(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \|\mathbf{S}^{-1/2}\mathbb{E}\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathbf{S}}
\end{aligned}$$

$$\leq \lambda\big(\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \|\mathbf{S}^{-1/2}\mathbb{E}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathbf{S}}$$
$$\leq \frac{\lambda}{2}\big(3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) + \frac{\lambda}{2}|\widehat{v}_0| + \frac{\sigma^2}{\tau}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathbf{S}},$$

where the last step uses the bound $\|\mathbf{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}\{\ell_\tau'(\varepsilon)\langle\boldsymbol{u}, \mathbf{S}^{-1/2}\boldsymbol{Z}\rangle\} \leq \sigma^2\tau^{-1}$. Together with (S3.1), this implies

$$\langle\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^*\rangle$$
$$\leq \frac{1}{2}\lambda\eta\big(3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{v}}_{\mathcal{S}^{\mathrm{c}}}\|_1\big) + \frac{1}{2}\lambda\eta|\widehat{v}_0| + \frac{\sigma^2}{\tau}\|\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}^*\|_{\mathbf{S}}. \tag{S3.4}$$

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 (S3.2), we also have $\widehat{\boldsymbol{\theta}}_\eta \in \mathcal{C}$ under the assumed scaling.

To bound the left-hand side of (S3.4), 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 S2.1. We leave the proof to Section S3.2.

Proposition S3.1. Let $m_4 = \sup_{\boldsymbol{u}\in\mathbb{S}^d}\mathbb{E}\langle\mathbf{S}^{-1/2}\boldsymbol{Z}, \boldsymbol{u}\rangle^4$. Let $\tau, r > 0$ satisfy

$$\tau \geq \max(4\sigma, 8m_4^{1/2}r) \quad \text{and} \quad n \geq c_0\underline{\lambda}_{\mathbf{S}}^{-1}\max_{1\leq j\leq d}\sigma_{jj}(A_0\tau/r)^2 s\log d, \tag{S3.5}$$

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

$$\langle\nabla\mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \geq \frac{1}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2 \ \text{ uniformly over } \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}. \tag{S3.6}$$

Let Ω_r be the event on which (S3.6) holds. Then $\mathbb{P}(\Omega_r^{\mathrm{c}}) \leq d^{-1}$ under the scaling (S3.5) and it holds on $\Omega_r \cap \{\lambda \geq 2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathbb{E}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty\}$ that

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

Substituting this lower bound into (S3.4) yields

$$\begin{aligned}\frac{1}{4}\|\widehat{\boldsymbol{\delta}}_\eta\|_{\mathbf{S}}^2 &\le \frac{1}{2}\lambda\eta\big(|\widehat{v}_0| + 3\|\widehat{\boldsymbol{v}}_{\mathcal{S}}\|_1\big) + \frac{\sigma^2}{\tau}\|\widehat{\boldsymbol{\delta}}_\eta\|_{\mathbf{S}} \\ &\le \frac{3}{2}(s+1)^{1/2}\lambda\|\eta\widehat{\boldsymbol{\delta}}\|_2 + \frac{\sigma^2}{\tau}\|\widehat{\boldsymbol{\delta}}_\eta\|_{\mathbf{S}} \\ &= \frac{3}{2}(s+1)^{1/2}\lambda\|\widehat{\boldsymbol{\delta}}_\eta\|_2 + \frac{\sigma^2}{\tau}\|\widehat{\boldsymbol{\delta}}_\eta\|_{\mathbf{S}}.\end{aligned}$$

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

$$\|\widehat{\boldsymbol{\delta}}_\eta\|_{\mathbf{S}} \le 6\underline{\lambda}_{\mathbf{S}}^{-1/2}(s+1)^{1/2}\lambda + 4\sigma^2\tau^{-1} \tag{S3.7}$$

$$\text{and} \quad \|\widehat{\boldsymbol{\delta}}_\eta\|_1 \le 24\underline{\lambda}_{\mathbf{S}}^{-1}(s+1)\lambda + 16\underline{\lambda}_{\mathbf{S}}^{-1/2}\sigma^2(s+1)^{1/2}\tau^{-1}$$

under the assumed scaling $\lambda \ge 2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$ and (S3.5) .

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

Proposition S3.2. Assume Condition 3.1 holds and let $\tau = \sigma\sqrt{n/t}$ for some $t > 0$. Then with probability at least $1 - 2d^{-1}$,

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty \le 2A_0 \max_{0\le j\le d}\sigma_{jj}^{1/2}\sigma\left(\sqrt{\frac{2\log\bar{d}}{n}} + \frac{\log\bar{d}}{\sqrt{nt}}\right). \tag{S3.8}$$

Applying Proposition S3.2 with $t = \log d$, we find that

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty \le c_1 A_0 \max_{0\le j\le d}\sigma_{jj}^{1/2}\sigma\sqrt{\frac{\log d}{n}}$$

with probability at least $1 - 2d^{-1}$, where $c_1 > 0$ is an absolute constant. We therefore choose $\lambda = c_2 A_0 \max_{0\le j\le d}\sigma_{jj}^{1/2}\sigma\sqrt{(\log d)/n}$ for some constant $c_2 \ge 2c_1$, such that $\lambda \ge 2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty$ with high probability.

Putting the above calculations together and taking $r = \sigma$, we conclude that

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

with probability at least $1-3d^{-1}$, assuming the scaling $n \gtrsim \underline{\lambda}_{\mathbf{S}}^{-1} A_0^2 m_4 \max_{0\le j\le d} \sigma_{jj}\, s \log d$. 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 (3.9) then follows from (S3.7). □

S3.1 Proof of (S3.2)

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. □

S3.2 Proof of Proposition S3.1

The proof is almost identical to that of Proposition S2.1. With slight abuse of notation, define the supremum of a random process indexed by $\boldsymbol{\Theta}_r \cap \mathcal{C}$:

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

Provided $\tau \geq \max(4\sigma, 8m_4^{1/2} r)$, it can be shown that

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathbf{S}}^2} \geq \frac{1}{2} - \Delta_r \quad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_r \cap \mathcal{C}. \tag{S3.9}$$

The following lemma is a slight modification of Lemma S2.1.

Lemma S3.1. For any $x > 0$,

$$\Delta_r \leq 10\sqrt{2}\underline{\lambda}_{\mathbf{S}}^{-1/2} A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2} \bar{s}^{1/2} \frac{\tau}{r} \left\{ \sqrt{\frac{\log(2\bar{d})}{n}} + \frac{\log(2\bar{d})}{n} \right\} + (2m_4)^{1/2} \sqrt{\frac{x}{n}} + \frac{\tau^2}{r^2} \frac{x}{3n}$$

with probability at least $1 - e^{-x}$, where $\bar{d} = d + 1$ and $\bar{s} = s + 1$.

Taking $x = \log d$ in Lemma S3.1, we obtain that with probability at least $1 - d^{-1}$,

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

as long as $n \gtrsim \underline{\lambda}_{\mathbf{S}}^{-1} \max_{1 \leq j \leq d} \sigma_{jj} (A_0 \tau / r)^2 s \log d$. This proves (S3.6). □

S3.3 Proof of Proposition S3.2

To begin with, recall that $\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*) = (1/n) \sum_{i=1}^n \xi_i \boldsymbol{Z}_i$, where $\xi_i = \ell_\tau'(\varepsilon_i)$ and $\boldsymbol{Z}_i = (X_{i0}, X_{i1}, \ldots, X_{id})^\intercal$ with $X_{i0} \equiv 1$. Hence,

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*) - \nabla \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\theta}^*)\|_\infty = \max_{0 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^n \xi_i X_{ij} - \mathbb{E}(\xi_i X_{ij}) \right|.$$

Next we use the union bound and Bernstein's inequality to bound the maximum. For every $0 \le j \le d$,

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

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

Moreover, for $k = 3, 4, \ldots$, $\mathbb{E}|\xi_i X_{ij}|^k \le \sigma^2\tau^{k-2}A_0^k\sigma_{jj}^{k/2}k! = \frac{k!}{2}2A_0^2\sigma_{jj}\sigma^2(A_0\sigma_{jj}^{1/2}\tau)^{k-2}$. Then it follows from Bernstein's inequality that, for any $x > 0$,

$$\left|\frac{1}{n}\sum_{i=1}^n(\xi_i X_{ij} - \mathbb{E}\xi_i X_{ij})\right| \le 2A_0\sigma_{jj}^{1/2}\sigma\sqrt{\frac{x}{n}} + A_0\sigma_{jj}^{1/2}\tau\frac{x}{n}$$

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

S3.4 Proof of Lemma S3.1

Following the proof of Lemma S2.1, we only need to bound the expectation $\mathbb{E}\Delta_r$. By Rademacher symmetrization and Talagrand's contraction principle, it suffices to bound

$$\mathbb{E}\left\{\sup_{\boldsymbol{\delta}=\boldsymbol{\theta}-\boldsymbol{\theta}^*:\boldsymbol{\theta}\in\Theta_r\cap\mathcal{C}}\frac{1}{n}\sum_{i=1}^n e_i\langle\boldsymbol{Z}_i, \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\mathbf{S}}\rangle\right\} \le 4\underline{\lambda}_{\mathbf{S}}^{-1/2}(s+1)^{1/2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{Z}_i\right\|_\infty,$$

where $\boldsymbol{Z}_i = (X_{i0}, X_{i1}, \ldots, X_{id})^\intercal$ with $X_{i0} \equiv 1$. For $j = 1, \ldots, d$, define $S_j = \sum_{i=1}^n e_i X_{ij}$ and note that $\mathbb{E}(e_i X_{ij}) = 0$ and $\mathbb{E}(e_i X_{ij})^2 = \sigma_{jj}$. For $k = 3, 4, \ldots$,

$$\mathbb{E}|e_i X_{ij}|^k = A_0^k\sigma_{jj}^{k/2}k\int_0^\infty t^{k-1}\mathbb{P}(|X_{ij}| \ge A_0\sigma_{jj}^{1/2}t)\,\mathrm{d}t \le k!A_0^k\sigma_{jj}^{k/2}.$$

The above moment inequalities, together with the symmetry of Rademacher random variable, yield

$$\begin{aligned}
\mathbb{E}e^{\lambda e_i X_{ij}} &= 1 + \frac{1}{2}\sigma_{jj}\lambda^2 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(e_i X_{ij})^k \\
&\leq 1 + \frac{1}{2}\sigma_{jj}\lambda^2 + \sum_{\ell=2}^{\infty} \frac{\lambda^{2\ell}}{(2\ell)!}(2\ell)! A_0^{2\ell}\sigma_{jj}^{\ell} \\
&= 1 + \frac{A_0^2\sigma_{jj}}{2}\sum_{k=2}^{\infty}\lambda^k(\sqrt{2}A_0\sigma_{jj}^{1/2})^{k-2} \\
&\leq 1 + \frac{1}{2}\frac{A_0^2\sigma_{jj}\lambda^2}{1-\sqrt{2}A_0\sigma_{jj}^{1/2}\lambda}, \quad \text{for all} \;\; 0 \leq \lambda < \frac{1}{\sqrt{2}A_0\sigma_{jj}^{1/2}}.
\end{aligned}$$

Let $v = A_0^2 \max_{0\leq j\leq d}\sigma_{jj} n$ and $c = \sqrt{2}A_0 \max_{0\leq j\leq d}\sigma_{jj}^{1/2}$. Following the proof of Theorems 2.10 and 2.5 in Boucheron, Lugosi, and Massart (2013), it can be shown that $\log \mathbb{E}e^{\lambda S_j} \leq \psi(\lambda) := \frac{v\lambda^2}{2(1-c\lambda)}$ for any $0 \leq j \leq d$ and $\lambda \in (0, 1/c)$, and hence

$$\mathbb{E}\max_{0\leq j\leq d}|S_j| \leq \inf_{\lambda\in(0,1/c)}\left\{\frac{\log(2\bar{d}) + \psi(\lambda)}{\lambda}\right\} = \sqrt{2v\log(2\bar{d})} + c\log(2\bar{d}).$$

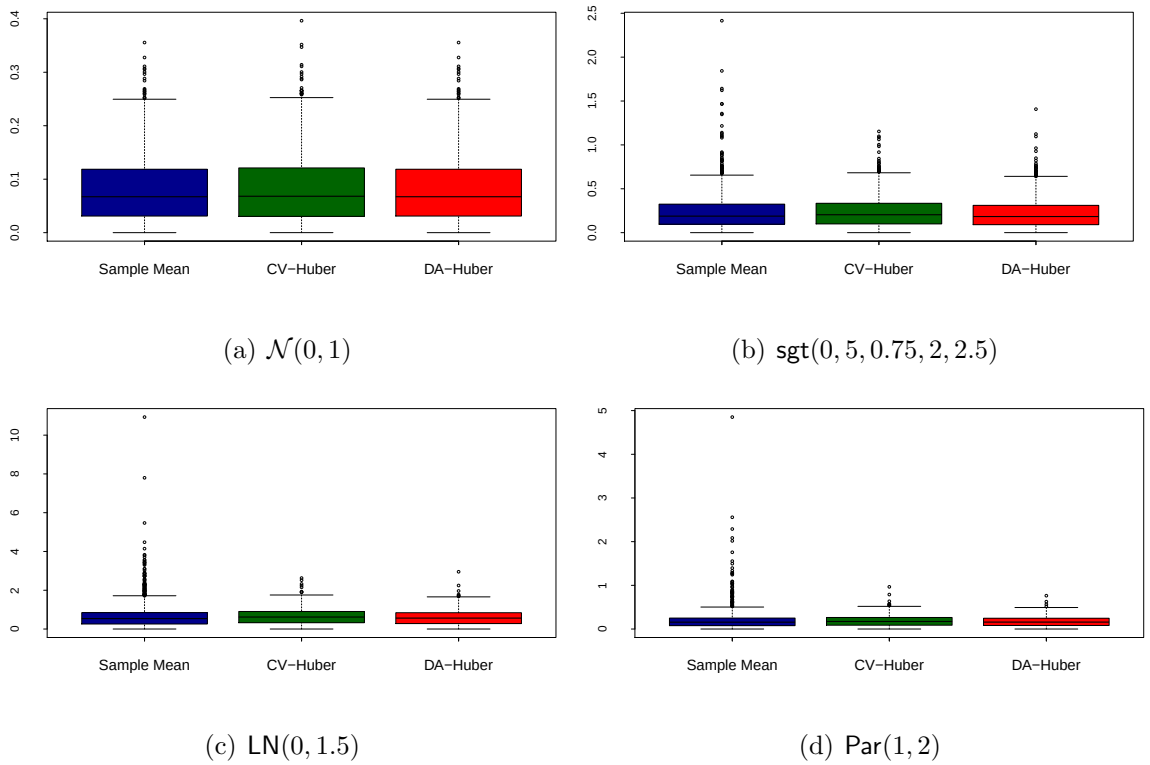
Putting together the pieces, we conclude that

$$\mathbb{E}\Delta_r \leq 8\sqrt{2}\underline{\lambda}_{\mathbf{S}}^{-1/2}A_0 \max_{0\leq j\leq d}\sigma_{jj}^{1/2}(s+1)^{1/2}\frac{\tau}{r}\left\{\sqrt{\frac{\log(2\bar{d})}{n}} + \frac{\log(2\bar{d})}{n}\right\}.$$

In conjunction with the concentration bound (S2.23), this proves the claimed result. □

S4 Additional Simulation Studies

Additional results from the numerical studies in Sections 4.1 and 4.3 are displayed in Figures S1 and S2.



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

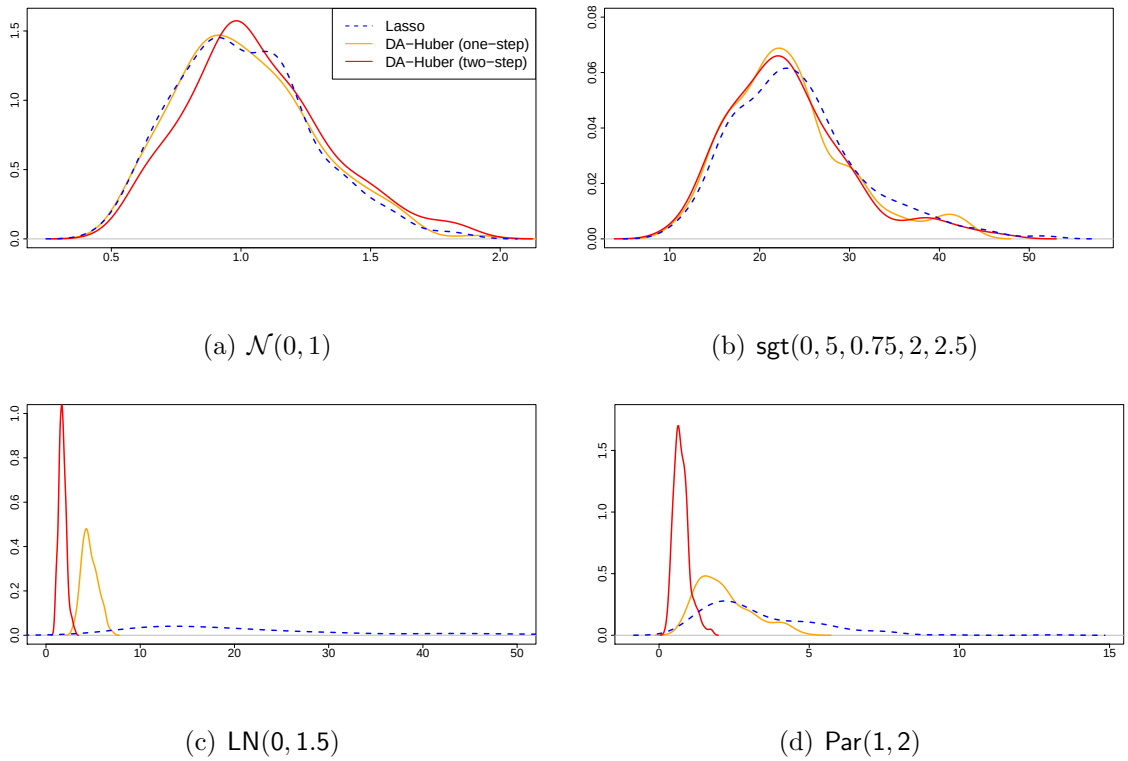
(c) $\mathsf{LN}(0,1.5)$ (d) $\mathsf{Par}(1,2)$

Figure S1: Estimation errors for the sample mean, CV-Huber, and DA-Huber estimators under different settings based on 2000 simulations.

S5 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



(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 S2: Distributions of the ℓ_2-errors of the Lasso and DA-Huber estimators.

adjusted. Linear model (3.1) in the main paper 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.

From the simulation studies in Section 4 of the main text 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

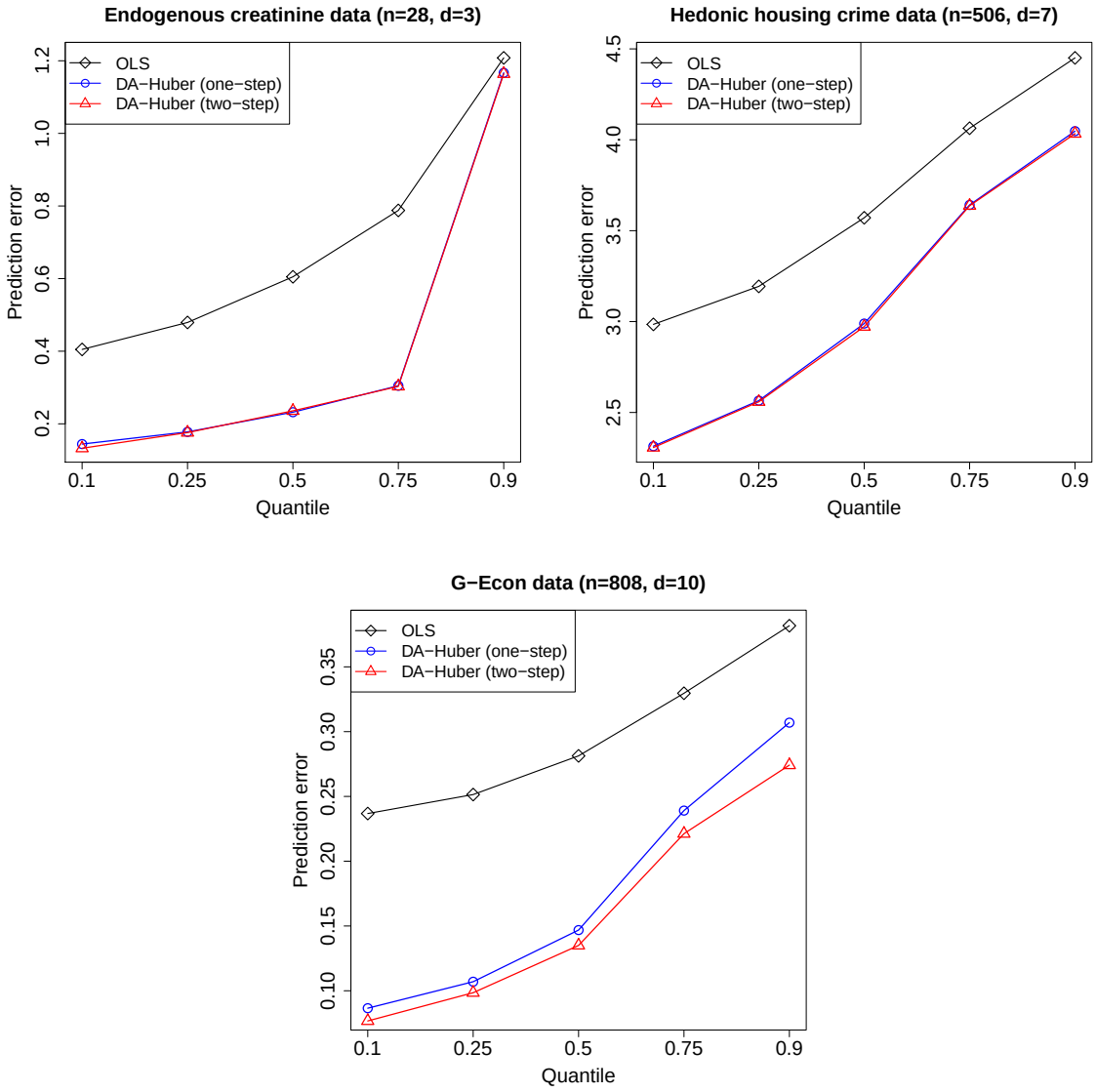


Figure S3: Comparison of the quantiles of mean absolute prediction errors for the OLS (black diamonds), one-step DA-Huber (blue circles), and two-step DA-Huber (red triangles). The results are based on 100 random splittings.

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 S3 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.

S6 Other Loss Functions

The analysis in the paper can be extended to a broader class of robust convex loss functions that include the Huber loss as a prototype. The key to achieve the tail-robustness is the global Lipschitz and local quadratic geometry of the loss function $\ell_\tau(x) = \tau^2\ell(x/\tau)$. The "mother" function $\ell : \mathbf{R} \to [0, \infty)$ satisfies the following condi-

tions with some constants c_1-c_4:

1. $\ell'(0) = 0$ and $|\ell'(x)| \leq c_1$ for all $x \in \mathbf{R}$;

2. $\ell''(0) = 1$ and $|\ell''(x)| \geq c_2$ for all $|x| \leq c_3$;

3. $|\ell'(x) - x| \leq c_4 x^2$ for all $x \in \mathbf{R}$.

Below we list a few examples of ℓ that satisfy conditions (1)-(3).

1. Huber Loss: $\ell(x) = x^2/2 \cdot I(|x| \leq 1) + (|x| - 1/2) \cdot I(|x| > 1)$ with $\ell'(x) = xI(|x| \leq 1) + \text{sign}(x)I(|x| > 1)$ and $\ell''(x) = I(|x| \leq 1)$. Moreover,

$$|\ell'(x) - x| = |x - \text{sign}(x)|I(|x| > 1) \leq x^2.$$

2. Pseudo-Huber loss I : $\ell(x) = \sqrt{1 + x^2} - 1$, whose first and second derivatives are

$$\ell'(x) = \frac{x}{\sqrt{1 + x^2}} \text{ and } \ell''(x) = \frac{1}{(1 + x^2)^{3/2}}$$

respectively.

3. Pseudo-Huber loss II: $\ell(x) = \log\left\{\left(e^x + e^{-x}\right)/2\right\}$, whose first and second derivatives are, respectively,

$$\ell'(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \text{ and } \ell''(x) = \frac{4}{\left(e^x + e^{-x}\right)^2}.$$

4. Smoothed Huber loss I: The Huber loss is twice differentiable in $\mathcal{R}$, except at ± 1. Modifying the Huber loss gives rise to the following function that is twice

differentiable everywhere:

$$\ell(x) = \begin{cases} x^2/2 - |x|^3/6 & \text{if } |x| \leq 1 \\ |x|/2 - 1/6 & \text{if } |x| > 1, \end{cases}$$

whose first and second derivatives are

$$\ell'(x) = \begin{cases} x - \text{sign}(x) \cdot x^2/2 & \text{if } |x| \leq 1 \\ \text{sign}(x)/2 & \text{if } |x| > 1, \end{cases} \qquad \ell''(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases}$$

5. Smoothed Huber loss II: Another smoothed version of the Huber loss function is

$$\ell(x) = \begin{cases} x^2/2 - x^4/24 & \text{if } |x| \leq \sqrt{2} \\ (2\sqrt{2}/3)|x| - 1/2 & \text{if } |x| > \sqrt{2}. \end{cases}$$

References

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, **70**(350), 428–434.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, London.

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

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 48(4), 1148–1185.

Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems.* **32** 13188–13198.

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

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**(1), 247–265.

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**(2), 866–896.

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.

Hahn, M. G., Kuelbs, J. and Weiner, D. C. (1990). The asymptotic joint distribu-

tion of self-normalized censored sums and sums of squares. *The Annals of Probability*, 18(3), 1284–1341.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, Second Edition. Wiley, New York.

Maronna, R. A., Marton, R. D., Yohai, V. J., and Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods (with R)*. Wiley, New York.

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.

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

Yohai, Y.J. and Zamar, R.H. (1997) Optimal locally robust M-estimates of regression. *Journal of Statistical Planning and Inference.* **64**(2) 309–323.

Sun, Q., Zhou, W.-X. and Fan, J. (2019). Adaptive Huber regression. *Journal of the American Statistical Association*, in press. *arXiv preprint arXiv:1706.06991.*

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**(5), 1904–1931.