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 the theoretical results in the main text and 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 3.3. 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

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

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

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

By definition, it is straightforward that $Q(t) = P(t) + t^2G(t)$ 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

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

and

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

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,

$$q(s) = \mathbb{P}(X \neq 0) \quad \text{for all} \quad 0 \leq s \leq \Delta := \inf\{y > 0 : G(y) < \mathbb{P}(X \neq 0)\}, \quad (S1.5)$$
Proof of Lemma S1.1. Notice $Q(t) = \mathbb{E}\{(|X| \wedge t)^2\}$ and it holds almost surely that

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

$$
= 2 \int_0^t I(|X| > t) y \, dy + 2 \int_0^t I(|X| > y) I(|X| \leq t) y \, dy
$$

$$
= 2 \int_0^t I(|X| > y) y \, dy.
$$

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 (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 $0 < s < t$,

$$
q(t) = q(s) - 2 \int_s^t p(y)y^{-1} \, dy.
$$

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,

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

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 (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 $\Delta$ in (S1.5), we have $\mathbb{P}(0 < |X| \leq y) = 0$ and thus $p(y) = 0$ for all

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

q(s) decreases strictly and continuously on $(\Delta, \infty)$, 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

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

$$
= 2 \int_0^t I(|X| > t) y \, dy + 2 \int_0^t I(|X| > y) I(|X| \leq t) y \, dy
$$

$$
= 2 \int_0^t I(|X| > y) y \, dy.
$$

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 (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 $0 < s < t$,

$$
q(t) = q(s) - 2 \int_s^t p(y)y^{-1} \, dy.
$$

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,

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

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 (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 $\Delta$ in (S1.5), we have $\mathbb{P}(0 < |X| \leq y) = 0$ and thus $p(y) = 0$ for all
0 < y < Δ. 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 \((\Delta, \infty)\). Finally, note that

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

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

\[\square\]

**S1.2 Proof of Proposition 2.1**

To be self-contained, we first formally define the sub-Gaussian estimator as follows.

Let \( X \) be a real-valued random variable with mean \( \mu = \mathbb{E}(X) \) and variance \( \sigma^2 = \text{var}(X) > 0 \), 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 \( \hat{\mu} \) (which possibly depends on \( \alpha \)) is a sub-Gaussian estimator of \( \mu \) if it satisfies the bound

\[
|\hat{\mu} - \mu| \leq C \sqrt{\frac{\log(1/\delta)}{n}}
\]

with probability at least \( 1 - \alpha \), where \( C > 0 \) is an absolute constant. For the sparse linear regression model \( Y_i = X_i^\top \beta^* + \varepsilon_i \) under Condition 3.1 in the main paper, that is, \( \mathbb{E}(\varepsilon_i^2 | X_i) = \sigma^2 \) and \( \beta^* = (\beta^*_1, \ldots, \beta^*_d)^\top \in \mathbb{R}^d \) is sparse with \( \|\beta^*\|_0 := \sum_{j=1}^d I(\beta^*_j \neq 0) = s \ll n \), a sub-Gaussian estimator \( \hat{\beta} \) admits, for \( \alpha \in (0, 1) \),

\[
\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma \sqrt{\frac{s \{ \log d + \log(1/\alpha) \}}{n}}
\]

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

$$
\mathbb{P} \left[ \sum_{i=1}^n \{ \tau \psi_1(X_i/\tau) - \mu \} \geq y \right] \leq \exp \left\{ -(y + n\mu)/\tau \right\} \mathbb{E} \left[ \exp \left\{ \sum_{i=1}^n \psi_1(X_i/\tau) \right\} \right] 
$$

$$
= \exp \left\{ -(y + n\mu)/\tau \right\} \prod_{i=1}^n \mathbb{E} \exp \left\{ \psi_1(X_i/\tau) \right\} 
$$

$$
\leq \exp \left\{ -(y + n\mu)/\tau \right\} \prod_{i=1}^n \mathbb{E} \exp \left\{ \log(1 + X_i/\tau + X_i^2/\tau^2) \right\} 
$$

$$
\leq \exp \left\{ -(y + n\mu)/\tau \right\} \prod_{i=1}^n \exp \left\{ \mu/\tau + \mathbb{E}(X_i^2)/\tau^2 \right\} 
$$

$$
\leq \exp \left\{ -(y/\tau + nv^2/\tau^2) \right\} 
$$

$$
= \exp \left\{ nv^2 \left( \frac{1}{\tau} - \frac{y}{2nv^2} \right)^2 - \frac{y^2}{4nv^2} \right\}. 
$$

Similarly,

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

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

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

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 2.2

Proof of (i). Using the notation in Section S1.1, equation (2.4) can be written as

\[ q(\tau) = \frac{z}{n}. \]

By Lemma S1.1, the function \( q \) satisfies \( \max_{t \geq 0} q(t) = \lim_{t \to 0} q(t) = P(|X| > 0), \lim_{t \to \infty} q(t) = 0 \) and is strictly decreasing on \((\Delta, \infty)\). Provided \( \frac{z}{n} < P(|X| > 0) \), equation (2.4) has a unique solution that lies in \((\Delta, \infty)\).

By definition, this unique solution \( \tau_z \) 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 \( \mathbb{P}(|X| > \tau_z) \leq \frac{z}{n} \), which implies \( \tau_z \geq q_z/n \). Substituting this into (S1.6) gives

\[
\tau_z^2 \geq \mathbb{E}(X^2 \wedge q_z^2/n)(n/z).
\]

Proof of (ii). Recall that \( q(\tau) = \frac{z}{n} \). Since \( \frac{z}{n} \to 0 \) 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. \( \square \)

S1.4  Proof of Proposition 2.3

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 \( \Delta \) 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 $\(\Delta_n, \infty\)$. Therefore, equation (2.2) has a unique solution on $\(\Delta_n, \infty\)$ if and only if $\frac{z}{n} < G_n(0).$

\[
\Box
\]

S1.5 Proof of Theorem 2.1

Keep the notation used in the proof of Proposition 2.3. Recall that $\hat{\tau}_z$ 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 $\hat{\tau}_z$ is positive with probability one. We divide the rest of the proof into four steps.

**Step 1 (Preliminaries).** Define the function

\[
p_n(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i^2 I(|X_i| \leq t)}{t^2}
\]

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

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

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

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

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

\[
q_n(\tau_z) - q(\tau_z) \geq 2 \int_0^{r} \frac{p_n(\tau_z + \tau_z u)}{1 + u} du
\]

\[
= 2 \int_0^{r} \frac{p_n(\tau_z + \tau_z u) - p(\tau_z + \tau_z u)}{1 + u} du + 2 \int_0^{r} \frac{p(\tau_z + \tau_z u)}{1 + u} du
\]
\[ = 2 \int_0^r \frac{p_n(\tau_z + \tau_z u) - p(\tau_z + \tau_z u)}{1 + u} \, du + \{q(\tau_z) - q(\tau_z + \tau_z r)\} \]
\[ =: R_1 + D_1. \]

Similarly, on the event \( \{ (\tau_z - \tau_z)/\tau_z \leq -r \} \), it holds
\[ q_n(\tau_z) - q(\tau_z) \]
\[ \leq - \{q(\tau_z - \tau_z r) - q(\tau_z)\} - 2 \int_{-r}^0 \frac{p_n(\tau_z + \tau_z u) - p(\tau_z + \tau_z u)}{1 + u} \, du \]
\[ =: -D_2 + R_2. \]

Putting the above calculations together, we arrive at
\[ \mathbb{P}(|\tau_z/\tau_z - 1| \geq r) \leq \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \geq D_1 + R_1\} + \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \leq -D_2 + R_2\}. \]  
\[ \text{(S1.7)} \]

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 \( \zeta_i \)’s are bounded random variables satisfying \( 0 \leq \zeta_i \leq \min\{1, (|X_i| \wedge \tau_z)/\tau_z\} \) and \( \mathbb{E}(\zeta_i^2) \leq \mathbb{E}(X_i^2 \wedge \tau_z^2)/\tau_z^2 = z/n. \) By Bernstein’s inequality, for any \( u > 0 \) it holds
\[ \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \geq u/n\} \leq \exp\{-u^2/(2z + 2u/3)\}. \]  
\[ \text{(S1.8)} \]

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 \( 0 < u < z, \)
\[ \mathbb{P}\{q_n(\tau_z) - q(\tau_z) \leq -u/n\} \leq \exp\{-u^2/(2z)\}. \]  
\[ \text{(S1.9)} \]

**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

\[
\xi_i = \int_{0}^{r} \frac{X_i^2 I\{ |X_i| \leq \tau_z (1 + u) \}}{\tau_z^2 (1 + u)^3} \, du \quad \text{and} \quad \eta_i = \int_{-r}^{0} \frac{X_i^2 I\{ |X_i| \leq \tau_z (1 + u) \}}{\tau_z^2 (1 + u)^3} \, du
\]

are bounded, nonnegative random variables satisfying

\[
\xi_i \leq \int_{0}^{r} \frac{du}{1 + u} \leq r, \quad \eta_i \leq \int_{-r}^{0} \frac{du}{1 + u} \leq \frac{r}{1 - r}.
\]

In addition,

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

and

\[
\mathbb{E}(\eta_i^2) \leq \mathbb{E}\left\{ \frac{X_i^2 I\{ |X_i| \leq \tau_z \}}{\tau_z^2} \right\} \leq \frac{q(\tau_z) r^2}{(1 - r)^2}.
\]

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

\[
P(R_1 \leq -2rv/n) \leq \exp\{-v^2/(2z)\} \tag{S1.10}
\]

and

\[
P(R_2 \geq 2rv/(1 - r)n) \leq \exp\{-v^2/(2z)\}. \tag{S1.11}
\]

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

\[
D_1 = q(\tau_z) - q(\tau_z + \tau_z r) = 2 \int_{\tau_z}^{\tau_z (1+r)} \frac{P(u)}{u^3} \, du \geq 2P(\tau_z) \int_{\tau_z}^{\tau_z (1+r)} \frac{du}{u^3} = \frac{r^2 + 2r}{(1 + r)^2} \frac{P(\tau_z)}{\tau_z^2}. \tag{S1.12}
\]

Similarly,

\[
D_2 = q(\tau_z - \tau_z r) - q(\tau_z) = 2 \int_{\tau_z (1-r)}^{\tau_z} \frac{P(u)}{u^3} \, du \geq \frac{2r - r^2}{(1 - r)^2} \frac{P(\tau_z - \tau_z r)}{\tau_z^2}. \tag{S1.13}
\]
STEP 4. Together, (S1.7) and (S1.10)-(S1.13) imply that, for any $0 < r < 1$ and $v > 0$,

$$
\mathbb{P}(|\hat{\tau}_z/\tau_z - 1| \geq r) \\
\leq \mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \geq \frac{r^2 + 2rP(\tau_z)}{(1 + r)^2} \cdot \frac{2rv}{n} \right\} \\
+ \mathbb{P}\left\{ q_n(\tau_z) - q(\tau_z) \leq -\frac{2r - r^2}{(1 - r)^2} \cdot \frac{P(\tau_z - \tau_z r)}{\tau_z^2} + \frac{2rv}{(1 - r)n} \right\} + 2 \exp\{-v^2/(2z)\}. \tag{S1.14}
$$

Note that

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

and

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

Taking $v = (a_1 \land a_2)z/2$ for $a_1$ and $a_2$ as in (2.6), the right-hand side of (S1.14) can further be bounded by

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

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

S1.6 Proof of Theorem 2.2

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

$$
\sup_{\tau_z^*/2 \leq \tau \leq 3\tau_z^*/2} |m_\tau - \mu| \leq \max_{1 \leq k \leq n} |m_{\tau_{z,k}^*} - \mu| + \frac{v_2}{\sqrt{n}z}. \tag{S1.15}
$$
For $1 \leq k < n/2$ so that $1/2 \leq c_k < 1$, by Proposition 2.1-(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 2.1-(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

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

(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 2.2, Theorem 2.1 and Remark 2.1, we find that $\tau^*_z/2 \leq \hat{\tau}_z \leq 3\tau^*_z/2$ with probability at least $1 - 4n^{-c}$ for all sufficiently large $n$. The desired result then follows from (2.7).

\[\square\]

### S1.7 Location and Scale Equivariant of Our Estimator

As noticed in Remark 2.2 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, $\tau$ 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

$$\sum_{i=1}^{n} \text{sign}(aX_i + b - \theta) \min\{|aX_i + b - \theta|, \tau\} = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} \min\{(aX_i + b - \theta)^2, \tau^2\}/\tau^2 = \frac{z}{n} = 0,$$
which is equivalent to

\[
\begin{aligned}
\sum_{i=1}^{n} \text{sign}(X_i - \tilde{\theta}) \min\{|X_i - \tilde{\theta}|, \tilde{\tau}\} &= 0 \\
\frac{1}{n} \sum_{i=1}^{n} \min\{(X_i - \tilde{\theta})^2, \tilde{\tau}^2\}/\tilde{\tau}^2 - \frac{z}{n} &= 0
\end{aligned}
\]  

by letting \( \tilde{\theta} = a^{-1}(\theta - b) \) and \( \tilde{\tau} = |a|^{-1}\tau \). 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 \( \theta \) 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 2.2 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

\[
\begin{aligned}
\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta_n}{\tau_n}\right) &= 0 \\
\sum_{i=1}^{n} \chi\left(\frac{X_i - \theta_n}{\tau_n}\right) &= 0
\end{aligned}
\]  

(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 \( \text{IC}(x; F, \theta) \) and \( \text{IC}(x; F, \tau) \) of our estimator \( (\theta_n, \tau_n) \) 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

\[
\text{IC}(x; F, \theta) = \frac{\psi(x/\tau(F))\tau(F)}{\mathbb{E}_F[\psi'(X_i/\tau(F))]} = \frac{\text{sign}(x) \min\{|x|, \tau(F)\}}{\mathbb{P}_F\{X_i \leq \tau(F)\}}
\]

and

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

The gross error sensitive of our estimator $\text{GES}(\tau; F) = \sup_x |\text{IC}(x; F, \tau)|$ is bounded for each fixed \( n \). This reflects the relative influence of individual outlier toward our estimator for finite sample. On the other hand, $\text{GES}(\tau; F) \geq \tau(F) \sup_x \left| \frac{\min\{x^2, \tau^2(F)\} / 2 - z\tau^2(F)/(2n)}{\mathbb{E}_F(X_i^2)} \right| \geq \tau^3(F) \left( \frac{1}{2} - \frac{z}{2n} \right)$.

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 $n \to \infty$, $\text{GES}(\tau; F)$ 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 $\epsilon$-contamination model in the traditional robust statistics. In fact, for the fixed $\tau$ such as $\tau = 1.345\sigma$, it guarantees a bounded $\text{GES}(\tau; F)$ 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 3.1 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 $n^\beta$ 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 3.1

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

\[ G(\theta) = \mathbb{E}\{\ell_\tau(\varepsilon - (\beta_0 - \beta_0^*) - (X, \beta - \beta^*))\} \]

\[ = \mathbb{E}[\mathbb{E}\{\ell_\tau(\varepsilon_i - (\beta_0 - \beta_0^*) - (X, \beta - \beta^*)) | X\}] \]

\[ \geq \mathbb{E}\{\ell_\tau(\varepsilon - \alpha_\tau)\} = G(\tilde{\theta}_\tau^*) , \]

where \( \tilde{\theta}_\tau^* := (\beta_0^* + \alpha_\tau, \beta^*)^T \in \mathbb{R}^d \). This implies that \( G(\beta_0^* + \alpha_\tau, \beta^*) = \min_{\theta \in \mathbb{R}^d} G(\theta) \).

Moreover, consider the Hessian matrix \( \nabla^2 G(\theta) = \mathbb{E}\{I(|Y - Z^T \theta| \leq \tau) Z Z^T\} \in \mathbb{R}^d \).

By (3.4), \( \nabla^2 G(\tilde{\theta}_\tau^*) = \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau) \mathbb{E}(Z Z^T) \) is positive definite, such that \( \tilde{\theta}_\tau^* \) is the unique minimizer of the function \( \theta \mapsto G(\theta) \). This enforces \( \beta_0^* + \alpha_\tau = \beta_0^* \) and \( \beta^* = \beta^* \).

Next we prove (3.5). By the optimality of \( \alpha_\tau \) and the mean value theorem, we have

\[ h'(\alpha_\tau) = \frac{dh(\alpha)}{d\alpha}|_{\alpha = \alpha_\tau} = 0 \]

and

\[ h''(\tilde{\alpha}_\tau)\alpha_\tau = h'(\alpha_\tau) - h'(0) = -h'(0) = \mathbb{E}\{\ell'_\tau(\varepsilon)\} . \]  \hspace{1cm} (S2.1)

where \( \tilde{\alpha}_\tau = \lambda_0 + (1 - \lambda) \alpha_\tau \) for some \( 0 \leq \lambda \leq 1 \). Note that

\[ h''(\tilde{\alpha}_\tau) = 1 - \mathbb{P}(|\varepsilon - \tilde{\alpha}_\tau| > \tau) . \]  \hspace{1cm} (S2.2)

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

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

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

\[ \tau \mathbb{E}|\varepsilon - \tilde{\alpha}_\tau| I(|\varepsilon - \tilde{\alpha}_\tau| > \tau) \leq \frac{\tau^2}{2} \mathbb{P}(|\varepsilon - \tilde{\alpha}_\tau| > \tau) + \frac{\sigma^2}{2} . \]
Further, by Markov’s inequality,

\[
P(|\varepsilon - \tilde{\alpha}_\tau| > \tau) \leq \tau^{-1}E|\varepsilon - \tilde{\alpha}_\tau|I(|\varepsilon - \tilde{\alpha}_\tau| > \tau) \leq \frac{1}{2}P(|\varepsilon - \tilde{\alpha}_\tau| > \tau) + \frac{\sigma^2}{2\tau^2},
\]

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

\[
h''(\tilde{\alpha}_\tau) \geq 1 - \tau^{-2}\sigma^2. \quad (S2.3)
\]

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

\[
|E\{\ell'_\tau(\varepsilon)\}| \leq E(|\varepsilon| - \tau)I(|\varepsilon| > \tau) \leq \tau^{-1}E(\varepsilon^2 - \tau^2)I(|\varepsilon| > \tau) = \frac{\sigma^2}{\tau} - \frac{E\{\psi^2_\tau(\varepsilon)\}}{\tau}
\]

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

\[\square\]

### S2.2 Proof of Theorem 3.1

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

\[
\Theta_r = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_S \leq r\}, \quad (S2.4)
\]
where $\| \cdot \|_S$ denotes the rescaled $\ell_2$-norm $\| u \|_S = \| S^{1/2} u \|_2$ for $u \in \mathbb{R}^d$. If $\tilde{\Theta}_r \notin \Theta_r$, there exists some $\eta \in (0, 1)$ such that $\tilde{\theta} = \theta^* + \eta(\hat{\theta}_r - \theta^*) \in \Theta_r$; otherwise if $\tilde{\theta} \in \Theta_r$, we can simply take $\eta = 1$. By the optimality of $\hat{\theta}_r$, we have $\nabla L(\hat{\theta}_r) = 0$. Applying Lemma 2 in Sun, Zhou, and Fan (2019) to $L(\theta) = (1/n) \sum_{i=1}^n \ell_\tau(Y_i - Z_i^\top \theta)$ gives

$$\langle \nabla L(\tilde{\theta}) - \nabla L(\theta^*), \tilde{\theta} - \theta^* \rangle \leq \eta \langle -\nabla L(\theta^*), \hat{\theta}_r - \theta^* \rangle \leq \| S^{-1/2} \nabla L(\theta^*) \|_2 \times \| \tilde{\theta} \|_S. \quad \text{(S2.5)}$$

In what follows, we bound the left-hand and right-hand sides of (S2.5) separately, starting with the former. Proposition S2.1 below shows that $L_\tau$ is strictly convex on $\Theta_r$ with high probability.

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

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

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

$$\langle \nabla L_\tau(\theta) - \nabla L_\tau(\theta^*), \theta - \theta^* \rangle \geq \frac{1}{4}\| \theta - \theta^* \|_S^2 \quad \text{uniformly over} \quad \theta \in \Theta_r. \quad \text{(S2.7)}$$

**Proof of Theorem 3.1.** Since $\tilde{\theta} \in \Theta_r$ by construction, it holds under the scaling (S2.6) that

$$\langle \nabla L_\tau(\tilde{\theta}) - \nabla L_\tau(\theta^*), \tilde{\theta} - \theta^* \rangle \geq \frac{1}{4}\| \tilde{\theta} - \theta^* \|_S^2 \quad \text{(S2.8)}$$
with probability at least $1 - e^{-z}$.

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

$$
\left\| \frac{1}{n} \sum_{i=1}^n \left\{ \xi_i z_i - \mathbb{E}(\xi_i z_i) \right\} \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\xi_i z_i) \right\|_2,
$$

(S2.9)

where $\xi_i = \ell'_i(\epsilon_i)$ and $z_i = S^{-1/2}Z_i$. To bound $\|\gamma\|_2 = \sup_{u \in \mathbb{S}^d} u^T \gamma$, by a standard covering argument, we can find a $(1/2)$-net $\mathcal{N}_{1/2}$ of $\mathbb{S}^d$ with $|\mathcal{N}_{1/2}| \leq 5^d$ such that $\|\gamma\|_2 \leq 2 \max_{u \in \mathcal{N}_{1/2}} u^T \gamma$. For every $u \in \mathbb{S}^d$, note that $u^T \gamma = \sum_{i=1}^n (\xi_i u^T z_i - \mathbb{E}(\xi_i u^T z_i))$.

Under Condition 3.1, we have

$$
\mathbb{E}|u^T z_i|^k = A_0^k k \int_0^\infty t^{k-1} \mathbb{P}(|u^T z_i| \geq A_0 t) \, dt \leq A_0^k k \int_0^\infty t^{k-1} e^{-t} \, dt = A_0^k k!, \quad k \geq 1,
$$

from which it follows that $\mathbb{E}(\xi_i u^T z_i)^2 = \mathbb{E}\{\mathbb{E}(\xi_i^2 | z_i)(u^T z_i)^2\} \leq \sigma^2$ and

$$
\mathbb{E}|\xi_i u^T z_i|^k \leq \frac{k!}{2} 2A_0^2 \sigma^2 (A_0 \tau)^{k-2} \quad \text{for all } k = 3, 4, \ldots.
$$

By Bernstein’s inequality,

$$
\mathbb{P}\left( u^T \gamma \geq 2A_0 \sigma \sqrt{\frac{x}{n} + A_0 \tau \frac{x}{n}} \right) \leq e^{-x} \quad \text{for any } x > 0.
$$

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

$$
\mathbb{P}\left( \|\gamma\|_2 \geq 4A_0 \sigma \sqrt{\frac{2\bar{d} + z}{n}} + 2A_0 \tau \frac{2\bar{d} + z}{n} \right) \leq e^{-z}.
$$
For the deterministic part $\|E(\xi_i z_i)\|_2$ in (S2.9), it holds $\|E(\xi_i z_i)\|_2 = \sup_{u \in S^d} E(\xi_i u^T z_i) \leq \sigma^2 \tau^{-1}$. Putting the above calculations together yields that with probability greater than $1 - e^{-z}$,

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

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 $1 - 2e^{-z}$, $\|\tilde{\theta} - \theta^*\|_S \leq 4\|S^{-1/2} \nabla \mathcal{L}_\tau(\theta^*)\|_2 < 4r_0$ under the assumed scaling (S2.6). Provided $n \gtrsim d + z$ so that $4r_0 \leq r$, the intermediate estimator $\tilde{\theta}$ will lie in the interior of $\Theta_r$, which enforces $\eta = 1$ and $\tilde{\theta} = \tilde{\theta}_r$ (otherwise $\tilde{\theta}$ will lie on the boundary). Putting together the pieces, we arrive at the desired result.

### S2.3 Proof of Theorem 3.2

In view of the proof of Theorem 3.1, 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 $\tau$ is set as a constant rather than a function of $(n, d)$. Since the target parameter now is $\theta^*_r$, we slightly change the notation and set

$$\tilde{\theta} = \theta^*_r + \eta(\tilde{\theta}_r - \theta^*_r) \quad \text{and} \quad \tilde{\Theta}_r = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*_r\|_S \leq r\}$$

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

We start with the deviation bound. Recalling $\theta^*_r = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n E\ell_{\tau}(Y_i - Z_i^T \theta)$,
it follows from Proposition 3.1 that
\[ 0 = \nabla \mathbb{E} L_{\tau}(\theta^*_\tau) = \mathbb{E} \nabla L_{\tau}(\theta^*_\tau) = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \{ \ell'_\tau(\varepsilon_i - \alpha_\tau) Z_i \}, \]
where \( L_{\tau}(\theta) = (1/n) \sum_{i=1}^{n} \ell_{\tau}(Y_i - Z_i^\top \theta) \). Recalling that \( z_i = S^{-1/2} Z_i \),
\[ \| S^{-1/2} \nabla L_{\tau}(\theta^*_\tau) \|_2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \ell'_\tau(\varepsilon_i - \alpha_\tau) z_i \right\|_2. \] (S2.11)
Since \( \mathbb{E} \{ \ell'_\tau(\varepsilon_i - \alpha_\tau) \} = 0 \) and by the optimality of \( \alpha_\tau \),
\[ \text{var}(\ell'_\tau(\varepsilon_i - \alpha_\tau)) = \mathbb{E} \{ \ell'_\tau(\varepsilon_i - \alpha_\tau) \}^2 = \mathbb{E}(\varepsilon_i - \alpha_\tau)^2 I(|\varepsilon_i - \alpha_\tau| \leq \tau) \leq 2 \mathbb{E} \ell_{\tau}(\varepsilon_i - \alpha_\tau) \leq 2 \mathbb{E} \ell_{\tau}(\varepsilon) \leq \sigma^2. \]
Following the same argument as in the proof of Theorem 3.1, it can be shown that with probability at least \( 1 - 5^d e^{-x} \),
\[ \left\| \frac{1}{n} \sum_{i=1}^{n} \ell'_\tau(\varepsilon_i - \alpha_\tau) z_i \right\|_2 < 4 A_0 \sigma \sqrt{\frac{x}{n}} + 2 A_0 \tau \frac{x}{n}. \]
Taking \( x = 2d + z \) in the last display, we obtain from (S2.11) that
\[ \| S^{-1/2} \nabla L_{\tau}(\theta^*_\tau) \|_2 < r_1 := 4 A_0 \sigma \sqrt{\frac{2d + z}{n}} + 2 A_0 \tau \frac{2d + z}{n} \] (S2.12)
with probability at least \( 1 - e^{-z} \).

The next proposition provides the restricted strong convexity around \( \theta^*_\tau \) when \( \tau \) is treated as a constant.

Proposition S2.2. Let \( m_4 = \sup_{u \in S^d} \mathbb{E}(S^{-1/2} Z, u)^4 \). Let \( \tau, r > 0 \) satisfy
\[ \tau \geq 8 m_4^{1/2} r \quad \text{and} \quad n \geq c_0 \rho_\tau^{-2} (\tau/r)^2 (d + z), \] (S2.13)
where \( c_0 > 0 \) is an absolute constant. Then with probability at least \( 1 - e^{-z} \),

\[
\langle \nabla \mathcal{L}_\tau(\theta) - \nabla \mathcal{L}_\tau(\theta^*), \theta - \theta^* \rangle \geq \frac{\rho_\tau}{2} \| \theta - \theta^* \|_S^2 \text{ uniformly over } \theta \in \tilde{\Theta}_r. \tag{S2.14}
\]

According to (S2.12) and (S2.14), we take \( r = \tau/(8m_4^{1/2}) \) so that with probability at least \( 1 - 2e^{-z} \), \( \| \tilde{\theta} - \theta^* \|_S \leq 2\rho_\tau^{-1}\|S^{-1/2}\nabla \mathcal{L}_\tau(\theta^*)\|_2 < 2\rho_\tau^{-1}r_1 \) 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 \( \tilde{\theta} \) will lie in the interior of \( \tilde{\Theta}_r \), which enforces \( \eta = 1 \) and \( \tilde{\theta} = \tilde{\theta}_r \), as desired.

\[\square\]

**S2.4 Proof of Proposition S2.1**

Since the Huber loss is convex and differentiable, we have

\[
\mathcal{T}(\theta) := \langle \nabla \mathcal{L}_\tau(\theta) - \nabla \mathcal{L}_\tau(\theta^*), \theta - \theta^* \rangle \\
= \frac{1}{n} \sum_{i=1}^{n} \left\{ \ell'_\tau(Y_i - Z_i^\top \theta^*) - \ell'_\tau(Y_i - Z_i^\top \theta) \right\} Z_i^\top (\theta - \theta^*) \\
\geq \frac{1}{n} \sum_{i=1}^{n} \left\{ \ell'_\tau(\varepsilon_i) - \ell'_\tau(Y_i - Z_i^\top \theta) \right\} Z_i^\top (\theta - \theta^*) I_{E_i}, \tag{S2.15}
\]

where \( I_{E_i} \) is the indicator function of the event

\[
E_i := \{ |\varepsilon_i| \leq \tau/2 \} \cap \left\{ \frac{| \langle Z_i, \theta - \theta^* \rangle |}{\| \theta - \theta^* \|_S} \leq \frac{\tau}{2r} \right\}, \tag{S2.16}
\]
on which $|Y_i - Z_i^\top \theta| \leq \tau$ for all $\theta \in \Theta_r$. Also, recall that $\ell''_r(u) = 1$ for $|u| \leq \tau$. For any $R > 0$, define functions

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

In particular, $\varphi_R$ is $R$-Lipschitz and satisfies

$$u^2 I(|u| \leq R/2) \leq \varphi_R(u) \leq u^2 I(|u| \leq R). \quad (S2.17)$$

It then follows that

$$\mathcal{T}(\theta) \geq g(\theta) := \frac{1}{n} \sum_{i=1}^n \varphi_{\tau \|\theta - \theta^*\|_S/2r}(\langle Z_i, \theta - \theta^* \rangle) \psi_{\tau/2}(\varepsilon_i). \quad (S2.18)$$

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

$$\Delta_r := \sup_{\theta \in \Theta_r} \frac{-g(\theta) + \mathbb{E}g(\theta)}{\|\theta - \theta^*\|^2_S}. \quad (S2.19)$$

Write $\delta = \theta - \theta^*$ for $\theta \in \Theta_r$. By (S2.17),

$$\mathbb{E}g(\theta) \geq \mathbb{E}\langle Z_i, \delta \rangle^2 - \mathbb{E}\{\langle Z_i, \delta \rangle^2 I(|\langle Z_i, \delta \rangle| \geq \frac{\tau}{4r} \|\delta\|_S)\} - \mathbb{E}\{\langle Z_i, \delta \rangle^2 I(|\varepsilon_i| > \tau/2)\} \geq \|\delta\|^2_S - \frac{4}{\tau^2} \left( \frac{4r^2}{\|\delta\|^2_S} \mathbb{E}\langle Z_i, \delta \rangle^4 + \mathbb{E}\langle Z_i, \delta \rangle^2 \varepsilon_i^2 \right). \quad (S2.20)$$
Recall that $\mathbb{E}(\delta, Z_i)^4 \leq m_4\|\delta\|_S^4$ for all $u \in \mathbb{R}^\tilde{d}$ ($\tilde{d} = d + 1$). In conjunction with (S2.20), this implies

$$\mathbb{E}g(\theta) \geq \|\delta\|_S^2 - \|\delta\|_S^2(\sigma^2 + 4m_4r^2)\frac{4}{\tau^2} \geq \frac{1}{2}\|\delta\|_S^2 \text{ for all } \theta \in \Theta_r,$$  

(S2.21)

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

$$\frac{T(\theta)}{\|\theta - \theta^*\|_S^2} \geq \frac{1}{2} - \Delta_r \text{ for all } \theta \in \Theta_r.$$  

(S2.22)

The following lemma provides an upper bound on the stochastic term $\Delta_r$.

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

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

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

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

$$\frac{T(\theta)}{\|\theta - \theta^*\|_S^2} \geq \frac{1}{4} \text{ uniformly over } \theta \in \Theta_r$$

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

\[\Box\]

### S2.5 Proof of Proposition S2.2

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

$$\langle \nabla \mathcal{L}_\tau(\theta) - \nabla \mathcal{L}_\tau(\theta^*_\tau), \theta - \theta^*_\tau \rangle$$
\[
\geq \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_r(Y_i - Z_i^\top \theta_r^*) - \ell'_r(Y_i - Z_i^\top \theta) \} Z_i^\top (\theta - \theta_r^*) I_{E_{r,i}},
\]

where
\[
E_{r,i} = \{ |\varepsilon_i - \alpha_r| \leq \tau/2 \} \cap \left\{ \frac{|\langle Z_i, \theta - \theta_r^* \rangle|}{\|\theta - \theta_r^*\|_S} \leq \frac{\tau}{2r} \right\}
\]

On \( E_{r,i} \), \(|Y_i - Z_i^\top \theta_r^*| = |\varepsilon_i + \beta_0^* - \beta_{0,r}^*| = |\varepsilon_i - \alpha_r| \leq \tau \) and \(|Y_i - Z_i^\top \theta| \leq \tau \) for all \( \theta \in \Theta_r \).

Moreover, let \( g(\theta) \) be as in (S2.18) except with \( \theta^* \) replaced by \( \theta_r^* \). By assumption
\[
\rho_r := P(|\varepsilon - \alpha_r| \leq \tau/2) > 0 \quad \text{and Markov's inequality, we obtain that for every } \theta \in \Theta_r,
\]
\[
Eg(\theta) \geq \rho_r \left( 1 - 16m_4 r^2 \tau^{-2} \right) \| \theta - \theta_r^* \|_S^2 \geq \frac{3}{4} \rho_r \| \theta - \theta_r^* \|_S^2,
\]

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

### S2.6 Proof of Lemma S2.1

For \( g(\theta) \) given in (S2.18), we write \( g(\theta) = (1/n) \sum_{i=1}^{n} g_i(\theta). \) Observing that \( \varphi_R(u) \leq R^2/4 \) and \( \psi_R(u) \leq 1 \) for all \( u \in \mathbb{R} \), we have \( 0 \leq g_i(\theta) \leq (\tau/4r)^2 \| \theta - \theta^* \|_S^2. \)

It then follows from Theorem 7.3 in Bousquet (2003), a variant of Talagrand’s inequality, that for any \( x > 0 \),
\[
\Delta_r \leq E\Delta_r + (E\Delta_r)^{1/2} \frac{\tau}{2r} \sqrt{\frac{x}{n}} + (2m_4)^{1/2} \sqrt{\frac{x}{n}} + \frac{\tau^2 x}{48r^2 n}
\]
\[
\leq 1.25 E\Delta_r + (2m_4)^{1/2} \sqrt{\frac{x}{n}} + \frac{\tau^2 x}{3r^2 n} \quad \text{(S2.23)}
\]

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

$$E_i(\delta) := \frac{1}{\|\delta\|_S^2} \varphi_{\tau/2}(\langle Z_i, \delta \rangle) \psi_{\tau/2}(\varepsilon_i) = \varphi_{\tau/2}(\langle Z_i, \delta \rangle) \psi_{\tau/2}(\varepsilon_i), \quad \delta = \theta - \theta^*.$$  

Applying the symmetrization inequality for empirical processes yields

$$\mathbb{E} \Delta_r \leq 2 \mathbb{E} \left\{ \sup_{\theta \in \Theta_r} \frac{1}{n} \sum_{i=1}^{n} e_i E_i(\delta) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since $\varphi_R$ is $R$-Lipschitz, $E_i(\delta)$ is a $(\tau/2r)$-Lipschitz function in $\langle Z_i, \delta/\|\delta\|_S \rangle$. By Talagrand’s contraction principle (see, e.g. Theorem 4.4, Theorem 4.12 and (4.20) in Ledoux and Talagrand (1991)),

$$\mathbb{E} \Delta_r \leq \frac{\tau}{r} \mathbb{E} \left\{ \sup_{\theta \in \Theta_r} \frac{1}{n} \sum_{i=1}^{n} e_i \langle Z_i, \delta/\|\delta\|_S \rangle \right\} \leq \frac{\tau}{r} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} e_i S^{-1/2} Z_i \right\|_2 \leq \frac{\tau}{r} \sqrt{\frac{d+1}{n}}.$$

This, together with (S2.23), proves the stated result. \hfill \Box

### S3 Proof of Theorem 3.3

This proof is based on an argument similar to that used in the proof of Theorem 3.1. 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 $\beta^*$ is not inherited by $\beta^*_\tau$. In this proof, we follow a different route to directly establish the convergence around $\theta^*$ instead of $\theta^*_\tau$. For simplicity, we write $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}^\top) = \hat{\theta}_H(\tau, \lambda) \in \mathbb{R}^d$ with $d = d + 1$. For some $r > 0$ to be specified, we use $\Theta_r$ and $\|\cdot\|_S$ to denote the local neighborhood
and rescaled $\ell_2$-norm as in (S2.4). As before, let $\hat{\theta}_\eta (0 < \eta \leq 1)$ be an intermediate estimator satisfying (i) $\hat{\theta}_\eta \in \Theta_r$, (ii) $\hat{\theta}_\eta$ lies on the boundary of $\Theta_r$ with $\eta \in (0, 1)$ if $\hat{\theta} \notin \Theta_r$, and (iii) $\hat{\theta}_1 = \hat{\theta}$. Moreover, $\hat{\theta}$ and $\hat{\theta}_\eta$ fulfill

$$
\langle \nabla L_{\tau}(\hat{\theta}_\eta) - \nabla L_{\tau}(\theta^*), \hat{\theta}_\eta - \theta^* \rangle \leq \eta \langle \nabla L_{\tau}(\hat{\theta}) - \nabla L_{\tau}(\theta^*), \hat{\theta} - \theta^* \rangle. \quad (S3.1)
$$

Write $\hat{\delta} = (\hat{v}_0, \hat{v}^\top) = \hat{\theta} - \theta^*$ and denote by $S \subseteq \{1, \ldots, d\}$ the the support of $\beta^*$. Define the $\ell_1$-cone $C \subseteq \mathbb{R}^d$ as

$$
C = \{ \theta \in \mathbb{R}^d : \|v_S\|_1 \leq 3\|v_S\|_1 + |v_0| \text{ for } (v_0, v^\top) = \theta - \theta^* \}.
$$

It can be shown that the optimal solution $\hat{\theta}$ to program (3.9) satisfies

$$
\hat{\theta} \in C \text{ on the event } \{ \lambda \geq 2\|\nabla L_{\tau}(\theta^*)\|_\infty \}, \quad (S3.2)
$$

from which it follows

$$
\|\hat{\delta}\|_1 = |\hat{v}_0| + \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1 \leq 2|\hat{v}_0| + 4\|\hat{v}_S\|_1 \leq 4(s + 1)^{1/2}\|\hat{\delta}\|_2. \quad (S3.3)
$$

By necessary conditions of the optima for convex problem (3.9),

$$
\langle \nabla L_{\tau}(\hat{\theta}) + \lambda \hat{z}, \hat{\theta} - \theta^* \rangle \leq 0,
$$

where $\hat{z} = (0, \hat{u}^\top)^\top$ with $\hat{u} \in \partial \beta$ satisfies $\langle \hat{z}, \theta^* - \hat{\theta} \rangle \leq \|\beta^*\|_1 - \|\beta\|_1$. Under the scaling $\lambda \geq 2\|\nabla L_{\tau}(\theta^*)\|_\infty$, it holds

$$
\langle \nabla L_{\tau}(\hat{\theta}) - \nabla L_{\tau}(\theta^*), \hat{\theta} - \theta^* \rangle
$$

$$
\leq \lambda \|\beta^*\|_1 - \|\beta\|_1 + \frac{\lambda}{2}\|\theta - \theta^*\|_1 + \|S^{-1/2}\nabla L_{\tau}(\theta^*)\|_2 \|\hat{\theta} - \theta^*\|_S
$$
\[
\leq \lambda (\|s_1 \hat{v}_S - s_1 \hat{v}_S\|_1) + \frac{\lambda}{2} \|\hat{v} - \theta^*\|_1 + S^{-1/2}\mathbb{E}\nabla \mathcal{L}_\tau(\theta^*)\|_2 \|\hat{v} - \theta^*\|_S
\]
\[
\leq \frac{\lambda}{2} (3\|s_1 \hat{v}_S - s_1 \hat{v}_S\|_1) + \frac{\lambda}{2} \|\hat{v}_0\| + \frac{\sigma^2}{\tau} \|\hat{v} - \theta^*\|_S,
\]
where the last step uses the bound \(\|S^{-1/2}\nabla \mathcal{L}_\tau(\theta^*)\|_2 = \sup_{u \in \mathbb{S}} \mathbb{E}\{\ell'_r(\varepsilon) \langle u, S^{-1/2}Z \rangle\} \leq \sigma^2 \tau^{-1}\). Together with (S3.1), this implies

\[
\langle \nabla \mathcal{L}_\tau(\hat{\theta}_\eta) - \nabla \mathcal{L}_\tau(\theta^*), \hat{\theta}_\eta - \theta^* \rangle \leq \frac{1}{4} \|\hat{v} - \theta^*\|_S.
\]

Moreover, we introduce \(\hat{\delta}_\eta = \hat{\theta}_\eta - \theta^*\) and note that \(\hat{\delta}_\eta = \eta \hat{\delta}\). By (S3.2), we also have \(\hat{\theta}_\eta \in \mathcal{C}\) 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 \(\Theta_r \cap \mathcal{C}\) with high probability. It is a straightforward extension of Proposition S2.1. We leave the proof to Section S3.2.

**Proposition S3.1.** Let \(m_4 = \sup_{u \in \mathbb{S}} \mathbb{E}(S^{-1/2}Z, u)^4\). Let \(\tau, r > 0\) satisfy

\[
\tau \geq \max(4\sigma, 8m_4^{1/2}r) \quad \text{and} \quad n \geq c_0 A_0 \lambda^{-1} \max_{1 \leq j \leq d} \sigma_{jj} (A_0 \tau/r)^2 s \log d,
\]

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

\[
\langle \nabla \mathcal{L}_\tau(\theta) - \nabla \mathcal{L}_\tau(\theta^*), \theta - \theta^* \rangle \geq \frac{1}{4} \|\theta - \theta^*\|_S^2 \quad \text{uniformly over } \theta \in \Theta_r \cap \mathcal{C}.
\]

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

\[
\langle \nabla \mathcal{L}_\tau(\hat{\theta}_\eta) - \nabla \mathcal{L}_\tau(\theta^*), \hat{\theta}_\eta - \theta^* \rangle \geq \frac{1}{4} \|\hat{\delta}_\eta\|_S^2.
\]
Substituting this lower bound into (S3.4) yields
\[
\frac{1}{4} \| \hat{\delta}_\eta \|_S^2 \leq \frac{1}{2} \lambda \eta (|\hat{v}_0| + 3 \| \hat{v}_S \|_1) + \frac{\sigma^2}{\tau} \| \hat{\delta}_\eta \|_S \\
\leq \frac{3}{2} (s + 1)^{1/2} \lambda \eta \| \hat{\delta}_\eta \|_2 + \frac{\sigma^2}{\tau} \| \hat{\delta}_\eta \|_S \\
= \frac{3}{2} (s + 1)^{1/2} \lambda \| \hat{\delta}_\eta \|_2 + \frac{\sigma^2}{\tau} \| \hat{\delta}_\eta \|_S.
\]

Canceling \( \| \hat{\delta}_\eta \|_S \) on both sides delivers
\[
\| \hat{\delta}_\eta \|_S \leq 6 \lambda^{-1/2} (s + 1)^{1/2} \lambda + 4 \sigma^2 \tau^{-1} \tag{S3.7}
\]
and
\[
\| \hat{\delta}_\eta \|_1 \leq 24 \lambda^{-1} (s + 1) \lambda + 16 \lambda^{-1/2} \sigma^2 (s + 1)^{1/2} \tau^{-1}
\]
under the assumed scaling \( \lambda \geq 2 \| \nabla L(T^\ast) - \nabla E L(T^\ast) \|_\infty \) and (S3.5).

It remains to tune the parameters \( \tau, \lambda \) and \( r \). The following result provides a concentration inequality for \( \| \nabla L(T^\ast) - \nabla E L(T^\ast) \|_\infty \).

**Proposition S3.2.** Assume Condition 3.1 holds and let \( \tau = \sigma \sqrt{n/t} \) for some \( t > 0 \).

Then with probability at least \( 1 - 2d^{-1} \),
\[
\| \nabla L(T^\ast) - \nabla E L(T^\ast) \|_\infty \leq 2 A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2} \sigma \left( \frac{2 \log d}{n} + \frac{\log d}{\sqrt{nt}} \right) \tag{S3.8}
\]

Applying Proposition S3.2 with \( t = \log d \), we find that
\[
\| \nabla L(T^\ast) - \nabla E L(T^\ast) \|_\infty \leq c_1 A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2} \sigma \sqrt{\log d/n}
\]
with probability at least \( 1 - 2d^{-1} \), where \( c_1 > 0 \) is an absolute constant. We therefore choose \( \lambda = c_2 A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2} \sigma \sqrt{(\log d)/n} \) for some constant \( c_2 \geq 2c_1 \), such that
\[
\lambda \geq 2 \| \nabla L(T^\ast) - \nabla E L(T^\ast) \|_\infty \text{ with high probability.}
\]
Putting the above calculations together and taking $r = \sigma$, we conclude that

$$\|\hat{\theta}_\eta - \theta^*\|_S \leq 6c_2\Lambda_S^{-1/2}A_0 \max_{1 \leq j \leq d} \sigma_{jj}^{1/2} \sigma \sqrt{\frac{(s + 1) \log d}{n}} < \sigma$$

with probability at least $1 - 3d^{-1}$, assuming the scaling $n \gtrsim \Lambda_S^{-1} A_0^2 m_4 \max_{0 \leq j \leq d} \sigma_{jj} s \log d$. By the construction of $\hat{\theta}_\eta$, with the same probability we must have $\eta = 1$ and therefore $\hat{\theta} = \hat{\theta}_\eta$. The stated result (3.10) then follows from (S3.7).

\[\square\]

**S3.1 Proof of (S3.2)**

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

$$\mathcal{L}_r(\hat{\theta}) - \mathcal{L}_r(\theta^*) \leq \lambda \left( \|\beta^*\|_1 - \|\hat{\beta}\|_1 \right).$$

By direct calculation, we have

$$\|\hat{\beta}\|_1 - \|\beta^*\|_1 \geq \|\beta^*_S + \hat{\beta}_{S^c}\|_1 - \|\beta^*_S\|_1 - \|\hat{\beta}_{S^c}\|_1 - \left( \|\beta^*_S\|_1 + \|\beta^*_{S^c}\|_1 \right)$$

$$\geq \|\hat{\beta}_{S^c}\|_1 - \|\hat{\beta}_S\|_1.$$

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

$$\mathcal{L}_r(\hat{\theta}) - \mathcal{L}_r(\theta^*) \geq \langle \nabla \mathcal{L}_r(\theta^*), \hat{\delta} \rangle \geq -\|\nabla \mathcal{L}_r(\theta^*)\|_\infty \|\hat{\delta}\|_1$$

$$\geq -\frac{\lambda}{2} \left( |\hat{\nu}_0| + \|\hat{\nu}_{S^c}\|_1 + \|\hat{\nu}_S\|_1 \right).$$

Putting the above calculations together delivers

$$0 \leq \frac{\lambda}{2} \left( |\hat{\nu}_0| + 3\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^c}\|_1 \right),$$

from which the conclusion follows. \[\square\]
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 $\Theta_r \cap C$:

$$\Delta_r := \sup_{\theta \in \Theta_r \cap C} \frac{-g(\theta) + \mathbb{E}g(\theta)}{\|\theta - \theta^*\|_S^2}.$$ 

Provided $\tau \geq \max(4\sigma, 8m_1^{1/2}r)$, it can be shown that

$$\mathcal{T}(\theta) \geq \frac{1}{2} - \Delta_r$$

for all $\theta \in \Theta_r \cap C$. (S3.9)

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

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

$$\Delta_r \leq 10\sqrt{2}A_0^{-1/2}A_0 \max_{0 \leq j \leq d} \sigma_{jj}^{1/2} \mathbb{E}_{\Theta_r \cap C} \left\{ \sqrt{\frac{\log(2d)}{n}} + \frac{\log(2\bar{d})}{n} \right\} + (2m_4)^{1/2} \sqrt{\frac{x}{n}} + \frac{\tau^2}{r^2} \frac{x}{3n}$$

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 $1 - d^{-1}$,

$$\mathcal{T}(\theta) \geq \frac{1}{4}$$

uniformly over $\theta \in \Theta_r \cap C$ as long as $n \geq \Delta_s^{-1} \max_{1 \leq j \leq d} \sigma_{jj}(A_0\tau/r)^{2s} \log d$. This proves (S3.6).

S3.3 Proof of Proposition S3.2

To begin with, recall that $\nabla \mathcal{L}_\tau(\theta^*) = (1/n) \sum_{i=1}^n \xi_i Z_i$, where $\xi_i = \ell'_\tau(\epsilon_i)$ and $Z_i = (X_{i0}, X_{i1}, \ldots, X_{id})^\top$ with $X_{i0} \equiv 1$. Hence,

$$\|\nabla \mathcal{L}_\tau(\theta^*) - \nabla \mathbb{E}\mathcal{L}_\tau(\theta^*)\|_{\infty} = \max_{0 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^n \xi_i X_{ij} - \mathbb{E}(\xi_i X_{ij}) \right|.$$
Next we use the union bound and Bernstein’s inequality to bound the maximum. For every $0 \leq j \leq d$,

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

and

$$\mathbb{E}(\xi_i X_{ij})^2 = \mathbb{E}\{(\xi_i^2 | X_{ij}) X_{ij}^2\} \leq \sigma_{jj}.$$

Moreover, for $k = 3, 4, \ldots$, \(\mathbb{E}|\xi_i X_{ij}|^k \leq \sigma^2 r^{-k-2} A_0 \sigma_{jj}^{k/2} k! = \frac{k!}{2} A_0^2 \sigma_{jj} \sigma^2 (A_0 \sigma_{jj}^{1/2} r)^{k-2} \).

Then it follows from Bernstein’s inequality that, for any $x > 0$,

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

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

$$\mathbb{E} \left\{ \sup_{\delta = \theta - \bar{\theta}, \bar{\delta} \in \Theta, \bar{\delta} \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n e_i \langle Z_i, \bar{\delta} / \|\bar{\delta}\| \rangle \right\} \leq 4A_{1/2}^{-1/2} (s + 1)^{1/2} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n e_i Z_i \right\|_{\infty},$$

where $Z_i = (X_{i0}, X_{i1}, \ldots, X_{id})^\top$ 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 $k = 3, 4, \ldots$,

$$\mathbb{E}|e_i X_{ij}|^k = A_0^k \sigma_{jj}^{k/2} k \int_0^\infty t^{k-1} \mathbb{P}(|X_{ij}| \geq A_0 \sigma_{jj}^{1/2} t) \, dt \leq k! A_0^k \sigma_{jj}^{k/2}.$$
The above moment inequalities, together with the symmetry of Rademacher random variable, yield
\[
Ee^{\lambda e_{i}X_{ij}} = 1 + \frac{1}{2} \sigma_{jj} \lambda^2 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!} E(e_{i}X_{ij})^k
\]
\[
\leq 1 + \frac{1}{2} \sigma_{jj} \lambda^2 + \sum_{\ell=2}^{\infty} \frac{\lambda^{2\ell}}{(2\ell)!} A_0^2 \sigma_{jj}^\ell
\]
\[
= 1 + \frac{A_0^2 \sigma_{jj}}{2} \sum_{k=2}^{\infty} \lambda^k \left( \sqrt{2} A_0 \sigma_{jj}^{1/2} \right)^{k-2}
\]
\[
\leq 1 + \frac{1}{2} A_0^2 \sigma_{jj} \lambda^2 \left( 1 - \frac{1}{\sqrt{2} A_0 \sigma_{jj}^{1/2}} \right), \quad \text{for all } 0 \leq \lambda < \frac{1}{\sqrt{2} A_0 \sigma_{jj}^{1/2}}.
\]

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
\[
\log E e^{\lambda S_j} \leq \psi(\lambda) := \frac{v \lambda^2}{2(1-c \lambda)} \text{ for any } 0 \leq j \leq d \text{ and } \lambda \in (0, 1/c), \text{ and hence}
\]
\[
E \max_{0 \leq j \leq d} |S_j| \leq \inf_{\lambda \in (0, 1/c)} \left\{ \frac{\log(2d) + \psi(\lambda)}{\lambda} \right\} = \sqrt{2v \log(2d) + c \log(2d)}.
\]

Putting together the pieces, we conclude that
\[
E \Delta_r \leq 8 \sqrt{2} A_0^{-1/2} \max_{0 \leq j \leq d} \sigma_{jj}^{-1/2} (s + 1)^{1/2} \tau \left\{ \sqrt{\frac{\log(2d)}{n} + \frac{\log(2d)}{n}} \right\}.
\]

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

S4 Additional Simulation Studies

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

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.

Figure S1: Estimation errors for the sample mean, CV-Huber and DA-Huber estimators under different settings based on 2000 simulations.
Figure S2: Distributions of the $\ell_2$-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 \( \alpha \)-quantiles of the mean absolute prediction errors for the three data sets, where \( \alpha \) 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 \( \ell : \mathbb{R} \to [0, \infty) \) satisfies the following condi-
tions with some constants $c_1$-$c_4$:

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

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 $x \in \mathbb{R}$.

Below we list a few examples of $\ell$ 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

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

respectively.

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

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

4. Smoothed Huber loss I: The Huber loss is twice differentiable in $\mathcal{R}$, except at $\pm 1$. 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

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

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


HAHN, M. G., KUELBS, J. and WEINER, D. C. (1990). The asymptotic joint distribu-


