Submitted to the Annals of Statistics

SUPPLEMENT TO "ROBUST INFERENCE VIA MULTIPLIER BOOTSTRAP"

By Xi Chen* and Wen-Xin Zhou†

*New York University** *and University of California, San Diego*†

This supplemental material contains (1) the proofs of Theorems 2.1–2.6 and Theorem 3.1 in the main text, (2) implementation of the proposed methods, and (3) additional simulation studies.

APPENDIX A: NOTATION AND PRELIMINARIES

A.1. Notation. Recall that the error variable ε has mean zero and variance $\sigma^2 = \mathbb{E}(\varepsilon^2) > 0$. For every $\tau > 0$, we define the truncated mean and second moment of ε to be

$$m_\tau = \mathbb{E}\{\psi_\tau(\varepsilon)\} \quad \text{and} \quad \sigma_\tau^2 = \mathbb{E}\{\psi_\tau^2(\varepsilon)\}, \tag{A.1}$$

where $\psi_\tau(u) := \ell_\tau'(u) = \mathrm{sgn}(u)\min(|u|, \tau)$, $u \in \mathbb{R}$. For IID random variables $\varepsilon_1, \ldots, \varepsilon_n$ from ε, we define truncated variables

$$\xi_i = \psi_\tau(\varepsilon_i), \quad i = 1, \ldots, n. \tag{A.2}$$

The dependence of ξ_i on τ will be assumed without displaying.

Moreover, define

$$\mathbf{S}_n = \mathbf{S}_{n,\tau} = \frac{1}{n}\sum_{i=1}^n \xi_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i^\intercal \quad \text{and} \quad M_{n,4} = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n (\boldsymbol{u}^\intercal \boldsymbol{Z}_i)^4. \tag{A.3}$$

Throughout, we use $\mathbb{P}^\dagger$-probability to denote the probability measure over $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ and use $\mathbb{P}^*$-probability to denote the probability measure over $\{U_i\}_{i=1}^n$ conditioning on $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. In general, $\mathbb{P}$ denotes the probability measure over all the random variables involved.

A.2. Technical lemmas. In this section, we provide several technical lemmas that will be used repeatedly to prove the main theorems. Recall the isotropic random vectors $\boldsymbol{Z}_i$ given in (2.3). The first two lemmas provide concentration properties for $M_{n,4}$ and $\mathbf{S}_n$, respectively.

LEMMA A.1. Assume Condition 2.1 holds. Then for any $x > 0$,

$$M_{n,4} \le 2 \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E}(\boldsymbol{u}^\intercal \boldsymbol{Z})^4 + C\left\{\sqrt{\frac{3d+x}{n}} + \frac{(3d+x)^2}{n}\right\} \tag{A.4}$$

with probability at least $1 - 2e^{-x}$, where $C > 0$ depends only on A_0.

PROOF. The proof is based on the standard covering argument. For any $\epsilon \in (0,1)$, let $\mathcal{N}_\epsilon$ be an ϵ-net of the unit sphere $\mathbb{S}^{d-1}$ with $\mathrm{card}(\mathcal{N}_\epsilon) \le (1+2/\epsilon)^d$ such that for every $\boldsymbol{u} \in \mathbb{S}^{d-1}$, there exists some $\boldsymbol{v} \in \mathcal{N}_\epsilon$ satisfying $\|\boldsymbol{u} - \boldsymbol{v}\|_2 \le \epsilon$. Define the map $f : \mathbb{R}^d \to \mathbb{R}^n$ as

$$f(\boldsymbol{u}) = n^{-1/4}(\boldsymbol{u}^\intercal \boldsymbol{Z}_1, \ldots, \boldsymbol{u}^\intercal \boldsymbol{Z}_n)^\intercal, \quad \boldsymbol{u} \in \mathbb{R}^d.$$

By the triangle inequality,

$$\begin{aligned}\|f(\boldsymbol{u})\|_4 &\le \|f(\boldsymbol{v})\|_4 + \|f(\boldsymbol{u}) - f(\boldsymbol{v})\|_4 \\ &= \|f(\boldsymbol{v})\|_4 + \left(\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}-\boldsymbol{v}, \boldsymbol{Z}_i\rangle^4\right)^{1/4} \le \|f(\boldsymbol{v})\|_4 + \epsilon M_{n,4}^{1/4}.\end{aligned}$$

Taking the maximum over $\boldsymbol{v} \in \mathcal{N}_\epsilon$ and then taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we arrive at

$$M_{n,4}^{1/4} \le N_{n,\epsilon}^{1/4} + \epsilon M_{n,4}^{1/4},$$

where $N_{n,\epsilon} = \max_{\boldsymbol{v}\in\mathcal{N}_\epsilon}(1/n)\sum_{i=1}^n (\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4$. Solving this inequality yields

$$M_{n,4} \le (1-\epsilon)^{-4} N_{n,\epsilon}. \tag{A.5}$$

For every $\boldsymbol{v} \in \mathcal{N}_\epsilon$, note that $\mathbb{P}\{(\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4 \ge y\} \le 2e^{-\sqrt{y}/A_0^2}$ for any $y > 0$. Hence, by inequality (3.6) in Adamczak et al. (2011) with $s = 1/2$, we obtain that for any $z > 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4 - \mathbb{E}(\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4\right| \ge z\right\} \le 2\exp\left\{-c\min\left(\frac{z^2}{nC_1^2}, \sqrt{\frac{z}{C_1}}\right)\right\},$$

where $c > 0$ is a universal constant and $C_1 > 0$ depends only on A_0. Taking the union bound over all vectors $\boldsymbol{v}$ in $\mathcal{N}_\epsilon$ gives

$$\begin{aligned}&\mathbb{P}\left\{\max_{\boldsymbol{v}\in\mathcal{N}_\epsilon}\left|\sum_{i=1}^n (\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4 - \mathbb{E}(\boldsymbol{v}^\intercal \boldsymbol{Z}_i)^4\right| \ge z\right\} \\ &\le 2\exp\left\{d\log(1+2/\epsilon) - c\min\left(\frac{z^2}{nC_1^2}, \sqrt{\frac{z}{C_1}}\right)\right\}.\end{aligned}$$

It follows that

$$N_{n,\epsilon} \le \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathbb{E}(\boldsymbol{u}^\intercal \boldsymbol{Z})^4 + C_2\left[\sqrt{\frac{d\log(1+2/\epsilon)+x}{n}} + \frac{\{d\log(1+2/\epsilon)+x\}^2}{n}\right] \tag{A.6}$$

with probability at least $1-2e^{-x}$, where $C_2>0$ depends only on A_0.

Finally, taking $\epsilon = 1/8$ in (A.5) and (A.6) implies (A.4). □

LEMMA A.2. Assume Condition 2.1 holds with $\delta = 2$. For any $x>0$,

$$\|\mathbf{S}_n - \sigma_\tau^2 \mathbf{I}_d\|_2 \le 4A_0^2 \upsilon_4^{1/2}\sqrt{\frac{3d+x}{n}} + \sqrt{2}A_0^2\tau^2\frac{3d+x}{n} \tag{A.7}$$

with probability at least $1-2e^{-x}$.

PROOF. Define random variables $w_i = \xi_i/\sigma_\tau$ so that $\mathbb{E}(w_i^2)=1$. We will bound $\|\boldsymbol{\Delta}\|_2$ via a standard covering argument, where

$$\boldsymbol{\Delta} = \frac{1}{n}\sum_{i=1}^n w_i^2 \boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal - \mathbf{I}_d = \frac{1}{\sigma_\tau^2}(\mathbf{S}_n - \sigma_\tau^2\mathbf{I}_d).$$

Proceed similarly to the proof of Lemma 4.4.1 in Vershynin (2018), it can be shown that for any $0<\epsilon<1/2$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of the unit sphere $\mathbb{S}^{d-1}$ with $\mathrm{card}(\mathcal{N}_\epsilon) \le (1+2/\epsilon)^d$ such that

$$\|\boldsymbol{\Delta}\|_2 \le \frac{1}{1-2\epsilon}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}|\boldsymbol{u}^\intercal\boldsymbol{\Delta}\boldsymbol{u}| = \frac{1}{1-2\epsilon}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\left|\frac{1}{n}\sum_{i=1}^n w_i^2(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2 - 1\right|.$$

For any $\boldsymbol{u}\in\mathbb{S}^{d-1}$, by (B.7) we have $\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z})^{2k} \le 2A_0^{2k}\,k!$ for all $k\ge 1$. This implies

$$\mathbb{E}\{w_i^4(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^4\} = \mathbb{E}\{(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^4\mathbb{E}(w_i^4|\boldsymbol{X}_i)\} \le 4A_0^4\,\sigma_\tau^{-4}\upsilon_4,$$
$$\text{and}\quad \mathbb{E}\{w_i^2(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2\}^k \le \frac{k!}{2}4A_0^4\,\sigma_\tau^{-4}\upsilon_4(A_0^2\,\sigma_\tau^{-2}\tau^2)^{k-2}\quad \text{for all } k\ge 3.$$

It then follows from Bernstein's inequality that for any $x\ge 0$,

$$\mathbb{P}\left\{|\boldsymbol{u}^\intercal\boldsymbol{\Delta}\boldsymbol{u}| \ge 2\sqrt{2}A_0^2\,\sigma_\tau^{-2}\upsilon_4^{1/2}\sqrt{\frac{x}{n}} + A_0^2\,\sigma_\tau^{-2}\tau^2\frac{x}{n}\right\} \le 2e^{-x}.$$

Taking the union bound over all $\boldsymbol{u} \in \mathcal{N}_\epsilon$ with $\epsilon = (1 - 1/\sqrt{2})/2$ yields

$$\|\boldsymbol{\Delta}\|_2 \le 4A_0^2\, \sigma_\tau^{-2} \upsilon_4^{1/2} \sqrt{\frac{x}{n}} + \sqrt{2} A_0^2\, \sigma_\tau^{-2} \tau^2 \frac{x}{n}$$

with probability at least $1 - 2e^{3d-x}$. Reinterpret this we reach (A.7). □

The next lemma gives a deviation bound for the ℓ_2-norm of the d-variate random vector $\boldsymbol{\xi}^\flat = -\sum_{i=1}^n \xi_i U_i \boldsymbol{Z}_i$, where ξ_i are given in (A.2). Recall that $\mathbb{P}^*$ is the conditional probability measure over the random multipliers given $\mathcal{D}_n = \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$.

LEMMA A.3. Assume Condition 2.2 is fulfilled. For every $x > 0$, it holds with $\mathbb{P}^*$-probability at least $1 - e^{-x}$ that

$$\frac{1}{\sqrt{n}} \|\boldsymbol{\xi}^\flat\|_2 \le B_U \|\mathbf{S}_n\|_2^{1/2} (2d + x)^{1/2}, \tag{A.8}$$

where $B_U > 0$ is a constant depending only on A_U.

PROOF. The proof is based on an argument similar to that leads to (B.8). For any $0 < \epsilon < 1$, There exists an ϵ-net $\mathcal{N}_\epsilon$ of $\mathbb{S}^{d-1}$ with cardinality $|\mathcal{N}_\epsilon| \le (1 + 2/\epsilon)^d$ such that

$$\|\boldsymbol{\xi}^\flat\|_2 \le \frac{1}{1-\epsilon} \max_{\boldsymbol{u} \in \mathcal{N}_\epsilon} \boldsymbol{u}^\intercal \boldsymbol{\xi}^\flat = \frac{1}{1-\epsilon} \max_{\boldsymbol{u} \in \mathcal{N}_\epsilon} \left(\sum_{i=1}^n \xi_i \boldsymbol{u}^\intercal \boldsymbol{Z}_i U_i \right).$$

For every $\boldsymbol{u} \in \mathbb{S}^{d-1}$, applying Proposition 2.5 in Wainwright (2019) gives

$$\mathbb{P}^* \left[\sum_{i=1}^n \xi_i \boldsymbol{u}^\intercal \boldsymbol{Z}_i U_i \ge C \left\{ \sum_{i=1}^n \xi_i^2 (\boldsymbol{u}^\intercal \boldsymbol{Z}_i)^2 \right\}^{1/2} \sqrt{x} \right] \le e^{-x} \text{ for every } x \ge 0,$$

where $C = C(A_U) > 0$ and $\xi_i = \psi_\tau(\varepsilon_i)$ are as in (A.2). Taking the union bound over all vectors $\boldsymbol{u} \in \mathcal{N}_\epsilon$ with $\epsilon = 1/3$, we obtain that with $\mathbb{P}^*$-probability greater than $1 - e^{2d-x}$,

$$\|\boldsymbol{\xi}^\flat\|_2 \le 1.5C \left\| \sum_{i=1}^n \xi_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i^\intercal \right\|_2^{1/2} \sqrt{x}.$$

Reinterpret this inequality to obtain the stated result (A.8). □

Recall the random process $\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) = -\sum_{i=1}^n \ell_\tau'(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}) U_i \boldsymbol{Z}_i$, $\boldsymbol{\theta} \in \mathbb{R}^d$ defined in (2.16). The following lemma gives an upper bound on the local fluctuation $\sup_{\boldsymbol{\theta} \in \Theta_0(r)} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2$ for $r > 0$.

LEMMA A.4. Assume Condition 2.2 holds. For any $x > 0$, it holds with $\mathbb{P}^*$-probability at least $1 - e^{-x}$ that

$$\sup_{\boldsymbol{\theta} \in \Theta_0(r)} \frac{1}{\sqrt{n}} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \leq C M_{n,4}^{1/2} (4d + x)^{1/2} r, \tag{A.9}$$

where $C > 0$ depends only on A_U, and $M_{n,4}$ is given in (A.3).

PROOF. To begin with, note that

$$\sup_{\boldsymbol{\theta} \in \Theta_0(r)} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 = \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \sup_{\boldsymbol{\omega} \in \mathbb{B}^d(r)} \boldsymbol{\omega}^\intercal \{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\}/r,$$

where $\mathbb{B}^d(r) = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 \leq r\}$. Define a new process $\boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega}) = \boldsymbol{\omega}^\intercal \{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\}/(2r\sqrt{n})$ for $\boldsymbol{\theta} \in \Theta_0(r)$ and $\boldsymbol{\omega} \in \mathbb{B}^d(r)$, so that

$$\sup_{\boldsymbol{\theta} \in \Theta_0(r)} \frac{1}{\sqrt{n}} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 = 2 \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \sup_{\boldsymbol{\omega} \in \mathbb{B}^d(r)} \boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{A.10}$$

It is easy to see that $\mathbb{E}^*\{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega})\} = 0$ and

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^\intercal \nabla \boldsymbol{\xi}^\flat(\boldsymbol{\theta})}{2r\sqrt{n}}, \quad \nabla_{\boldsymbol{\omega}} \boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)}{2r\sqrt{n}}.$$

For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, by Hölder's inequality,

$$\begin{aligned}
&\log \mathbb{E}^* \exp\left\{\lambda \frac{(\boldsymbol{u}^\intercal, \boldsymbol{v}^\intercal) \nabla \boldsymbol{\xi}^\flat(\boldsymbol{\theta}, \boldsymbol{\omega})}{\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal, \boldsymbol{v})\|_2}\right\} \\
&\leq \frac{1}{2} \log \mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}} \frac{\boldsymbol{\omega}^\intercal \nabla \boldsymbol{\xi}^\flat(\boldsymbol{\theta}) \boldsymbol{u}}{\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal, \boldsymbol{v})\|_2}\right\} \\
&\quad + \frac{1}{2} \log \mathbb{E}^* \exp\left[\frac{\lambda}{r\sqrt{n}} \frac{\boldsymbol{v}^\intercal \{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\}}{\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal, \boldsymbol{v})\|_2}\right].
\end{aligned} \tag{A.11}$$

Write $\overline{\boldsymbol{u}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}/\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal, \boldsymbol{v})\|_2$ and $\overline{\boldsymbol{v}} = \boldsymbol{v}/\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal, \boldsymbol{v})\|_2$. For the first term on the right-hand side of (A.11), it follows from (2.16) and Con-

dition 2.2 that

$$\begin{aligned}
&\mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}}\frac{\boldsymbol{\omega}^\intercal\nabla\boldsymbol{\xi}^\flat(\boldsymbol{\theta})\boldsymbol{u}}{\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal,\boldsymbol{v})\|_2}\right\}\\
&= \mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}}\sum_{i=1}^n \ell_\tau''(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta})\boldsymbol{\omega}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\overline{\boldsymbol{u}}\,U_i\right\}\\
&= \prod_{i=1}^n \mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}}\ell_\tau''(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta})\boldsymbol{\omega}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\overline{\boldsymbol{u}}\,U_i\right\}\\
&\le \prod_{i=1}^n \exp\left\{\frac{\lambda^2 B_U^2}{2r^2 n}(\boldsymbol{\omega}^\intercal\boldsymbol{Z}_i)^2(\overline{\boldsymbol{u}}^\intercal\boldsymbol{Z}_i)^2\right\}\\
&\le \exp\left(\frac{\lambda^2}{2}B_U^2 M_{n,4}\right) \qquad \text{(A.12)}
\end{aligned}$$

almost surely. For the second term, by the mean value theorem and taking $\boldsymbol{\delta}_r = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/r$, we get

$$\begin{aligned}
&\mathbb{E}^* \exp\left[\frac{\lambda}{r\sqrt{n}}\overline{\boldsymbol{v}}^\intercal\{\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\}\right]\\
&= \prod_{i=1}^n \mathbb{E}^* \exp\left\{\frac{\lambda}{\sqrt{n}} I(|Y_i - \boldsymbol{X}_i^\intercal\widetilde{\boldsymbol{\theta}}| \le \tau)\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r\,U_i\right\}\\
&\le \exp\left(\frac{\lambda^2}{2}B_U^2 M_{n,4}\right) \qquad \text{(A.13)}
\end{aligned}$$

almost surely, where $\widetilde{\boldsymbol{\theta}}$ is a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$, and $M_{n,4}$ is given in (A.3). Putting (A.11), (A.12) and (A.13) together yields

$$\log \mathbb{E}^* \exp\left\{\lambda\frac{(\boldsymbol{u}^\intercal,\boldsymbol{v}^\intercal)\nabla\boldsymbol{\xi}^\flat(\boldsymbol{\theta},\boldsymbol{\omega})}{\|((\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\intercal,\boldsymbol{v})\|_2}\right\} \le \frac{\lambda^2}{2}B_U^2 M_{n,4}.$$

Applying a conditional version of Theorem A.1 in Spokoiny (2013) with $p = 2d$,

$$H_0 = \begin{pmatrix} \boldsymbol{\Sigma}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}, \quad \mathrm{g} = \infty \ \text{ and } \ \nu_0^2 = B_U^2 M_{n,4}$$

to the process $\{\boldsymbol{\xi}^\flat(\boldsymbol{\theta},\boldsymbol{\omega}) : \boldsymbol{\theta} \in \Theta_0(r), \boldsymbol{\omega} \in \mathbb{B}^d(r)\}$ in (A.10), we arrive at

$$\mathbb{P}^*\left\{\sup_{\boldsymbol{\theta}\in\Theta_0(r)}\frac{1}{\sqrt{n}}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \ge 6B_U M_{n,4}^{1/2}(8d + 2x)^{1/2}r\right\} \le e^{-x}$$

almost surely. This is the bound stated in (A.9). □

The following lemma provides moderate deviation results for the robust estimators $\widehat{\mu}_k$ given in (4.1).

LEMMA A.5. Assume Condition 4.1 holds. Let $\{a_n\}_{n\geq 1}$ be a sequence of positive numbers satisfying $a_n \to \infty$ and $a_n = o(n^{1/2})$ as $n \to \infty$. For each $1 \leq k \leq m$, the robust estimator $\widehat{\mu}_k$ with $\tau_k = v_k\{n/(s+a_n)\}^{1/3}$ for some $v_k \geq v_{k,4}^{1/4}$ satisfies

$$\frac{\mathbb{P}(\sqrt{n}\,|\widehat{\mu}_k - \mu_k| \geq z)}{2\{1-\Phi(z/\sigma_k)\}} \to 1 \tag{A.14}$$

uniformly for $0 \leq z \leq o\{\sigma_k \min(n^{1/6}, \sqrt{n}a_n^{-1})\}$ and $z \leq \sigma_k\sqrt{a_n}$.

PROOF. Let $1 \leq k \leq m$ be fixed and write $\tau = \tau_k$ for simplicity. Define truncated mean and variance of ε_k by $m_{k,\tau} = \mathbb{E}\{\psi_\tau(\varepsilon_k)\}$ and $\sigma_{k,\tau}^2 = \mathbb{E}\{\psi_\tau^2(\varepsilon_k)\}$. Moreover, define

$$T_k = \sum_{i=1}^n \psi_\tau(\varepsilon_{ik}) \quad \text{and} \quad T_{0k} = \sum_{i=1}^n \{\psi_\tau(\varepsilon_{ik}) - m_{k,\tau}\}.$$

Taking $\boldsymbol{X} = (1, \boldsymbol{x}^\intercal)^\intercal$, $\boldsymbol{\theta}^* = (\mu_k, \boldsymbol{\beta}_k^\intercal)^\intercal$ and $\varepsilon = \varepsilon_k$ in Theorem 2.1, we obtain that with probability at least $1 - 4e^{-a_n}$,

$$|\sqrt{n}\,(\widehat{\mu}_k - \mu_k) - T_k/\sqrt{n}| \leq C_1 v_k (s+a_n) n^{-1/2} \tag{A.15}$$

as long as $n \geq C_2(s+a_n)$. We prove (A.14) by considering the following two cases.

CASE 1: Assume $0 \leq z/\sigma_k \leq 1$. Applying Theorem 2.2 in Zhou et al. (2018) to T_k gives

$$\sup_{x\in\mathbb{R}} |\mathbb{P}(T_k/\sqrt{n} \leq x) - \Phi(x/\sigma_k)| \leq C\left(\frac{v_{k,3}}{\sigma_k^3\sqrt{n}} + \frac{v_{k,4}}{\sigma_k^2\tau^2} + \frac{v_{k,4}\sqrt{n}}{\sigma_k\tau^3}\right),$$

where $C > 0$ is an absolute constant. The stated result (A.14) then holds uniformly for $0 \leq z \leq \sigma_k$.

CASE 2: Assume $1 \leq z/\sigma_k \leq \sqrt{a_n}$. It follows from Proposition A.2 with $\kappa = 4$ in the supplement of Zhou et al. (2018) that $|m_{k,\tau}| \leq v_{k,4}\tau^{-3}$. Together with (A.15), this implies that with probability at least $1 - 4e^{-a_n}$,

$$\begin{aligned} &|\sqrt{n}\,(\widehat{\mu}_k - \mu_k) - T_{0k}/\sqrt{n}| \\ &\leq \delta_1 := C_1 v_k (s+a_n) n^{-1/2} + v_{k,4}\tau^{-3}\sqrt{n}. \end{aligned} \tag{A.16}$$

It follows that

$$\begin{aligned}&\mathbb{P}(|T_{0k}|/\sqrt{n} \geq z + \delta_1) - 4e^{-a_n} \\ &\leq \mathbb{P}(\sqrt{n}\,|\widehat{\mu}_k - \mu_k| \geq z) \leq \mathbb{P}(|T_{0k}|/\sqrt{n} \geq z - \delta_1) + 4e^{-a_n}.\end{aligned} \tag{A.17}$$

Next we focus on T_{0k}. Recall that $\tau = \upsilon_k\{n/(s+a_n)\}^{1/3}$ with $\upsilon_k \geq \upsilon_{k,4}^{1/4}$. To apply Lemma 3.1 in the supplement of Liu and Shao (2014), we take

$$d = 1, \quad B_n = \sigma_k^2 n, \quad c_n = \frac{\tau}{\sigma_k\sqrt{n}} = \frac{\upsilon_k}{\sigma_k(s+a_n)^{1/3}n^{1/6}}$$

and note that

$$\left|\frac{\operatorname{cov}(T_{0k})}{B_n} - 1\right| \leq \frac{|\sigma_{k,\tau}^2 - \sigma_k^2| + m_{k,\tau}^2}{\sigma_k^2} \leq b_n := \frac{\upsilon_{k,4}}{\sigma_k^2\tau^2} + \frac{\sigma_k^2}{\tau^2},$$

$$\beta_n := \frac{1}{B_n^{3/2}}\sum_{i=1}^n \mathbb{E}|\psi_\tau(\varepsilon_{ik}) - m_{k,\tau}|^3 \leq C_2\frac{\upsilon_{k,3}}{\sigma_k^3\sqrt{n}},$$

where $C_2 > 0$ is an absolute constant. Consequently, taking $d_n = n^{-1/6}$, $x = \sqrt{n}z$ and $t_n = (C_{3,1}^{-1/2} \vee 4)\{z/\sigma_k + (\log n)^{1/2}\}$ in Lemma 3.1 implies that for all sufficiently large n,

$$\begin{aligned}&|\mathbb{P}(|T_{0k}|/\sqrt{n} \geq z) - \mathbb{P}(|Z| \geq z/\sigma_k)| \\ &\leq C_3\{\beta_n t_n^3 + n^{-1/6}(1 + z/\sigma_k)\}\mathbb{P}(|Z| \geq z/\sigma_k) \\ &\quad + 7n^{-1}e^{-(z/\sigma_k)^2} + 9e^{-c_1 n^{1/3}}\end{aligned} \tag{A.18}$$

uniformly over $0 \leq z/\sigma_k \leq c_2\min(c_n^{-1}, \beta_n^{-1/3}, d_n^{-1})$, where $Z \sim \mathcal{N}(0,1)$, $c_1 > 0$ depends only on $(\sigma_k, \upsilon_{k,3})$ and $C_3, c_2 > 0$ are absolute constants. For normal distribution, it is known that for any $w > 0$,

$$\frac{w}{1+w^2}\frac{1}{\sqrt{2\pi}}e^{-w^2/2} \leq 1 - \Phi(w) \leq \min\left(\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right)e^{-w^2/2}.$$

Combining this with (A.18), we obtain that for $z > \sigma_k$ and δ_1 in (A.16),

$$\begin{aligned}&|\mathbb{P}(\sigma_k|Z| \geq z - \delta_1) - \mathbb{P}(|Z| \geq z/\sigma_k)| \\ &\leq \frac{2\delta_1}{\sqrt{2\pi}}e^{-(z-\delta_1)^2/(2\sigma_k^2)} \leq \delta_1(1 + z/\sigma_k)e^{\delta_1 z/\sigma_k^2}\,\mathbb{P}(|Z| \geq z/\sigma_k)\end{aligned} \tag{A.19}$$

and

$$\begin{aligned}&|\mathbb{P}(\sigma_k|Z| \geq z + \delta_1) - \mathbb{P}(|Z| \geq z/\sigma_k)| \\ &\leq \frac{2\delta_1}{\sqrt{2\pi}}e^{-z^2/(2\sigma_k^2)} \leq \delta_1(1 + z/\sigma_k)\,\mathbb{P}(|Z| \geq z/\sigma_k).\end{aligned} \tag{A.20}$$

Finally, observe that $e^{-a_n} \le e^{-a_n/2-(z/\sigma_k)^2/2}$ for $z/\sigma_k \le \sqrt{a_n}$. Then it follows from (A.16)–(A.20) that (A.14) holds uniformly for $1 \le z/\sigma_k \le o\{\min(n^{1/6}, \sqrt{n}a_n^{-1})\}$, which completes the proof. □

APPENDIX B: PROOFS FOR SECTION 2

Without loss of generality, we assume $t \ge \log 2$, or equivalently $2e^{-t} \le 1$ throughout the proof; otherwise if $2e^{-t} > 1$, the conclusion is trivial. Let $\|\cdot\|_{\boldsymbol{\Sigma},2}$ denote the rescaled ℓ_2-norm on $\mathbb{R}^d$, i.e. $\|\boldsymbol{u}\|_{\boldsymbol{\Sigma},2} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u} \in \mathbb{R}^d$.

B.1. Proof of Theorem 2.1. Proof of (2.4). To begin with, define the parameter set

$$\Theta_0(r) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} \le r\}, \quad r > 0. \tag{B.1}$$

For any prespecified $r > 0$, we can find an intermediate estimator $\widehat{\boldsymbol{\theta}}_{\tau,\eta} = \boldsymbol{\theta}^* + \eta(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)$ for some $\eta \in [0,1]$, satisfying $w(\eta) := \|\widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} \le r$. In fact, if $\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r)$, we can simply take $\eta = 1$; otherwise, since the function $w(\cdot) : [0,1] \mapsto (0,\infty)$ is continuous with $w(0) = 0$ and $w(1) > r$, there always exists some $\eta \in (0,1)$ such that $w(\eta) = r$. Applying Lemma F.2 in Fan et al. (2018) to the loss function $\overline{\mathcal{L}}_\tau := (1/n)\mathcal{L}_\tau$, we obtain

$$\begin{aligned}\langle \nabla\overline{\mathcal{L}}_\tau(\widehat{\boldsymbol{\theta}}_{\tau,\eta}) - \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\rangle &\le \eta\langle \nabla\overline{\mathcal{L}}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\rangle \\ &= -\eta\langle \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\rangle, \end{aligned} \tag{B.2}$$

where the last step uses the first order condition $\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{0}$.

In what follows, we bound the two sides of (B.2) separately. Proposition B.1 below shows that $\overline{\mathcal{L}}_\tau$ is strongly convex on $\Theta_0(r)$ with high probability.

Proposition B.1. Assume that kurtosises of the linear forms $\langle \boldsymbol{u}, \boldsymbol{Z}\rangle$ are uniformly bounded by κ^4 for some $\kappa > 0$, i.e. $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{Z}\rangle^4 \le \kappa^4\|\boldsymbol{u}\|_2^4$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Let (τ, r) and (n, d, t) satisfy

$$\tau \ge 2\max\left\{(4\upsilon_{2+\delta})^{1/(2+\delta)}, 4\kappa^2 r\right\} \quad \text{and} \quad n \ge C(\tau/r)^2(d+t), \tag{B.3}$$

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

$$\langle \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \ge \frac{1}{4}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^2 \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta_0(r). \tag{B.4}$$

By construction, $\widehat{\boldsymbol{\theta}}_{\tau,\eta} \in \Theta_0(r)$ and therefore under the scaling (B.3),

$$\langle \overline{\mathcal{L}}_\tau(\widehat{\boldsymbol{\theta}}_{\tau,\eta}) - \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\rangle \ge \frac{1}{4}\|\widehat{\boldsymbol{\theta}}_{\tau,\eta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^2 \tag{B.5}$$

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

Next we bound the quadratic form $\|\boldsymbol{\Sigma}^{-1/2}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2$. Define the centered random vector $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1/2}\{\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*) - \mathbb{E}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\}$ so that

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2 \le \|\boldsymbol{\gamma}\|_2 + \|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2. \tag{B.6}$$

To bound $\|\boldsymbol{\gamma}\|_2$, by a standard covering argument, there exits an ϵ-net $\mathcal{N}_\epsilon$ $(0 < \epsilon < 1)$ of $\mathbb{S}^{d-1}$ with $|\mathcal{N}_\epsilon| \le (1 + 2/\epsilon)^d$ such that $\|\boldsymbol{\gamma}\|_2 \le (1 - \epsilon)^{-1} \max_{\boldsymbol{u}\in\mathcal{N}_\epsilon} \boldsymbol{u}^\intercal\boldsymbol{\gamma}$. For every $\boldsymbol{u} \in \mathbb{S}^{d-1}$, note that $\boldsymbol{u}^\intercal\boldsymbol{\gamma} = (1/n)\sum_{i=1}^n\{\xi_i\boldsymbol{u}^\intercal\boldsymbol{Z}_i - \mathbb{E}\xi_i\boldsymbol{u}^\intercal\boldsymbol{Z}_i\}$, where $\xi_i = \ell'_\tau(\varepsilon_i)$ and $\boldsymbol{Z}_i$ are IID from $\boldsymbol{Z}$ given in (2.3). Since $\boldsymbol{u}^\intercal\boldsymbol{Z}$ is sub-Gaussian, it follows from the proof of Proposition 2.5.2 in Vershynin (2018) that

$$\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{Z}|^k \le A_0^k k\Gamma(k/2) \quad \text{for all } k \ge 2, \tag{B.7}$$

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

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

By the above calculations, we obtain

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

Applying Bernstein's inequality, we see that

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

Taking the union bound over all vectors $\boldsymbol{u} \in \mathcal{N}_\epsilon$, we obtain that with probability at least $1 - e^{d\log(1+2/\epsilon)-x}$, $\|\boldsymbol{\gamma}\|_2 \le (1 - \epsilon)^{-1}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\boldsymbol{u}^\intercal\boldsymbol{\gamma} \le (1 - \epsilon)^{-1}\{2\sigma A_0\sqrt{x/n} + A_0\tau x/(2n)\}$. Taking $x = \log(1 + 2/\epsilon)(d + t)$ with $\epsilon = 0.27$ yields

$$\mathbb{P}\left\{\|\boldsymbol{\gamma}\|_2 \ge 4\sigma A_0\sqrt{\frac{d+t}{n}} + 1.5A_0\tau\frac{d+t}{n}\right\} \le e^{-2t}. \tag{B.8}$$

For the second term $\|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2$ in (B.6), it holds

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbb{E}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^n\mathbb{E}(\xi_i\boldsymbol{u}^\intercal\boldsymbol{Z}_i) \le \frac{\upsilon_{2+\delta}}{\tau^{1+\delta}}.$$

Together, the last two displays imply with probability at least $1 - e^{-t}$,

$$\begin{aligned} &\|\boldsymbol{\Sigma}^{-1/2}\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2 \\ &\le r_0 := \frac{\upsilon_{2+\delta}}{\tau^{1+\delta}} + 4\sigma A_0\sqrt{\frac{d+t}{n}} + 1.5A_0\tau\frac{d+t}{n}. \end{aligned} \tag{B.9}$$

Finally, in view of (B.3), (B.5) and (B.9), we choose $r = \tau/(4\kappa^2)$. Under Condition 2.1, κ scales as A_0. Then with probability at least $1-2e^{-t}$, $\widehat{\boldsymbol{\theta}}_{\tau,\eta} \in \Theta_0(4r_0)$ under the scaling (B.3). Provided $n \gtrsim A_0^4(d+t)$, we have $r > 4r_0$ so that $\widehat{\boldsymbol{\theta}}_{\tau,\eta}$ lies in the interior of $\Theta_0(r)$, which enforces $\eta = 1$ and $\widehat{\boldsymbol{\theta}}_{\tau,\eta} = \widehat{\boldsymbol{\theta}}_\tau$ (otherwise $\widehat{\boldsymbol{\theta}}_{\tau,\eta}$ will lie on the boundary). Putting together the pieces, we arrive at (2.4).

Proof of (2.5). Next we prove (2.5). Define random processes

$$\boldsymbol{B}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \nabla\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{B.10}$$

and $\boldsymbol{\zeta}(\boldsymbol{\theta}) = \overline{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta})$. In this notation, we have

$$\begin{aligned} &\boldsymbol{B}(\boldsymbol{\theta}) \\ &= \boldsymbol{\Sigma}^{-1/2}\{\nabla\boldsymbol{\zeta}(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*) + \nabla\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \nabla\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*) - \boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\} \\ &\text{and}\ \ \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\} = \boldsymbol{\Sigma}^{-1/2}\{\nabla\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}) - \nabla\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \end{aligned}$$

In the following, we will deal with $\boldsymbol{B}(\boldsymbol{\theta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\}$ and $\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\}$ separately, starting with the latter. By the mean value theorem for vector-valued functions (see, e.g. Theorem 12 in Section 2 of Pugh (2015)),

$$\begin{aligned} \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\} &= \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\int_0^1 \nabla^2\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}_t^*)\,dt(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \left\{\boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}_t^*)\,dt\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d\right\}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \end{aligned}$$

where $\boldsymbol{\theta}_t^* = (1-t)\boldsymbol{\theta}^* + t\boldsymbol{\theta}$. Note that

$$\nabla^2\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}_t^*) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{1/2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\{I(|Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}_t^*| > \tau)\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\}\boldsymbol{\Sigma}^{1/2}.$$

For every $t \in [0,1]$, since $\boldsymbol{\theta} \in \Theta_0(r)$ and $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have $\boldsymbol{\theta}_t^* \in \Theta_0(r)$ so that $\boldsymbol{\delta}_t := \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_t^* - \boldsymbol{\theta}^*)$ satisfies $\|\boldsymbol{\delta}_t\|_2 \le r$. Consequently, by Markov's

inequality and (B.7),

$$\begin{aligned}
&|\boldsymbol{u}^\intercal\{\boldsymbol{\Sigma}^{-1/2}\nabla^2\mathbb{E}\overline{\mathcal{L}}_\tau(\boldsymbol{\theta}_t^*)\boldsymbol{\Sigma}^{-1/2}-\mathbf{I}_d\}\boldsymbol{u}|\\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\{I(|Y_i-\boldsymbol{X}_i^\intercal\boldsymbol{\theta}_t^*|>\tau)(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2\}\\
&\le \upsilon_{2+\delta}(2/\tau)^{2+\delta}+2(\mathbb{E}|\boldsymbol{\delta}_t^\intercal\boldsymbol{Z}|^2)^{1/2}(\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{Z}|^4)^{1/2}\tau^{-1}\\
&\le \delta(r):=2^{2+\delta}\upsilon_{2+\delta}\tau^{-2-\delta}+4\kappa^2\tau^{-1}r,
\end{aligned}$$

where $\kappa>0$ is as in Proposition B.1. Putting together the pieces implies

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)}\|\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\}\|_2\le\delta(r)r. \tag{B.11}$$

Turing to $\boldsymbol{B}(\boldsymbol{\theta})-\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\}=\boldsymbol{\Sigma}^{-1/2}\{\nabla\boldsymbol{\zeta}(\boldsymbol{\theta})-\nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)\}$, we set

$$\overline{\boldsymbol{B}}(\boldsymbol{\delta})=\boldsymbol{B}(\boldsymbol{\theta})-\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\theta})\}\text{ with }\boldsymbol{\delta}=\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*).$$

It is easy to see that $\overline{\boldsymbol{B}}(\mathbf{0})=\mathbf{0}$, $\mathbb{E}\{\overline{\boldsymbol{B}}(\boldsymbol{\delta})\}=\mathbf{0}$ and

$$\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})=\frac{1}{n}\sum_{i=1}^n[\ell_\tau''(\varepsilon_i-\boldsymbol{Z}_i^\intercal\boldsymbol{\delta})\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal-\mathbb{E}\{\ell_\tau''(\varepsilon_i-\boldsymbol{Z}_i^\intercal\boldsymbol{\delta})\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\}].$$

In addition, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}$ and $\lambda\in\mathbb{R}$, using the inequality $|e^z-1-z|\le z^2e^{|z|}/2$ for all $z\in\mathbb{R}$ gives

$$\begin{aligned}
&\mathbb{E}\exp\{\lambda\sqrt{n}\,\boldsymbol{u}^\intercal\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\}\\
&\le\prod_{i=1}^n\left[1+\frac{\lambda^2}{n}\mathbb{E}\{(\boldsymbol{u}^\intercal\boldsymbol{Z}_i\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2+(\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{Z}\boldsymbol{v}^\intercal\boldsymbol{Z}|)^2\}e^{\frac{|\lambda|}{\sqrt{n}}(|\boldsymbol{u}^\intercal\boldsymbol{Z}_i\boldsymbol{v}^\intercal\boldsymbol{Z}_i|+\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{Z}\boldsymbol{v}^\intercal\boldsymbol{Z}|)}\right].
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\mathbb{E}|\boldsymbol{u}^\intercal\boldsymbol{Z}\boldsymbol{v}^\intercal\boldsymbol{Z}|\le\{\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z})^2\}^{1/2}\{\mathbb{E}(\boldsymbol{v}^\intercal\boldsymbol{Z})^2\}^{1/2}=1$$

and

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z}_i\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2e^{\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\intercal\boldsymbol{Z}_i\boldsymbol{v}^\intercal\boldsymbol{Z}_i|}\\
&\le\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z}_i\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2e^{\frac{|\lambda|}{2\sqrt{n}}(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2+\frac{|\lambda|}{2\sqrt{n}}(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2}\\
&\le\{\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^4e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2}\}^{1/2}\{\mathbb{E}(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^4e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2}\}^{1/2}.
\end{aligned}$$

Combining the last three displays, we arrive at

$$\begin{aligned}
&\mathbb{E}\exp\{\lambda\sqrt{n}\,\boldsymbol{u}^\top\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\}\\
&\le\prod_{i=1}^n\bigg[1+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\mathbb{E}(e^{\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{Z}_i\boldsymbol{v}^\top\boldsymbol{Z}_i|})\\
&\qquad\qquad+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\mathbb{E}\{(\boldsymbol{u}^\top\boldsymbol{Z}_i)^2(\boldsymbol{v}^\top\boldsymbol{Z}_i)^2e^{\frac{|\lambda|}{\sqrt{n}}|\boldsymbol{u}^\top\boldsymbol{Z}_i\boldsymbol{v}^\top\boldsymbol{Z}_i|}\}\bigg]\\
&\le\prod_{i=1}^n\bigg[1+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}\{e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\top\boldsymbol{Z})^2}\}\\
&\qquad\qquad+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}\{(\boldsymbol{u}^\top\boldsymbol{Z})^4e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\top\boldsymbol{Z})^2}\}\bigg]\\
&\le\exp\bigg[e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}\{e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\top\boldsymbol{Z})^2}\}+e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}\{(\boldsymbol{u}^\top\boldsymbol{Z})^4e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\top\boldsymbol{Z})^2}\}\bigg].
\end{aligned}$$

Under Condition 2.1, there exists a constant $A_1 = A_1(A_0) > 0$ such that, for all $|\lambda| \le \sqrt{n}/A_1$ and $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}}\mathbb{E}\exp\{\lambda\sqrt{n}\,\boldsymbol{u}^\top\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\}\le\exp(C^2\lambda^2/2), \tag{B.12}$$

where $C > 0$ is an absolute constant. With the above preparations, applying Theorem A.3 in Spokoiny (2013) which is a direct consequence of Corollary 2.2 in the supplement to Spokoiny (2012), yields

$$\mathbb{P}\bigg\{\sup_{\boldsymbol{\delta}\in\mathbb{B}^d(r)}\|\sqrt{n}\,\overline{\boldsymbol{B}}(\boldsymbol{\delta})\|_2\ge 6C(8d+2t)^{1/2}r\bigg\}\le e^{-t}$$

as long as $n \ge A_1^2(8d + 2t)$. Combining this and (B.11), we reach

$$\Delta(r) := \sup_{\boldsymbol{\theta}\in\Theta_0(r)}\|\boldsymbol{B}(\boldsymbol{\theta})\|_2\le\delta(r)r+6C(8d+2t)^{1/2}n^{-1/2}r \tag{B.13}$$

with probability at least $1-e^{-t}$, where $\boldsymbol{B}(\boldsymbol{\theta})$ is given in (B.10). Recalling the paragraph below (B.9), we have $\mathbb{P}\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(4r_0)\} \ge 1-2e^{-t}$ and $\nabla\overline{\mathcal{L}}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \boldsymbol{0}$. On the event $\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(4r_0)\}$, it holds $\|\boldsymbol{B}(\widehat{\boldsymbol{\theta}}_\tau)\|_2 \le \Delta(4r_0)$. Consequently, taking $r = 4r_0$ in (B.13) proves (2.5). □

B.2. Proof of Theorem 2.2. Keeping the notation appeared in the proof of Theorem 2.1, we consider the following local quadratic approximation of the Huber loss. For any $r > 0$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)$, define

$$R(\boldsymbol{\theta},\boldsymbol{\theta}') = \mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\boldsymbol{\theta}') - (\boldsymbol{\theta}-\boldsymbol{\theta}')^\top\nabla\mathcal{L}_\tau(\boldsymbol{\theta}') - \frac{n}{2}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}')\|_2^2.$$

Taking the gradient with respect to $\boldsymbol{\theta}$, we get $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\theta}') = \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}') - n\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}')$. Then, by the mean value theorem, $R(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\intercal \{\nabla \mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}') - n\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}')\}$, where $\widetilde{\boldsymbol{\theta}}$ is a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ and hence $\widetilde{\boldsymbol{\theta}} \in \Theta_0(r)$. It follows that

$$
\begin{aligned}
&|R(\boldsymbol{\theta}, \boldsymbol{\theta}')| \\
&\le n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2 \\
&\quad \times \sup_{\boldsymbol{\theta}'' \in \Theta_0(r)} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}'' - \boldsymbol{\theta}') - \boldsymbol{\Sigma}^{-1/2} n^{-1} \{\nabla \mathcal{L}_\tau(\boldsymbol{\theta}'') - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}')\}\|_2 \\
&\le 2n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2 \times \Delta(r), \qquad \text{(B.14)}
\end{aligned}
$$

where $\Delta(r)$ is as in (B.13). Recall from the proof of Theorem 2.1 that $\mathbb{P}\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(4r_0)\} \ge 1 - 2e^{-t}$ for r_0 given in (B.9). Taking $r = 4r_0$ in (B.13), $(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_\tau)$ in (B.14) and using the fact $\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{0}$, we obtain that with probability greater than $1 - 3e^{-t}$,

$$
\left| \mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \frac{n}{2}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_\tau)\|_2^2 \right| \le 8nr_0\Delta(4r_0).
$$

Write $\widehat{\boldsymbol{\delta}} = \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*)$ and $\boldsymbol{\Gamma}^* = \boldsymbol{\Sigma}^{-1/2} \nabla \overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)$. By (B.9) and (B.13), we have $\|\boldsymbol{\Gamma}^*\|_2 \le r_0$, $\|\widehat{\boldsymbol{\delta}} + \boldsymbol{\Gamma}^*\|_2 \le \Delta(4r_0)$ and

$$
|\|\widehat{\boldsymbol{\delta}}\|_2^2 - \|\boldsymbol{\Gamma}^*\|_2^2| \le \|\widehat{\boldsymbol{\delta}} + \boldsymbol{\Gamma}^*\|_2^2 + 2\|\boldsymbol{\Gamma}^*\|_2 \|\widehat{\boldsymbol{\delta}} + \boldsymbol{\Gamma}^*\|_2 \le \Delta(4r_0)\{\Delta(4r_0) + 2r_0\}.
$$

Together, the last two displays imply that with probability at least $1 - 3e^{-t}$,

$$
\left| \mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) - \frac{n}{2}\|\boldsymbol{\Sigma}^{-1/2} \nabla \overline{\mathcal{L}}_\tau(\boldsymbol{\theta}^*)\|_2^2 \right| \le 9nr_0\Delta(4r_0) + \frac{n}{2}\Delta^2(4r_0),
$$

which, together with (B.13), proves (2.6)

For the square-root Wilks' expansion, on $\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(4r_0)\}$ it holds

$$
\begin{aligned}
&\left| \sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}} - \sqrt{n}\,\|\widehat{\boldsymbol{\delta}}\|_2 \right| \\
&\le \frac{|2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\} - n\|\widehat{\boldsymbol{\delta}}\|_2^2|}{\sqrt{n}\,\|\widehat{\boldsymbol{\delta}}\|_2} = \frac{2R(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_\tau)}{\sqrt{n}\,\|\widehat{\boldsymbol{\delta}}\|_2} \le 4\sqrt{n}\,\Delta(4r_0),
\end{aligned}
$$

where the last step follows from (B.14). Moreover, note that

$$
|\|\widehat{\boldsymbol{\delta}}\|_2 - \|\boldsymbol{\Gamma}^*\|_2| \le \|\widehat{\boldsymbol{\delta}} + \boldsymbol{\Gamma}^*\|_2 \le \Delta(4r_0).
$$

Combining the last two displays with (B.13) proves (2.7). □

B.3. Proof of Proposition B.1. Since the Huber loss is convex and differentiable, we have

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

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

$$\mathcal{E}_i := \big\{ |\varepsilon_i| \leq \tau/2 \big\} \cap \big\{ |\langle \boldsymbol{X}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle| \leq (\tau/2r) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} \big\},$$

on which $|Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta}| \leq \tau$ for all $\boldsymbol{\theta} \in \Theta_0(r)$. Also, recall that $\ell''_\tau(u) = 1$ for $|u| \leq \tau$. For any $R > 0$, define functions

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

In particular, φ_R is R-Lipschitz and satisfies

$$u^2 I(|u| \leq R/2) \leq \varphi_R(u) \leq u^2 I(|u| \leq R). \tag{B.16}$$

It then follows that

$$\mathcal{T}(\boldsymbol{\theta}) \geq g(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n \varphi_{\tau\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{X}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle) \psi_{\tau/2}(\varepsilon_i). \tag{B.17}$$

To bound the right-hand side of (B.17), consider the supremum of a random process indexed by $\Theta_0(r)$:

$$\Delta_r := \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \frac{|g(\boldsymbol{\theta}) - \mathbb{E}g(\boldsymbol{\theta})|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma},2}}. \tag{B.18}$$

For any $\boldsymbol{\theta}$ fixed, write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. By (B.16),

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

Provided $\tau \geq 2\max\{(4\upsilon_{2+\delta})^{1/(2+\delta)}, 4\kappa^2 r\}$, it follows that

$$\mathbb{E}g(\boldsymbol{\theta}) \geq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2 - \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2\left(\frac{16\kappa^4 r^2}{\tau^2} + \frac{2^{2+\delta}\upsilon_{2+\delta}}{\tau^{2+\delta}}\right) \geq \frac{1}{2}\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2 \tag{B.19}$$

for all $\boldsymbol{\theta} \in \Theta_0(r)$. From (B.17)–(B.19), we conclude that

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^2} \geq \frac{1}{2} - \Delta_r \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta_0(r). \tag{B.20}$$

Next we deal with the stochastic term Δ_r defined in (B.18). For $g(\boldsymbol{\theta})$ given in (B.17), we write $g(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n g_i(\boldsymbol{\theta})$. Recalling that $0 \leq \varphi_R(u) \leq R^2/4$ and $0 \leq \psi_R(u) \leq 1$ for all $u \in \mathbb{R}$, it is easy to see that $0 \leq g_i(\boldsymbol{\theta}) \leq (\tau/4r)^2\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^2$. By Talagrand's inequality (see, e.g. Theorem 7.3 in Bousquet (2003)), we have for any $x > 0$,

$$\Delta_r \leq \mathbb{E}\Delta_r + (\mathbb{E}\Delta_r)^{1/2}\frac{\tau}{2r}\sqrt{\frac{x}{n}} + \sigma_n\sqrt{\frac{2x}{n}} + + \frac{\tau^2}{16r^2} \times \frac{x}{3n} \tag{B.21}$$

where $\sigma_n^2 = \sup_{\boldsymbol{\theta}\in\Theta_0(r)} \mathbb{E}g_i^2(\boldsymbol{\theta})/\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^4$. By (B.16), $\mathbb{E}g_i^2(\boldsymbol{\theta}) \leq \mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle^4 \leq \kappa^4\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^4$, implying $\sigma_n \leq \kappa^2$.

To bound the expectation $\mathbb{E}\Delta_r$, applying the symmetrization inequality for empirical processes, and by the connection between Gaussian and Rademacher complexities, we have $\mathbb{E}\Delta_r \leq 2\sqrt{\pi/2}\,\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\mathbb{G}_{\boldsymbol{\theta}}|\}$, where

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

and g_i are IID standard normal random variables that are independent of the observed data. For any $\boldsymbol{\theta}_0 \in \Theta_0(r)$, it holds

$$\mathbb{E}^*\left\{\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\mathbb{G}_{\boldsymbol{\theta}}|\right\} \leq \mathbb{E}^*|\mathbb{G}_{\boldsymbol{\theta}_0}| + 2\mathbb{E}^*\left\{\sup_{\boldsymbol{\theta}\in\Theta_0(r)} \mathbb{G}_{\boldsymbol{\theta}}\right\}, \tag{B.22}$$

where $\mathbb{E}^*$ denotes the conditional expectation given $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Taking the expectation with respect to $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ on both sides, we see that (B.22) remains valid with $\mathbb{E}^*$ replaced by $\mathbb{E}$. To select a proper $\boldsymbol{\theta}_0$, first decompose $\boldsymbol{\theta}^*$ as $(\theta_0, \widetilde{\boldsymbol{\theta}}^{*\intercal})^\intercal$, where θ_0 denotes the first coordinate of $\boldsymbol{\theta}^*$ and $\widetilde{\boldsymbol{\theta}}^* \in \mathbb{R}^{d-1}$ consists of the remaining. Taking $\boldsymbol{\theta}_0 = (\theta_0 + \sigma_{11}^{-1/2}r, \widetilde{\boldsymbol{\theta}}^{*\intercal})^\intercal$, we observe that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} = r$. Since $\varphi_R(u) \leq \min(u^2, R^2/4)$, it holds

$$\mathbb{E}|\mathbb{G}_{\boldsymbol{\theta}_0}| \leq (\mathbb{E}\mathbb{G}_{\boldsymbol{\theta}_0}^2)^{1/2} \leq \frac{\tau}{4r\sqrt{n}}.$$

As in the proof of Lemma 11 in Loh and Wainwright (2015), we next use the Gaussian comparison theorem to bound the expectation of the (conditional) Gaussian supremum $\mathbb{E}^*\{\sup_{\Theta_0(r)} \mathbb{G}_{\boldsymbol{\theta}}\}$.

Let var* be the conditional variance given $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. For $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)$, write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and $\boldsymbol{\delta}' = \boldsymbol{\theta}' - \boldsymbol{\theta}^*$. Then

$$\begin{aligned}
&\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \\
&\le \frac{1}{n^2} \sum_{i=1}^n \psi_{\tau/2}^2(\varepsilon_i) \left\{ \frac{\varphi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}(\boldsymbol{X}_i^\top \boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2} - \frac{\varphi_{\tau\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}/(2r)}(\boldsymbol{X}_i^\top \boldsymbol{\delta}')}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}^2} \right\}^2.
\end{aligned}$$

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

$$\varphi_{\tau\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}/(2r)}(\boldsymbol{X}_i^\top \boldsymbol{\delta}') = \frac{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}^2}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2} \varphi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)} \left(\frac{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}' \right).$$

Putting the above calculations together, we obtain

$$\begin{aligned}
&\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \\
&\le \frac{1}{n^2} \sum_{i=1}^n \frac{1}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^4} \left\{ \varphi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}(\boldsymbol{X}_i^\top \boldsymbol{\delta}) - \varphi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)} \left(\frac{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}} \boldsymbol{X}_i^\top \boldsymbol{\delta}' \right) \right\}^2 \\
&\le \frac{1}{n^2} \sum_{i=1}^n \frac{1}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^4} \frac{\tau^2 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2}{4r^2} \left(\boldsymbol{Z}_i^\top \boldsymbol{\delta} - \frac{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}} \boldsymbol{Z}_i^\top \boldsymbol{\delta}' \right)^2 \\
&\le \frac{1}{n^2} \sum_{i=1}^n \frac{\tau^2}{4r^2} \left(\frac{\boldsymbol{X}_i^\top \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}} - \frac{\boldsymbol{X}_i^\top \boldsymbol{\delta}'}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}} \right)^2. \qquad \text{(B.23)}
\end{aligned}$$

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

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

where g_i' are IID standard normal random variables that are independent of all other random variables. By (B.23), $\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\theta}} - \mathbb{G}_{\boldsymbol{\theta}'}) \le \mathrm{var}^*(\mathbb{Z}_{\boldsymbol{\theta}} - \mathbb{Z}_{\boldsymbol{\theta}'})$. By the Gaussian comparison inequality (Ledoux and Talagrand, 1991),

$$\mathbb{E}^* \left\{ \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \mathbb{G}_{\boldsymbol{\theta}} \right\} \le 2\mathbb{E}^* \left\{ \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \mathbb{Z}_{\boldsymbol{\theta}} \right\} \le \frac{\tau}{r} \mathbb{E}^* \left\| \frac{1}{n} \sum_{i=1}^n g_i \boldsymbol{Z}_i \right\|_2.$$

Together with the unconditional version of (B.22), this implies

$$\mathbb{E}\Delta_r \le \sqrt{2\pi}\left(\frac{2\tau}{r}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i \boldsymbol{Z}_i\right\|_2 + \frac{\tau}{4r\sqrt{n}}\right) \le \sqrt{2\pi}\left(\frac{2\tau}{r}\sqrt{\frac{d}{n}} + \frac{\tau}{4r\sqrt{n}}\right).$$

Combining this with (B.21) and (B.20), we obtain that with probability at least $1 - e^{-t}$,

$$\frac{\mathcal{T}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2}^2} \ge \frac{1}{4} \quad \text{uniformly over } \boldsymbol{\theta} \in \Theta_0(r)$$

for all sufficiently large n that scales as $(\tau/r)^2(d+t)$ up to an absolute constant. This proves (B.4). □

B.4. Proof of Theorem 2.3. Throughout we assume $t \ge 1$ and keep the notation used in the proof of Theorem 2.1.

Proof of (2.14). Recall the weighted loss function $\mathcal{L}_\tau^\flat(\boldsymbol{\theta}) = \sum_{i=1}^n W_i \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and the parameter set $\Theta_0(r)$ defined in (B.1). Write $r_1 = 4r_0$ for r_0 as in (B.9), so that

$$\mathbb{P}\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)\} \ge 1 - 2e^{-t} \quad \text{provided } n \ge C_1(d+t)$$

for some $C_1 = C_1(A_0) > 0$. By the definition of $\widehat{\boldsymbol{\theta}}_\tau^\flat$, $\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) \le 0$ and

$$\|\widehat{\boldsymbol{\theta}}_\tau^\flat - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} \le \|\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau\|_{\boldsymbol{\Sigma},2} + \|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} \le R_1 + \|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2},$$

where $R_1 := \overline{\lambda}_{\boldsymbol{\Sigma}}^{1/2} R$. If we can show that, for some $r_2 \ge r_1$ to be specified,

$$\mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) > 0 \quad \text{for all } \boldsymbol{\theta} \in \partial\Theta_0(r) \text{ with } r_2 \le r \le r_2 + R_1$$

with high probability, then we must have $\widehat{\boldsymbol{\theta}}_\tau^\flat \in \Theta_0(r_2)$ with high probability. Here and below, we set $\partial\Theta_0(r) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma},2} = r\}$.

Centering the weighted Huber loss function, we define

$$\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) = \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \mathbb{E}^*\{\mathcal{L}_\tau^\flat(\boldsymbol{\theta})\} = \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \mathcal{L}_\tau(\boldsymbol{\theta}). \tag{B.24}$$

Note that

$$\begin{aligned}
&\mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) \\
&= \boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau) + \mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\boldsymbol{\theta}^*) + \mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) \\
&\ge \underbrace{\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)}_{\Pi_1(\boldsymbol{\theta})} + \underbrace{\mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\boldsymbol{\theta}^*)}_{\Pi_2(\boldsymbol{\theta})}.
\end{aligned} \tag{B.25}$$

In the following, we bound $\Pi_1(\boldsymbol{\theta})$ and $\Pi_2(\boldsymbol{\theta})$ separately, starting with the latter which only depends on the observed data. As before, define $\boldsymbol{\zeta}(\boldsymbol{\theta}) = \mathcal{L}_\tau(\boldsymbol{\theta}) - \mathbb{E}\{\mathcal{L}_\tau(\boldsymbol{\theta})\}$ and consider the decomposition

$$\begin{aligned}\Pi_2(\boldsymbol{\theta}) &= \underbrace{\boldsymbol{\zeta}(\boldsymbol{\theta}) - \boldsymbol{\zeta}(\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\intercal \nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)}_{\Pi_{21}(\boldsymbol{\theta})} \\ &\quad + \underbrace{(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\intercal \nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)}_{\Pi_{22}(\boldsymbol{\theta})} + \underbrace{\mathbb{E}\{\mathcal{L}_\tau(\boldsymbol{\theta}) - \mathcal{L}_\tau(\boldsymbol{\theta}^*)\}}_{\Pi_{23}(\boldsymbol{\theta})}. \end{aligned} \tag{B.26}$$

First we deal with $\Pi_{21}(\boldsymbol{\theta})$. For every $r > 0$, define the random process

$$U_r(\boldsymbol{\theta}) = \frac{1}{r\sqrt{n}}\{\boldsymbol{\zeta}(\boldsymbol{\theta}) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\intercal\nabla\boldsymbol{\zeta}(\boldsymbol{\theta}^*)\}, \quad \boldsymbol{\theta}\in\Theta_0(r). \tag{B.27}$$

We will use Theorem A.1 in Spokoiny (2013) to bound the local fluctuation $|U_r(\boldsymbol{\theta}) - U_r(\boldsymbol{\theta}^*)|$ over $\boldsymbol{\theta} \in \Theta_0(r)$. For any random variable X, we write $(\mathbb{I}-\mathbb{E})X = X - \mathbb{E}(X)$. For every $\boldsymbol{\theta}\in\Theta_0(r)$, $\boldsymbol{v}\in\mathbb{R}^d$ and $\lambda\in\mathbb{R}$, putting $\overline{\boldsymbol{v}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}/\|\boldsymbol{v}\|_{\boldsymbol{\Sigma},2}$ and $\boldsymbol{\delta}_r = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)/r$, and by the mean value theorem, we have

$$\begin{aligned}&\mathbb{E}\exp\left\{\lambda\frac{\boldsymbol{v}^\intercal\nabla U_r(\boldsymbol{\theta})}{\|\boldsymbol{v}\|_{\boldsymbol{\Sigma},2}}\right\} \\ &= \mathbb{E}\exp\left\{\frac{\lambda}{\sqrt{n}}\sum_{i=1}^n(\mathbb{I}-\mathbb{E})\ell_\tau''(Y_i - \boldsymbol{X}_i^\intercal\widetilde{\boldsymbol{\theta}}_i)\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r\right\} \\ &\le \prod_{i=1}^n\Bigg(1 + \frac{\lambda^2}{n}\mathbb{E}\Big[\{(\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r)^2 + (\mathbb{E}|\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r|)^2\} \\ &\qquad\qquad \times e^{\frac{|\lambda|}{\sqrt{n}}(|\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r| + \mathbb{E}|\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r|)}\Big]\Bigg) \\ &\le \prod_{i=1}^n\left\{1 + \frac{\lambda^2}{n}e^{\frac{|\lambda|}{\sqrt{n}}}\mathbb{E}e^{\frac{|\lambda|}{\sqrt{n}}|\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r|} + \frac{\lambda^2}{n}e^{\frac{|\lambda|}{\sqrt{n}}}\mathbb{E}(\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r)^2 e^{\frac{|\lambda|}{\sqrt{n}}|\overline{\boldsymbol{v}}^\intercal\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\delta}_r|}\right\} \\ &\le \exp\left\{e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\intercal\boldsymbol{Z})^2} + e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\mathbb{E}(\boldsymbol{u}^\intercal\boldsymbol{Z})^4 e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{u}^\intercal\boldsymbol{Z})^2}\right\}.\end{aligned}$$

Similarly to (B.12), it can be shown that for all $|\lambda| \le c_1\sqrt{n}$ and $\boldsymbol{\theta}\in\Theta_0(r)$,

$$\mathbb{E}\exp\left\{\lambda\frac{\boldsymbol{v}^\intercal\nabla U_r(\boldsymbol{\theta})}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2}\right\} \le \exp(C_2^2\lambda^2/2).$$

Using Theorem A.1 in Spokoiny (2013), we deduce that with $\mathbb{P}^\dagger$-probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)}|U_r(\boldsymbol{\theta}) - U_r(\boldsymbol{\theta}^*)| \le 3C_2(4d+2t)^{1/2}r.$$

as long as $n \geq c_1^{-2}(4d+2t)$. In view of (B.26) and (B.27), it holds for every $r > 0$ that

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_{21}(\boldsymbol{\theta})| \leq 3C_2(4d+2t)^{1/2}r^2\sqrt{n} \tag{B.28}$$

with $\mathbb{P}^\dagger$-probability at least $1-e^{-t}$. The bound in (B.28) holds for any given $r > 0$. Following the slicing argument similar to that used in the proof of Theorem A.2 in Spokoiny (2013), it can be shown that with $\mathbb{P}^\dagger$-probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_{21}(\boldsymbol{\theta})| \leq 6C_2\{4d+2t+2\log(2r/r_2)\}^{1/2}r^2\sqrt{n} \tag{B.29}$$

for all $r_2 \leq r \leq r_2 + R_1$ as long as $n \geq c_1^{-2}\{4d+2t+2\log(2+2R_1/r_2)\}$.

For $\Pi_{22}(\boldsymbol{\theta})$, note that

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_{22}(\boldsymbol{\theta})| \leq r\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathbb{E}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)\}\|_2 = rn\|\boldsymbol{\gamma}\|_2$$

for $\boldsymbol{\gamma}$ as given in (B.6). This, together with (B.8), implies that with $\mathbb{P}^\dagger$-probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_{22}(\boldsymbol{\theta})| \leq C_3 v(d+t)^{1/2}r\sqrt{n} \quad \text{for any } r > 0, \tag{B.30}$$

where $C_3 = C_3(A_0) > 0$.

Turning to $\Pi_{23}(\boldsymbol{\theta})$, we define the function $h(\boldsymbol{\theta}) = (1/n)\mathbb{E}\{\mathcal{L}_\tau(\boldsymbol{\theta})\}$, $\boldsymbol{\theta} \in \mathbb{R}^d$ so that $\Pi_{23}(\boldsymbol{\theta}) = n\{h(\boldsymbol{\theta}) - h(\boldsymbol{\theta}^*)\}$. Put $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. By (2.3) and the mean value theorem, it follows that

$$h(\boldsymbol{\theta}) = h(\boldsymbol{\theta}^*) - \mathbb{E}\{\psi_\tau(\varepsilon)\boldsymbol{Z}^\intercal\boldsymbol{\delta}\} + \frac{1}{2}\mathbb{E}\{\ell_\tau''(Y - \boldsymbol{X}^\intercal\widetilde{\boldsymbol{\theta}})(\boldsymbol{Z}^\intercal\boldsymbol{\delta})^2\},$$

where $\widetilde{\boldsymbol{\theta}}$ is a point lying between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Since $\mathbb{E}(\varepsilon|\boldsymbol{Z}) = 0$, we have $-\mathbb{E}\{\psi_\tau(\varepsilon)|\boldsymbol{Z}\} = \mathbb{E}[\{\varepsilon - \tau\mathrm{sgn}(\varepsilon)\}I(|\varepsilon| > \tau)|\boldsymbol{Z}]$, which further implies

$$|\mathbb{E}\{\psi_\tau(\varepsilon)\boldsymbol{Z}^\intercal\boldsymbol{\delta}\}| \leq \upsilon_4\tau^{-3}\mathbb{E}(|\boldsymbol{Z}^\intercal\boldsymbol{\delta}|) \leq \upsilon_4\tau^{-3}\|\boldsymbol{\delta}\|_2.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\{\ell_\tau''(Y - \boldsymbol{X}^\intercal\widetilde{\boldsymbol{\theta}})(\boldsymbol{Z}^\intercal\boldsymbol{\delta})^2\} \\
&= \mathbb{E}(\boldsymbol{Z}^\intercal\boldsymbol{\delta})^2 - \mathbb{E}\{I(|Y - \boldsymbol{X}^\intercal\widetilde{\boldsymbol{\theta}}| > \tau)(\boldsymbol{Z}^\intercal\boldsymbol{\delta})^2\} \\
&\geq (1 - \sigma^2\tau^{-2})\|\boldsymbol{\delta}\|_2^2 - \tau^{-2}\mathbb{E}\{\boldsymbol{X}^\intercal(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}})\}^2(\boldsymbol{Z}^\intercal\boldsymbol{\delta})^2 \\
&\geq (1 - \sigma^2\tau^{-2})\|\boldsymbol{\delta}\|_2^2 - \kappa^4\tau^{-2}\|\boldsymbol{\delta}\|_2^4,
\end{aligned}$$

where $\kappa > 0$ is as in Proposition B.1. Putting the above calculations together yields that for any $\boldsymbol{\theta} \in \partial\Theta_0(r)$,

$$\frac{1}{n}\Pi_{23}(\boldsymbol{\theta}) \geq (1 - \sigma^2\tau^{-2} - \kappa^4\tau^{-2}r^2)\frac{r^2}{2} - \upsilon_4\tau^{-3}r = \frac{1}{2}b(r)r^2, \tag{B.31}$$

where $b(r) := 1 - \sigma^2\tau^{-2} - \kappa^4\tau^{-2}r^2 - 2\upsilon_4\tau^{-3}r^{-1}$, $r > 0$.

Combining (B.26), (B.29), (B.30) and (B.31), it follows that with $\mathbb{P}^\dagger$-probability at least $1 - 2e^{-t}$,

$$\Pi_2(\boldsymbol{\theta}) \geq r^2 n\left\{\frac{1}{2}b(r) - 6C_2\sqrt{\frac{4d + 2t + 2\log(2r/r_2)}{n}} - C_3\frac{v(d+t)^{1/2}}{r\sqrt{n}}\right\} \tag{B.32}$$

for all $\boldsymbol{\theta} \in \partial\Theta_0(r)$ with $r \in [r_2, r_2 + R_1]$ provided $n \geq c_1^{-2}\{4d + 2t + 2\log(2 + 2R_1/r_2)\}$.

Next we deal with the process $\Pi_1(\boldsymbol{\theta}) = \boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)$ in (B.25), where

$$\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta})(W_i - 1), \quad \boldsymbol{\theta} \in \mathbb{R}^d.$$

Decompose $\Pi_1(\boldsymbol{\theta})$ as

$$\Pi_1(\boldsymbol{\theta}) = \underbrace{\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau) - (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)^\intercal\nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)}_{\Pi_{11}(\boldsymbol{\theta})} + \underbrace{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)^\intercal\nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)}_{\Pi_{12}(\boldsymbol{\theta})}. \tag{B.33}$$

We will use a conditional version of Theorem A.1 in Spokoiny (2013) to bound $\Pi_{11}(\boldsymbol{\theta})$. Similarly to (B.27), define, for each $r > 0$,

$$U_r^\flat(\boldsymbol{\theta}) = \frac{1}{r\sqrt{n}}\{\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)^\intercal\nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)\}, \quad \boldsymbol{\theta} \in \mathbb{R}^d. \tag{B.34}$$

Similarly to (B.1), define

$$\widehat{\Theta}_0(r) = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau\|_{\boldsymbol{\Sigma},2} \leq r\}, \quad r > 0. \tag{B.35}$$

For every $\boldsymbol{\theta} \in \widehat{\Theta}_0(r)$ and $\boldsymbol{v} \in \mathbb{R}^d$, by the mean value theorem we have

$$\boldsymbol{v}^\intercal\nabla U_r^\flat(\boldsymbol{\theta}) = \frac{1}{r\sqrt{n}}\boldsymbol{v}^\intercal\{\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)\} = \frac{1}{r\sqrt{n}}\boldsymbol{v}^\intercal\nabla^2\boldsymbol{\zeta}^\flat(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau),$$

where $\widetilde{\boldsymbol{\theta}}$ is a convex combination of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_\tau$. Putting $\overline{\boldsymbol{\delta}}_r = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)/r$ and $\overline{\boldsymbol{v}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2$, we deduce that

$$\begin{aligned}
&\mathbb{E}^* \exp\left\{\lambda \frac{\boldsymbol{v}^\intercal \nabla U_r^\flat(\boldsymbol{\theta})}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2}\right\} \\
&= \mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}} \overline{\boldsymbol{v}}^\intercal \boldsymbol{\Sigma}^{-1/2} \nabla^2 \boldsymbol{\zeta}^\flat(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)\right\} \\
&= \prod_{i=1}^n \mathbb{E}^* \exp\left\{\frac{\lambda}{r\sqrt{n}} \overline{\boldsymbol{v}}^\intercal \boldsymbol{\Sigma}^{-1/2} \nabla^2_{\boldsymbol{\theta}} \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal \widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau) U_i\right\} \\
&= \prod_{i=1}^n \mathbb{E}^* \exp\left\{\frac{\lambda}{\sqrt{n}} \eta_i(\boldsymbol{\theta}, \boldsymbol{v}) U_i\right\}, \qquad \text{(B.36)}
\end{aligned}$$

where

$$\begin{aligned}
&\eta_i(\boldsymbol{\theta}, \boldsymbol{v}) \\
&:= \overline{\boldsymbol{v}}^\intercal \boldsymbol{\Sigma}^{-1/2} \nabla^2_{\boldsymbol{\theta}} \ell_\tau(Y_i - \boldsymbol{X}_i^\intercal \widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau)/r = I(|Y_i - \boldsymbol{X}_i^\intercal \widetilde{\boldsymbol{\theta}}| \le \tau) \overline{\boldsymbol{v}}^\intercal \boldsymbol{Z}_i \boldsymbol{Z}_i^\intercal \overline{\boldsymbol{\delta}}_r.
\end{aligned}$$

Under Condition 2.2, it holds

$$\begin{aligned}
&\mathbb{E}^* \exp\left\{\frac{\lambda}{\sqrt{n}} \eta_i(\boldsymbol{\theta}, \boldsymbol{v}) U_i\right\} \\
&\le \exp\left\{\frac{\lambda^2}{2n} B_U^2 \eta_i^2(\boldsymbol{\theta}, \boldsymbol{v})\right\} \le \exp\left\{\frac{\lambda^2}{2n} B_U^2 (\overline{\boldsymbol{v}}^\intercal \boldsymbol{Z}_i)^2 (\overline{\boldsymbol{\delta}}_r^\intercal \boldsymbol{Z}_i)^2\right\},
\end{aligned}$$

where $B_U = B_U(A_U) > 0$. Plugging this into (B.36) shows

$$\mathbb{E}^* \exp\left\{\lambda \frac{\boldsymbol{v}^\intercal \nabla U_r^\flat(\boldsymbol{\theta})}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2}\right\} \le \exp\left(\frac{\lambda^2}{2} B_U^2 M_{n,4}\right).$$

With the above preparations in place, it follows from (B.34) and Theorem A.1 in Spokoiny (2013) that for any $r > 0$,

$$\mathbb{P}^*\left\{\sup_{\boldsymbol{\theta} \in \widehat{\Theta}_0(r)} |\Pi_{11}(\boldsymbol{\theta})| \ge 3 B_U M_{n,4}^{1/2} (4d + 2t)^{1/2} r^2 \sqrt{n}\right\} \le e^{-t}$$

almost surely, where $M_{n,4}$ is given in (A.3). Again, using the slicing technique and applying the preceding bound to each slice separately, we obtain that with $\mathbb{P}^*$-probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\theta} \in \widehat{\Theta}_0(r)} |\Pi_{11}(\boldsymbol{\theta})| \le 6 B_U M_{n,4}^{1/2} \{4d + 2t + 2\log(2r/r_1)\}^{1/2} r^2 \sqrt{n}$$

for all $r_1 \le r \le 2r_2 + R_1$. Note that, on the event $\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)\}$ that occurs with $\mathbb{P}^\dagger$-probability at least $1 - 2e^{-t}$,

$$\Theta_0(r) \subseteq \widehat{\Theta}_0(r + r_1).$$

Combining the last two displays and taking $x = 2t$ in Lemma A.1, we obtain that with $\mathbb{P}^\dagger$-probability at least $1 - 3e^{-t}$,

$$\mathbb{P}^*\Bigg\{ \sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_{11}(\boldsymbol{\theta})| \le C_4\sqrt{d + t + \log(2 + 2r/r_1)} \times (r + r_1)^2\sqrt{n} \ \text{ for all } r_2 \le r \le r_2 + R_1 \Bigg\} \ge 1 - e^{-t} \tag{B.37}$$

as long as $n \ge C_0(d + t)^2$, where $C_0 = C_0(A_0)$ and $C_4 = C_4(A_0, A_U)$.

For $\Pi_{12}(\boldsymbol{\theta})$ in (B.33), it holds on the event $\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)\}$ that, for every $\boldsymbol{\theta} \in \Theta_0(r)$,

$$\begin{aligned} |\Pi_{12}(\boldsymbol{\theta})| &= |(\boldsymbol{\theta} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_\tau)^\intercal \nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)| \\ &\le (r + r_1)\|\boldsymbol{\Sigma}^{-1/2}\nabla\boldsymbol{\zeta}^\flat(\widehat{\boldsymbol{\theta}}_\tau)\|_2 \\ &\le (r + r_1)\{\|\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 + \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\} \\ &\le (r + r_1)\Bigg\{ \sup_{\boldsymbol{\theta}\in\Theta_0(r_1)} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 + \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \Bigg\}, \end{aligned} \tag{B.38}$$

where $\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1/2}\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) = -\sum_{i=1}^n \psi_\tau(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta})U_i\boldsymbol{Z}_i$ is as in (2.16).

By Lemma A.4, it holds for each $r > 0$ that

$$\mathbb{P}^*\Bigg\{ \sup_{\boldsymbol{\theta}\in\Theta_0(r)} \frac{1}{\sqrt{n}}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \ge CM_{n,4}^{1/2}(8d + 2t)^{1/2}r \Bigg\} \le e^{-t}$$

almost surely. Combining this and Lemma A.1, we see that, conditioning on the same event where (B.37) holds,

$$\mathbb{P}^*\Bigg\{ \sup_{\boldsymbol{\theta}\in\Theta_0(r_1)} \frac{1}{\sqrt{n}}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \ge C_5(d + t)^{1/2}r_1 \Bigg\} \le e^{-t} \tag{B.39}$$

as long as $n \ge C_0(d + t)^2$, where $C_5 = C_5(A_0, A_U) > 0$.

For $\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2$, taking $x = 2t$ in Lemmas A.2 and A.3, we see that with $\mathbb{P}^\dagger$-probability at least $1 - e^{-t}$,

$$\mathbb{P}^*\{\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \ge C_6(d + t)^{1/2}\sqrt{n}\} \le e^{-t}, \tag{B.40}$$

where $C_6 = C_6(A_U) > 0$. Combining (B.33), (B.37), (B.38), (B.39) and (B.40), we conclude that conditioning on some event that occurs with $\mathbb{P}^\dagger$-probability at least $1 - 4e^{-t}$, it holds

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta_0(r)} |\Pi_1(\boldsymbol{\theta})| \\
&\le r^2 n \Bigg[C_4 (r + r_1)^2 \frac{\{d + t + \log(2 + 2r/r_1)\}^{1/2}}{r^2\sqrt{n}} + C_5 (r + r_1) \frac{r_1 (d + t)^{1/2}}{r^2\sqrt{n}} \\
&\qquad + C_6 (r + r_1) \frac{v(d + t)^{1/2}}{r^2\sqrt{n}} \Bigg] \quad \text{for all } r_2 \le r \le r_2 + R_1
\end{aligned} \tag{B.41}$$

with $\mathbb{P}^*$-probability at least $1 - 3e^{-t}$ provided $n \ge C_0(d + t)^2$.

Finally, combining (B.25), (B.32) and (B.41), and taking

$$r_2 = C_7 v \sqrt{(d + t)/n} \ge r_1$$

for some sufficiently large constant $C_7 = C_7(A_0, A_U) > 0$, we conclude that, conditioning on some event that occurs with $\mathbb{P}^\dagger$-probability at least $1 - 5e^{-t}$,

$$\mathbb{P}^* \Big\{ \widehat{\boldsymbol{\theta}}_\tau^\flat \in \Theta_0(r_1 + R_1) \text{ and } \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) > \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau), \forall \boldsymbol{\theta} \in \partial\Theta_0(r), r \in [r_2, r_2 + R_1] \Big\} \ge 1 - 3e^{-t}$$

provided $n \ge C_0(d+t)^2$ and $n \ge C_8 \overline{\lambda}_{\boldsymbol{\Sigma}}$, where $C_8 = C_8(A_0) > 0$. Reinterpret this we obtain (2.14).

Proof of (2.15). An argument similar to that given in the proof of Theorem 2.1 can be used to prove (2.15). Define the random process

$$\begin{aligned}
\boldsymbol{B}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \boldsymbol{\Sigma}^{-1/2} n^{-1} \{\nabla \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau^\flat(\boldsymbol{\theta}')\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}') \\
&= \boldsymbol{\Sigma}^{-1/2} n^{-1} \{\nabla \boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \nabla \boldsymbol{\zeta}^\flat(\boldsymbol{\theta}') + \nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}') - n\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}')\},
\end{aligned}$$

where $\boldsymbol{\zeta}^\flat(\cdot)$ is given in (B.24). The stated result follows from a bound on

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)} \|\boldsymbol{B}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 \\
&\le \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)} \|\boldsymbol{B}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') - \mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')\}\|_2 + \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)} \|\mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')\}\|_2,
\end{aligned}$$

and the facts that $\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)$ and $\widehat{\boldsymbol{\theta}}_\tau^\flat \in \Theta_0(r_2)$ with high probability.

Note that $\mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\} = \boldsymbol{B}(\boldsymbol{\theta}) - \boldsymbol{B}(\boldsymbol{\theta}')$ for $\boldsymbol{B}(\cdot)$ as in (B.10). It then follows from (B.13) that with $\mathbb{P}^\dagger$-probability greater than $1 - e^{-t}$,

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta_0(r)} \|\mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\}\|_2 \le 2\delta(r)r + C_9(d+t)^{1/2}n^{-1/2}r, \tag{B.42}$$

where $\delta(\cdot)$ is defined above (B.11) and $C_9 = C_9(A_0) > 0$.

For $\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}') - \mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\}$, note that

$$\begin{aligned}
&\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}') - \mathbb{E}^*\{\boldsymbol{B}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\} \\
&= \boldsymbol{\Sigma}^{-1/2}n^{-1}\{\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}')\} \\
&= \boldsymbol{\Sigma}^{-1/2}n^{-1}\{\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}^*)\} - \boldsymbol{\Sigma}^{-1/2}n^{-1}\{\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}') - \nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}^*)\}.
\end{aligned}$$

Since we are interested in the case where both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are in a neighborhood of $\boldsymbol{\theta}^*$, it suffices to focus on $\boldsymbol{\theta}$. To proceed, we change the variable by $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and define

$$\begin{aligned}
\overline{\boldsymbol{B}}^\flat(\boldsymbol{\delta}) &= \boldsymbol{\Sigma}^{-1/2}n^{-1}\{\nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}) - \nabla\boldsymbol{\zeta}^\flat(\boldsymbol{\theta}^*)\} \\
&= -\frac{1}{n}\sum_{i=1}^n \{\psi_\tau(\varepsilon_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\delta}) - \psi_\tau(\varepsilon_i)\}U_i\boldsymbol{Z}_i.
\end{aligned}$$

It is easy to see that $\overline{\boldsymbol{B}}^\flat(\boldsymbol{0}) = \boldsymbol{0}$, $\mathbb{E}^*\{\overline{\boldsymbol{B}}^\flat(\boldsymbol{\delta})\} = \boldsymbol{0}$ and

$$\nabla\overline{\boldsymbol{B}}^\flat(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n \ell_\tau''(\varepsilon_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\delta})U_i\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal.$$

For any $\boldsymbol{u},\boldsymbol{v} \in \mathbb{S}^{d-1}$ and $\lambda \in \mathbb{R}$, by Condition 2.2 we have

$$\begin{aligned}
&\mathbb{E}^* \exp\{\lambda\sqrt{n}\,\boldsymbol{u}^\intercal\nabla\overline{\boldsymbol{B}}^\flat(\boldsymbol{\delta})\boldsymbol{v}\} \\
&\le \prod_{i=1}^n \exp\left\{\frac{\lambda^2}{2n}B_U^2(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2\right\} \le \exp\left(\frac{\lambda^2}{2}B_U^2M_{n,4}\right),
\end{aligned}$$

where $B_U = B_U(A_U) > 0$. Applying a conditional version of Theorem A.1 in Spokoiny (2013) delivers

$$\mathbb{P}^*\left\{\sup_{\boldsymbol{\delta}\in\mathbb{B}^d(r)} \|\sqrt{n}\,\overline{\boldsymbol{B}}^\flat(\boldsymbol{\delta})\|_2 \ge 3B_UM_{n,4}^{1/2}(4d+2t)^{1/2}r\right\} \le e^{-t} \tag{B.43}$$

almost surely, where $M_{n,4}$ is given in (A.3).

Finally, we take $(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\widehat{\boldsymbol{\theta}}_\tau^\flat, \widehat{\boldsymbol{\theta}}_\tau)$. By (2.4) and (2.14), $\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)$ with probability at least $1 - 3e^{-t}$, and with $\mathbb{P}^\dagger$-probability at least $1 - 5e^{-t}$,

$$\mathbb{P}^*\{\widehat{\boldsymbol{\theta}}_\tau^\flat \in \Theta_0(r_2)\} \geq 1 - 3e^{-t}.$$

Since $\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{0}$, it holds $\boldsymbol{\Sigma}^{-1/2} \nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) = \boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau)$, where $\boldsymbol{\xi}^\flat(\cdot)$ is given in (2.16). Then, on the event $\{\widehat{\boldsymbol{\theta}}_\tau \in \Theta_0(r_1)\}$, it holds

$$\|\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 \leq \sup_{\boldsymbol{\theta} \in \Theta_0(r_1)} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2,$$

so that the bound in (B.39) can be applied. Moreover, by the triangle inequality, $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\|_2 \leq r_1 + r_2$ with high probability, which in turn implies that $\widehat{\boldsymbol{\theta}}_\tau^\flat$ falls in the interior of $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_\tau\|_2 \leq R\}$ for all sufficiently large n and hence $\nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) = \mathbf{0}$. This, together with (B.42), (B.43) and the definition of $\boldsymbol{B}^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat, \widehat{\boldsymbol{\theta}}_\tau)$, proves (2.15). □

B.5. Proof of Theorem 2.4. The proof is based on a similar argument to that used in the proof of Theorem 2.2. To begin with, define the bootstrap random process: for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)$,

$$\begin{aligned} &R^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') \\ &= \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \mathcal{L}_\tau^\flat(\boldsymbol{\theta}') - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\intercal \nabla \mathcal{L}_\tau^\flat(\boldsymbol{\theta}') - \frac{n}{2} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^2. \end{aligned} \tag{B.44}$$

By the mean value theorem, $R^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\intercal \nabla_{\boldsymbol{\theta}} R^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}}$, where $\widetilde{\boldsymbol{\theta}}$ is a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ and thus satisfies $\widetilde{\boldsymbol{\theta}} \in \Theta_0(r)$. It follows that

$$R^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 2r \sup_{\boldsymbol{\theta} \in \Theta_0(r)} \|\boldsymbol{G}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2, \tag{B.45}$$

where $\boldsymbol{G}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') := \boldsymbol{\Sigma}^{-1/2} \nabla_{\boldsymbol{\theta}} R^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{\Sigma}^{-1/2}\{\nabla \mathcal{L}_\tau^\flat(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau^\flat(\boldsymbol{\theta}')\} - n\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')$. To bound the right-hand side of (B.45), we will deal with

$$\boldsymbol{D}(\boldsymbol{\theta}, \boldsymbol{\theta}') := \boldsymbol{G}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}') - \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{\theta}') \quad \text{and} \quad \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

separately, where $\boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}^*\{\boldsymbol{G}^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}')\} = \boldsymbol{\Sigma}^{-1/2}\{\nabla \mathcal{L}_\tau(\boldsymbol{\theta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\theta}')\} - n\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')$. For the latter, we have

$$\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_0(r)} \|\boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 \leq 2n\Delta(r), \tag{B.46}$$

where $\Delta(r)$ is given in (B.13). For the former term, note that

$$\boldsymbol{D}(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{i=1}^{n}\{\psi_\tau(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}) - \psi_\tau(Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}')\}U_i\boldsymbol{Z}_i.$$

Define new variables $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\boldsymbol{\delta}' = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}' - \boldsymbol{\theta}^*)$, so that

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_0(r)}\|\boldsymbol{D}(\boldsymbol{\theta},\boldsymbol{\theta}')\|_2 = \sup_{\boldsymbol{\delta},\boldsymbol{\delta}'\in\mathbb{B}^d(r)}\|\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{\delta}')\|_2 \le 2\sup_{\boldsymbol{\delta}\in\mathbb{B}^d(r)}\|\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{0})\|_2, \tag{B.47}$$

where $\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{\delta}') = \sum_{i=1}^n\{\psi_\tau(\varepsilon_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\delta}) - \psi_\tau(\varepsilon_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\delta}')\}U_i\boldsymbol{Z}_i$. It is easy to see that the random process $\{\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{0}), \boldsymbol{\delta}\in\mathbb{B}^d(r)\}$ satisfies $\overline{\boldsymbol{D}}(\boldsymbol{0},\boldsymbol{0}) = \boldsymbol{0}$ and $\mathbb{E}^*\{\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{0})\} = \boldsymbol{0}$. Moreover, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}$ and $\lambda\in\mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}^*\exp\left\{\frac{\lambda}{\sqrt{n}}\boldsymbol{u}^\intercal\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{0})\boldsymbol{v}\right\}\\
&= \prod_{i=1}^n\mathbb{E}^*\exp\left\{-\frac{\lambda}{\sqrt{n}}\ell_\tau''(\varepsilon_i - \boldsymbol{Z}_i^\intercal\boldsymbol{\delta})(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)U_i\right\}\\
&\le \prod_{i=1}^n\exp\left\{\frac{\lambda^2}{2n}B_U^2(\boldsymbol{u}^\intercal\boldsymbol{Z}_i)^2(\boldsymbol{v}^\intercal\boldsymbol{Z}_i)^2\right\} \le \exp\left(\frac{\lambda^2}{2}B_U^2M_{n,4}\right).
\end{aligned}$$

It then follows from Theorem A.3 in Spokoiny (2013) that

$$\mathbb{P}^*\left\{\sup_{\boldsymbol{\delta}\in\mathbb{B}^d(r)}\|\overline{\boldsymbol{D}}(\boldsymbol{\delta},\boldsymbol{0})\|_2 \ge 6B_UM_{n,4}^{1/2}(8d+2t)^{1/2}r\sqrt{n}\right\} \le e^{-t}. \tag{B.48}$$

Together, the estimates (B.45)–(B.48) imply that, with $\mathbb{P}^*$-probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_0(r)}\|R^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\|_2 \le 4r\{n\Delta(r) + 6B_UM_{n,4}^{1/2}(8d+2t)^{1/2}r\sqrt{n}\}. \tag{B.49}$$

Recall the proof of Theorem 2.3 and note that

$$\begin{aligned}
&\left|(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)^\intercal\nabla\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) + \frac{n}{2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\|_2^2 + \frac{1}{2n}\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)\|_2^2\right|\\
&= \frac{1}{2n}\|\boldsymbol{G}^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat,\widehat{\boldsymbol{\theta}}_\tau)\|_2^2\\
&\le \frac{1}{2n}(\|\boldsymbol{D}(\widehat{\boldsymbol{\theta}}_\tau^\flat,\widehat{\boldsymbol{\theta}}_\tau)\|_2 + \|\boldsymbol{G}(\widehat{\boldsymbol{\theta}}_\tau^\flat,\widehat{\boldsymbol{\theta}}_\tau)\|_2)^2.
\end{aligned}$$

This, together with (B.46)–(B.49) and the proof of Theorem 2.3, yields that, conditioning on some event that occurs with $\mathbb{P}^\dagger$-probability at least $1-5e^{-t}$,

$$M_{n,4}^{1/2} \le C_{10}$$

for some $C_{10} = C_{10}(A_0) > 0$ and moreover, the following inequalities

$$\begin{aligned}&\left|\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - (\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)^\intercal \nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \frac{n}{2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\|_2^2\right| \\ &\qquad \le 4r_2\{n\Delta(r_2) + 6B_U C_{10}(8d+2t)^{1/2} r_2\sqrt{n}\}\end{aligned}$$

and

$$\begin{aligned}&\left|(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)^\intercal \nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) + \frac{n}{2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau^\flat - \widehat{\boldsymbol{\theta}}_\tau)\|_2^2 + \frac{1}{2n}\|\boldsymbol{\Sigma}^{-1/2}\nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)\|_2^2\right| \\ &\qquad \le \frac{2}{n}\{n\Delta(r_2) + 6B_U C_{10}(8d+2t)^{1/2} r_2\sqrt{n}\}^2\end{aligned}$$

hold with $\mathbb{P}^*$-probability greater than $1-4e^{-t}$. Recall that $\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) = \mathbf{0}$ and by (2.16), $\boldsymbol{\Sigma}^{-1/2}\nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) = \boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau)$. It then follows that

$$\begin{aligned}&|\|\boldsymbol{\Sigma}^{-1/2}\nabla \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)\|_2^2 - \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2^2| \\ &= |\|\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau)\|_2^2 - \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2^2| \\ &\le \|\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\{\|\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2 + 2\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\}.\end{aligned}$$

Putting $\Delta^\flat(r) = \sup_{\boldsymbol{\theta}\in\Theta_0(r)} \|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}) - \boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2$ and combining the last three displays, we conclude that

$$\begin{aligned}&\left|\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) - \frac{1}{2n}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2^2\right| \\ &\le 4r_2\{n\Delta(r_2) + 6B_U C_{10}(8d+2t)^{1/2} r_2\sqrt{n}\} \\ &\quad + \frac{2}{n}\{n\Delta(r_2) + 6B_U C_{10}(8d+2t)^{1/2} r_2\sqrt{n}\}^2 \\ &\quad + \frac{\Delta^\flat(r_1)}{2n}\{\Delta^\flat(r_1) + 2\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\} \le C_{11}v^2\frac{(d+t)^{3/2}}{\sqrt{n}},\end{aligned} \tag{B.50}$$

for some $C_{11} = C_{11}(A_0, A_U) > 0$. This proves (2.17) immediately.

For the square-root Wilks expansion, note that

$$\begin{aligned}&\left|\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}} - \sqrt{n}\,\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \widehat{\boldsymbol{\theta}}_\tau^\flat)\|_2\right| \\ &\le \frac{|2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\} - n\,\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \widehat{\boldsymbol{\theta}}_\tau^\flat)\|_2^2|}{\sqrt{n}\,\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \widehat{\boldsymbol{\theta}}_\tau^\flat)\|_2} = \frac{2}{\sqrt{n}}\frac{|R^\flat(\widehat{\boldsymbol{\theta}}_\tau, \widehat{\boldsymbol{\theta}}_\tau^\flat)|}{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau - \widehat{\boldsymbol{\theta}}_\tau^\flat)\|_2},\end{aligned}$$

where $R^\flat(\cdot,\cdot)$ is given in (B.44). For any $r>0$, similarly to (B.45), it holds

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_0(r)}\frac{|R^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')|}{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}')\|_2}\le\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_0(r)}\|\boldsymbol{G}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\|_2.$$

Again, the estimates (B.46)–(B.48) imply that, with $\mathbb{P}^*$-probability at least $1-e^{-t}$,

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_0(r)}\|\boldsymbol{G}^\flat(\boldsymbol{\theta},\boldsymbol{\theta}')\|_2\le 2n\Delta(r)+12B_UM_{n,4}^{1/2}(8d+2t)^{1/2}r\sqrt{n},$$

where $\Delta(r)$ is given in (B.13). Recall that $\boldsymbol{G}^\flat(\widehat{\boldsymbol{\theta}}_\tau,\widehat{\boldsymbol{\theta}}_\tau^\flat)=\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)-n\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau-\widehat{\boldsymbol{\theta}}_\tau^\flat)=\boldsymbol{\xi}^\flat(\widehat{\boldsymbol{\theta}}_\tau)-n\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\tau-\widehat{\boldsymbol{\theta}}_\tau^\flat)$. Following the same argument that delivers (B.50), we reach

$$\begin{aligned}&\left|\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)-\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}}-n^{-1/2}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\right|\\&\le 6\sqrt{n}\,\Delta(r_2)+36B_UC_{10}(8d+2t)^{1/2}r_2\\&\quad+\frac{1}{\sqrt{n}}\sup_{\boldsymbol{\theta}\in\Theta_0(r_1)}\|\boldsymbol{\xi}^\flat(\boldsymbol{\theta})-\boldsymbol{\xi}^\flat(\boldsymbol{\theta}^*)\|_2\le C_{12}v\frac{d+t}{\sqrt{n}},\end{aligned}$$

where $C_{12}=C_{12}(A_0,A_U)>0$. This is the bound stated in (2.18). □

B.6. Proof of Theorem 2.5. We divide the proof into three steps. In the first step, we revisit the non-asymptotic square-root Wilks approximations for the excess loss and its bootstrap counterpart. The second step is on Gaussian approximation for the ℓ_2-norm of the standardized score vector $\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*)$. The last step links the distributions of the excess loss and its bootstrap counterpart via a Gaussian comparison inequality. Without loss of generality, we assume $t\ge 1$ throughout the proof.

Step 1 (Wilks approximations). Define $\boldsymbol{\xi}^*=-\sum_{i=1}^n\xi_i\boldsymbol{Z}_i$ and recall that $\boldsymbol{\xi}^\flat=-\sum_{i=1}^n\xi_iU_i\boldsymbol{Z}_i$.

For any $x\ge 0$, it follows from (2.7) that

$$\mathbb{P}\left[\sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*)-\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}}\le x\right]\le\mathbb{P}\left(\frac{\|\boldsymbol{\xi}^*\|_2}{\sqrt{n}}\le x+R_1\right)+3e^{-t},\tag{B.51}$$

where $R_1>0$ satisfies $R_1\asymp v(d+t)n^{-1/2}$. Similarly, applying (2.18) yields that, with probability (over $\mathcal{D}_n$) at least $1-5e^{-t}$,

$$\begin{aligned}&\mathbb{P}\left[\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau)-\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}}\le x\middle|\mathcal{D}_n\right]\\&\ge\mathbb{P}\left\{\frac{\|\boldsymbol{\xi}^\flat\|_2}{\sqrt{n}}\le\max(x-R_2,0)\middle|\mathcal{D}_n\right\}-4e^{-t},\end{aligned}\tag{B.52}$$

where $R_2 > 0$ satisfies $R_2 \asymp v(d+t)n^{-1/2}$. In the following two steps, we validate the approximation of the distribution of $\|\boldsymbol{\xi}^\flat\|_2$ by that of $\|\boldsymbol{\xi}^*\|_2$ in the Kolmogorov distance. To that end, define random vectors

$$\boldsymbol{S}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{V}_i \quad \text{and} \quad \boldsymbol{S}_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\boldsymbol{V}_i \quad \text{with} \quad \boldsymbol{V}_i = \xi_i\boldsymbol{Z}_i, \quad \xi_i = \psi_\tau(\varepsilon_i).$$

In this notation, we have $\|\boldsymbol{\xi}^*\|_2 = \sqrt{n}\|\boldsymbol{S}_1\|_2$ and $\|\boldsymbol{\xi}^\flat\|_2 = \sqrt{n}\|\boldsymbol{S}_2\|_2$.

STEP 2 (Gaussian approximation for $\|\boldsymbol{S}_1\|_2$). Recall the truncated mean and second moment $m_\tau = \mathbb{E}(\xi_i)$ and $\sigma_\tau^2 = \mathbb{E}(\xi_i^2)$, and consider the centered sum

$$\overline{\boldsymbol{S}}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\boldsymbol{V}_i - \mathbb{E}\boldsymbol{V}_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\boldsymbol{V}_i - m_\tau\boldsymbol{\upsilon}) = \boldsymbol{S}_1 - \sqrt{n}\, m_\tau\boldsymbol{\upsilon},$$

where $\boldsymbol{\upsilon} = \mathbb{E}(\boldsymbol{Z})$ is such that $\|\boldsymbol{\upsilon}\|_2 \le 1$. Here, $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ are independent copies of the random vector $\boldsymbol{V} = \psi_\tau(\varepsilon)\boldsymbol{Z} \in \mathbb{R}^d$ with mean $m_\tau\boldsymbol{\upsilon}$ and covariance matrix $\boldsymbol{\Sigma}_1 = \sigma_\tau^2\,\mathbf{I}_d - m_\tau^2\,\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\intercal$. For (m_τ, σ_τ^2), applying Proposition A.2 with $\kappa = 4$ in the supplement of Zhou et al. (2018) gives

$$|m_\tau| \le \upsilon_4\tau^{-3} \quad \text{and} \quad \sigma^2 - \upsilon_4\tau^{-2} \le \sigma_\tau^2 \le \sigma^2.$$

Hence, for any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, it holds

$$\sigma^2(1 - \upsilon_4\sigma^{-2}\tau^{-2} - \upsilon_4^2\sigma^{-2}\tau^{-6}) \le \|\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{u}\|_2^2 \le \sigma^2.$$

Taking $\tau = v\{n/(d+t)\}^\eta$ for $v \ge \upsilon_4^{1/4}$, this implies $\overline{\lambda}_{\boldsymbol{\Sigma}_1} \le \sigma^2$ and

$$\underline{\lambda}_{\boldsymbol{\Sigma}_1} \ge \sigma^2\left\{1 - \frac{\upsilon_4^{1/2}}{\sigma^2}\left(\frac{d+t}{n}\right)^{2\eta} - \frac{\upsilon_4^{1/2}}{\sigma^2}\left(\frac{d+t}{n}\right)^{6\eta}\right\} \ge \frac{1}{2}\sigma^2$$

provided $n \ge (4\upsilon_4^{1/2}/\sigma^2)^{1/(2\eta)}(d+t)$. Also, under this scaling condition, it holds $\sigma_\tau^2 \ge 3\sigma^2/4$. It then follows from a multivariate central limit theorem (Bentkus, 2005) that

$$\begin{aligned}&\sup_{y\ge 0}|\mathbb{P}(\|\overline{\boldsymbol{S}}_1\|_2 \le y) - \mathbb{P}(\|\boldsymbol{G}_1\|_2 \le y)| \\ &\le \frac{C_1}{\sqrt{n}}\mathbb{E}\{\|\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{V} - m_\tau\boldsymbol{\upsilon})\|_2^3\} \le C_2 \max_{1\le j\le d}\mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}},\end{aligned} \tag{B.53}$$

where $\boldsymbol{G}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $C_1, C_2 > 0$ are absolute constants.

Let $\boldsymbol{G}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 := \sigma_\tau^2 \mathbf{I}_d$. Note that $\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1/2} - \mathbf{I}_d = m_\tau^2\sigma_\tau^{-2}\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\intercal$,

$$\|m_\tau^2\sigma_\tau^{-2}\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\intercal\|_2 \le \delta_\tau := \upsilon_4^2\sigma_\tau^{-2}\tau^{-6} \quad \text{and} \quad \operatorname{tr}\{(m_\tau^2\sigma_\tau^{-2}\boldsymbol{\upsilon}\boldsymbol{\upsilon}^\intercal)^2\} \le \delta_\tau^2.$$

Applying Lemma A.7 in the supplementary material of Spokoiny and Zhilova (2015) gives

$$\sup_{y\ge 0} |\mathbb{P}(\|\boldsymbol{G}_1\|_2 \le y) - \mathbb{P}(\|\boldsymbol{G}_0\|_2 \le y)| \le \delta_\tau/2 \tag{B.54}$$

provided $\delta_\tau \le 1/2$. In addition, the Gaussian random vector $\boldsymbol{G}_0$ satisfies the following anti-concentration inequality (Ball, 1993): for any $\epsilon \ge 0$,

$$\sup_{y\ge 0} \mathbb{P}(y \le \|\boldsymbol{G}_0\|_2 \le y + \epsilon) \le \frac{C_3}{\sigma_\tau}\epsilon, \tag{B.55}$$

where $C_3 > 0$ is an absolute constant.

For the deterministic term $\mathbb{E}(\boldsymbol{S}_1) = \sqrt{n}\, m_\tau \boldsymbol{\upsilon}$, we have

$$\gamma_1 := \|\mathbb{E}(\boldsymbol{S}_1)\|_2 \le \upsilon_4\tau^{-3}\sqrt{n} \le \upsilon_4^{1/4}\frac{(d+t)^{3\eta}}{n^{3\eta-1/2}}.$$

Combining this with (B.53)–(B.55), we arrive at

$$\begin{aligned}
&\mathbb{P}\left(\frac{\|\boldsymbol{\xi}^*\|_2}{\sqrt{n}} \le x + R_1\right) = \mathbb{P}(\|\boldsymbol{S}_1\|_2 \le x + R_1) \\
&\le \mathbb{P}(\|\overline{\boldsymbol{S}}_1\|_2 \le x + R_1 + \gamma_1) \\
&\le \mathbb{P}(\|\boldsymbol{G}_1\|_2 \le x + R_1 + \gamma_1) + C_2 \max_{1\le j\le d}\mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}} \\
&\le \mathbb{P}(\|\boldsymbol{G}_0\|_2 \le x+R_1+\gamma_1)+C_2 \max_{1\le j\le d}\mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}}+\frac{\upsilon_4^{1/2}}{2\sigma_\tau^2}\left(\frac{d+t}{n}\right)^{6\eta} \\
&\le \mathbb{P}\{\|\boldsymbol{G}_0\|_2 \le \max(x-R_2, 0)\} \\
&\quad + C_2 \max_{1\le j\le d}\mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}} + \frac{\upsilon_4^{1/2}}{2\sigma_\tau^2}\left(\frac{d+t}{n}\right)^{6\eta} + \frac{C_3}{\sigma_\tau}(R_1+R_2+\gamma_1).
\end{aligned} \tag{B.56}$$

Step 3 (Gaussian comparison). Note that, conditional on $\mathcal{D}_n$, $\boldsymbol{S}_2$ follows a multivariate normal distribution with mean $\mathbb{E}(\boldsymbol{S}_2|\mathcal{D}_n) = \mathbf{0}$ and covariance matrix

$$\mathbf{S}_n = \operatorname{cov}(\boldsymbol{S}_2|\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^n \xi_i^2 \boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal \in \mathbb{R}^{d\times d}.$$

Applying Lemma A.2 with $x = t$ yields that, with probability at least $1-e^{-t}$,

$$\|\boldsymbol{\Sigma}_0^{-1/2}\mathbf{S}_n\boldsymbol{\Sigma}_0^{-1/2}-\mathbf{I}_d\|_2 \le C_4\frac{v^2}{\sigma^2}\left(\frac{d+t}{n}\right)^{1-2\eta} \le \frac{1}{2}$$
$$\text{and } \operatorname{tr}\{(\boldsymbol{\Sigma}_0^{-1/2}\mathbf{S}_n\boldsymbol{\Sigma}_0^{-1/2}-\mathbf{I}_d)^2\} \le C_4^2\frac{v^4 d}{\sigma^2}\left(\frac{d+t}{n}\right)^{2-4\eta}$$

as long as n is sufficiently large, where $C_4 = C_4(A_0) > 0$. Hence, it follows from a conditional version of Lemma A.7 in the supplement of Spokoiny and Zhilova (2015) that, with probability (over $\mathcal{D}_n$) at least $1-e^{-t}$,

$$\sup_{y\ge 0}|\mathbb{P}(\|\boldsymbol{S}_2\|_2 \le y|\mathcal{D}_n) - \mathbb{P}(\|\boldsymbol{G}_0\|_2 \le y)| \le C_4\frac{v^2}{2\sigma^2}\sqrt{d}\left(\frac{d+t}{n}\right)^{1-2\eta}.$$

In particular, taking $y = \max(x-R_2, 0)$ gives

$$\begin{aligned}&\mathbb{P}\{\|\boldsymbol{G}_0\|_2 \le \max(x-R_2,0)\}\\ &\le \mathbb{P}\left\{\frac{\|\boldsymbol{\xi}^\flat\|_2}{\sqrt{n}} \le \max(x-R_2,0)\middle|\mathcal{D}_n\right\} + C_4\frac{v^2}{2\sigma^2}\sqrt{d}\left(\frac{d+t}{n}\right)^{1-2\eta}.\end{aligned} \tag{B.57}$$

Combining the inequalities (B.51), (B.52), (B.56) and (B.57), we conclude that with probability (over $\mathcal{D}_n$) at least $1-6e^{-t}$,

$$\begin{aligned}&\mathbb{P}\left[\sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}} \le x\right]\\ &\le \mathbb{P}\left[\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}} \le x\middle|\mathcal{D}_n\right] + 7e^{-t} + \frac{v_4^{1/2}}{2\sigma_\tau^2}\left(\frac{d+t}{n}\right)^{6\eta}\\ &\quad + C_2\max_{1\le j\le d}\mathbb{E}(|Z_j|^3)\frac{v_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}} + \frac{C_3}{\sigma_\tau}(R_1+R_2+\gamma_1) + C_4\frac{v^2}{2\sigma^2}\sqrt{d}\left(\frac{d+t}{n}\right)^{1-2\eta}.\end{aligned}$$

A similar argument leads to the reverse inequality and thus completes the proof by taking $z = x^2/2$. □

B.7. Proof of Theorem 2.6. For $\alpha \in (0,1)$, let $q_\alpha^\flat$ and q_α be the upper α-quantiles of

$$\sqrt{2\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat)\}} \text{ and } \sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}},$$

respectively, under $\mathbb{P}^*$ and $\mathbb{P}$. By the definitions of $z_\alpha^\flat$ and z_α in (2.21) and (2.13), it is easy to see that $z_\alpha^\flat = (q_\alpha^\flat)^2/2$ almost surely and $z_\alpha = q_\alpha^2/2$.

According to Theorem 2.5, there exists an event $\mathcal{E}_t$ satisfying $\mathbb{P}(\mathcal{E}_t) \geq 1-6e^{-t}$ such that

$$\begin{aligned}&\mathbb{P}^*\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) > q_{\alpha-\Delta_1}^2/2\}\\ &\begin{cases} = 0 < \alpha, & \text{if } \alpha \leq \Delta_1,\\ \leq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > q_{\alpha-\Delta_1}^2/2\} + \Delta_1 \leq \alpha, & \text{if } \alpha > \Delta_1, \end{cases}\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}^*\{\mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau) - \mathcal{L}_\tau^\flat(\widehat{\boldsymbol{\theta}}_\tau^\flat) > (q_{\alpha+\Delta_1} - \sigma/n)^2/2\}\\ &\geq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > (q_{\alpha+\Delta_1} - \sigma/n)^2/2\} - \Delta_1 \geq \alpha\end{aligned}$$

hold almost surely on $\mathcal{E}_t$, where $\Delta_1 = \Delta_1(n, d, t)$. Together, these inequalities imply

$$q_{\alpha+\Delta_1} - \sigma/n \leq q_\alpha^\flat \leq q_{\alpha-\Delta_1} \quad \text{almost surely on } \mathcal{E}_t. \tag{B.58}$$

Next, define the Lévy concentration function of the non-negative random variable $T := \sqrt{2\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau)\}}$:

$$L(\epsilon) = \sup_{x \geq 0} \mathbb{P}(|T - x| \leq \epsilon), \quad \epsilon \geq 0.$$

It then follows from (B.58) that

$$\begin{aligned}&\mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > z_\alpha^\flat\}\\ &\geq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > q_{\alpha-\Delta_1}^2/2\} - 6e^{-t}\\ &\geq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > (q_{\alpha-\Delta_1} - \sigma/n)^2/2\} - L(\sigma/n) - 6e^{-t}\\ &\geq \alpha - \Delta_1 - L(\sigma/n) - 6e^{-t}. \qquad \text{(B.59)}\end{aligned}$$

Similarly, using (B.9) and the definition of $L(\cdot)$, we get

$$\begin{aligned}&\mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > z_\alpha^\flat\}\\ &\leq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > (q_{\alpha+\Delta_1} - \sigma/n)^2/2\} + 6e^{-t}\\ &\leq \mathbb{P}\{\mathcal{L}_\tau(\boldsymbol{\theta}^*) - \mathcal{L}_\tau(\widehat{\boldsymbol{\theta}}_\tau) > q_{\alpha+\Delta_1}^2/2\} + L(\sigma/n) + 6e^{-t}\\ &\leq \alpha + \Delta_1 + L(\sigma/n) + 6e^{-t}. \qquad \text{(B.60)}\end{aligned}$$

To complete the proof, it remains to bound $L(\epsilon)$ for any given $\epsilon > 0$. Keeping the notation used in the proof of Theorem 2.5, and following (B.51),

(B.53) and (B.55), we obtain that for any $x \geq 0$,

$$\begin{aligned}
&\mathbb{P}(|T - x| \leq \epsilon) \\
&\leq \mathbb{P}(|\|\boldsymbol{S}_1\|_2 - x| \leq \epsilon + R_1) + 3e^{-t} \\
&\leq \mathbb{P}(|\|\overline{\boldsymbol{G}}_1\|_2 - x| \leq \epsilon + R_1) + 2C_2 \max_{1\leq j\leq d} \mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}} + 3e^{-t} \\
&\leq C\frac{\epsilon + R_1}{\underline{\lambda}_{\boldsymbol{\Sigma}_1}^{1/2}} + 2C_2 \max_{1\leq j\leq d} \mathbb{E}(|Z_j|^3)\frac{\upsilon_3}{\sigma^3}\frac{d^{3/2}}{\sqrt{n}} + 3e^{-t}, \qquad \text{(B.61)}
\end{aligned}$$

where $R_1 \asymp v(d+t)n^{-1/2}$ is as in (B.51), $\overline{\boldsymbol{G}}_1 \sim \mathcal{N}(\mathbb{E}(\boldsymbol{S}_1), \boldsymbol{\Sigma}_1)$ and $C > 0$ is an absolute constant.

Finally, combining (B.59), (B.60) and (B.61) to reach (2.22). □

APPENDIX C: PROOFS FOR SECTIONS 3 AND 4

C.1. Proof of Theorem 3.1. This proof is based on an argument similar to that used in the proof of Theorem 5.1 in Minsker (2018). Let $j^* = \min\{j \in \mathcal{J} : \upsilon_j \geq \sigma\}$ and note that $\upsilon_{j^*} \leq a\sigma$. From the definition of $\widehat{j}_{\mathrm{L}}$ in (3.1) with $c_0 \geq 2c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}$, we see that

$$\begin{aligned}
\{\widehat{j}_{\mathrm{L}} > j^*\} &\subseteq \bigcup_{k\in\mathcal{J}:k>j^*} \left\{\|\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}^{(j^*)}\|_2 > c_0\upsilon_k\sqrt{\frac{d+t}{n}}\right\} \\
&\subseteq \bigcup_{k\in\mathcal{J}:k\geq j^*} \left\{\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\|_2 > c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}\upsilon_k\sqrt{\frac{d+t}{n}}\right\}.
\end{aligned}$$

Define the event

$$\mathcal{B} = \bigcap_{k\in\mathcal{J}:k\geq j^*} \left\{\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\|_2 \leq c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}\upsilon_k\sqrt{\frac{d+t}{n}}\right\},$$

such that $\mathcal{B} \subseteq \{\widehat{j}_{\mathrm{L}} \leq j^*\}$. Recalling Theorem 2.1, we have for any $v \geq \sigma$, $\widehat{\boldsymbol{\theta}}_\tau$ with $\tau = v\sqrt{n/(d+t)}$ satisfies the bound

$$\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*\|_2 \leq c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}v\sqrt{\frac{d+t}{n}}$$

with probability at least $1 - 3e^{-t}$ as long as $n \gtrsim d + t$. Together with the union bound, this implies

$$\begin{aligned}
\mathbb{P}(\mathcal{B}^{\mathrm{c}}) &\leq \sum_{k\in\mathcal{J}:k\geq j^*} \mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\|_2 > c_1\underline{\lambda}_{\boldsymbol{\Sigma}}^{-1/2}\upsilon_k\sqrt{\frac{d+t}{n}}\right) \\
&\leq 3|\mathcal{J}|e^{-t} \leq 3\{1 + \log_a(\upsilon_{\max}/\upsilon_{\min})\}e^{-t}.
\end{aligned}$$

On the event $\mathcal{B}$, $\widehat{j}_{\mathrm{L}} \le j^*$ and thus

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}^{(\widehat{j}_{\mathrm{L}})} - \boldsymbol{\theta}^*\|_2 &\le \|\widehat{\boldsymbol{\theta}}^{(\widehat{j}_{\mathrm{L}})} - \widehat{\boldsymbol{\theta}}^{(j^*)}\|_2 + \|\widehat{\boldsymbol{\theta}}^{(j^*)} - \boldsymbol{\theta}^*\|_2 \\ &\le (c_0 + c_1\lambda_{\boldsymbol{\Sigma}}^{-1/2})\upsilon_{j^*}\sqrt{\frac{d+t}{n}} \le \frac{3a}{2}c_0\sigma\sqrt{\frac{d+t}{n}}.\end{aligned}$$

Together, the last two displays lead to the stated result. □

C.2. Proof of Theorem 3.2. To begin with, define $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$ to be the two independent samples $\{(Y_i^{(1)}, \boldsymbol{X}_i^{(1)})\}_{i=1}^n$ and $\{(Y_i^{(2)}, \boldsymbol{X}_i^{(2)})\}_{i=1}^n$, respectively, such that $\bar{\mathcal{D}}_n = \mathcal{D}_n^{(1)} \cup \mathcal{D}_n^{(2)}$. Under the assumption that $\mathbb{E}(|\varepsilon|^{4+\delta}) \le \upsilon_{4+\delta}$ for some $\delta > 0$, we have $\mathbb{E}|Y - \mu_Y|^{4+\delta} < \infty$. For each $j = 1, \ldots, m$ with m denoting the number of blocks, by Chebyshev's inequality, one can show that for any $\delta \in (0, 1/2]$,

$$|\widehat{\upsilon}_{Y,j} - \upsilon_Y| \lesssim \left(\frac{\mathbb{E}|Y - \mu_Y|^{4+\delta}}{\delta n^{\delta/4}}\right)^{4/(4+\delta)},$$

with probability at least $1 - \delta$. Then, it follows from a variant of Lemma 2 in Bubeck, Cesa-Bianchi and Lugosi (2013) that, with $m = \lfloor 8 \log n + 1 \rfloor$,

$$\mathbb{P}\left\{|\widehat{\upsilon}_{Y,\mathrm{mom}} - \upsilon_Y| \gtrsim (\mathbb{E}|Y - \mu_Y|^{4+\delta})^{4/(4+\delta)}\left(\frac{\log n}{n}\right)^{\delta/(4+\delta)}\right\} \lesssim n^{-1}$$

as long as $n \gtrsim \log n$, where the probability is over the training set $\mathcal{D}_n^{(1)}$. Therefore, with the same probability (over $\mathcal{D}_n^{(1)}$), $|\widehat{\upsilon}_{Y,\mathrm{mom}} - \upsilon_Y| \le \upsilon_Y/2$ for all sufficiently large n. Then, it follows that, with high probability over $\mathcal{D}_n^{(1)}$,

(C.1)
$$\upsilon_4^{1/4} < \upsilon_Y^{1/4} \le \upsilon_{\max} \le (3\upsilon_Y)^{1/4} \quad \text{and} \quad \upsilon_{\min} = \frac{\upsilon_{\max}}{a^K} \le \frac{(3\upsilon_Y)^{1/4}}{a^K} < \upsilon_4^{1/4},$$

where the last inequality holds provided $K \ge \lfloor \log_a(3\upsilon_Y/\upsilon_4)^{1/4} \rfloor + 1$.

Next, recall that $\widehat{\boldsymbol{\theta}}^{(1)} = \widehat{\boldsymbol{\theta}}^{(1)}(\mathcal{D}_n^{(1)})$ is constructed via Lepski's adaptive procedure with the robustification parameter chosen from $\mathcal{J} = \{j \in \mathbb{Z} : \upsilon_{\min} \le \upsilon_j = \upsilon_{\min} a^j < a\upsilon_{\max}\}$. Let $\mathcal{E}_1$ be the event on which (C.1) holds so that $\mathbb{P}(\mathcal{E}_1^{\mathrm{c}}) \lesssim n^{-1}$. Moreover, note that $\upsilon_{\min} < \upsilon_4^{1/4} < \upsilon_{\max} \le (3\upsilon_Y)^{1/4}$ on $\mathcal{E}_1$, and by construction, $\log_a(a\upsilon_{\max}/\upsilon_{\min}) = K + 1$. Using arguments similar to those in the proof of Theorem 3.1 and taking $K = \lfloor \log_a(3\upsilon_Y/\upsilon_4)^{1/4} \rfloor + 1$, it can be shown that with probability (over $\mathcal{D}_n^{(1)}$) at least $1 - O(Kn^{-1})$,

(C.2)
$$\|\widehat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^*\|_2 \lesssim \upsilon_4^{1/4}\sqrt{\frac{d + \log n}{n}} \quad \text{and} \quad \frac{1}{3^{1/4}a}\tau^* \le \widehat{\tau} \le a\tau^*,$$

where the second inequality uses the property that the selected index $\tilde{j}$ satisfies $\tilde{j} \le j^* = \min\{j \in \mathcal{J} : v_j \ge v_4^{1/4}\}$ with high probability, and $\tau^* := v_4^{1/4}(\frac{n}{d+\log n})^{1/4}$.

For the second step, write $\varepsilon_i^{(2)} = Y_i^{(2)} - \langle \boldsymbol{X}_i^{(2)}, \boldsymbol{\theta}^* \rangle$ for $i = 1, \ldots, n$. Define random vectors

$$\boldsymbol{\xi}^* = \sum_{i=1}^n \ell'_{\widehat{\tau}}(\varepsilon_i^{(2)}) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}_i^{(2)} \quad \text{and} \quad \boldsymbol{\xi}^\flat = \sum_{i=1}^n \ell'_{\widehat{\tau}}(\varepsilon_i^{(2)})(1 - W_i) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}_i^{(2)}.$$

Conditioning on the event that (C.2) holds (which ensures $\widehat{\tau} \asymp v_4^{1/4}(\frac{n}{d+\log n})^{1/4}$), applying Theorems 2.2 and 2.4 we obtain that as long as $n \gtrsim d + \log n$,

$$\left| \sqrt{2\{\widehat{\mathcal{L}}(\boldsymbol{\theta}^*) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}})\}} - \frac{\|\boldsymbol{\xi}^*\|_2}{\sqrt{n}} \right| \lesssim v_4^{1/4} \frac{d + \log n}{\sqrt{n}}$$

with probability (over $\mathcal{D}_n^{(2)}$) at least $1 - O(n^{-1})$, and

$$\left| \sqrt{2\{\widehat{\mathcal{L}}^\flat(\widehat{\boldsymbol{\theta}}) - \widehat{\mathcal{L}}^\flat(\widehat{\boldsymbol{\theta}}^\flat)\}} - \frac{\|\boldsymbol{\xi}^\flat\|_2}{\sqrt{n}} \right| \lesssim v_4^{1/4} \frac{d + \log n}{\sqrt{n}}$$

with probability (over $\mathcal{D}_n^{(2)}$ and $\{W_i\}_{i=1}^n$) at least $1 - O(n^{-1})$. With the above preparations, the stated result follows from the same argument as in the proof of Theorem 2.6. □

C.3. Proof of Theorem 4.1. The proof consists of two main steps.

STEP 1 (Accuracy of bootstrap approximations). For each $1 \le k \le m$, write $\mathcal{D}_{kn} = \{(y_{ik}, \boldsymbol{x}_i)\}_{i=1}^n$ and $T_k^\flat = \sum_{i=1}^n \psi_{\tau_k}(\varepsilon_{ik}) U_i$. Then, applying Theorem 2.3 with $\boldsymbol{X} = (1, \boldsymbol{x}^\intercal)^\intercal$ and $\boldsymbol{\theta}^* = (\mu_k, \boldsymbol{\beta}_k^\intercal)^\intercal$ gives that, with probability (over $\mathcal{D}_{kn}$) at least $1 - 6/(nm)^2$,

$$\mathbb{P}\{|\sqrt{n}\,(\widehat{\mu}_k^\flat - \widehat{\mu}_k) - T_k^\flat/\sqrt{n}| \ge \delta_{2k} | \mathcal{D}_{kn}\} \le 4(nm)^{-2}, \tag{C.3}$$

where $\delta_{2k} := C_{1k}\, v_k \{s + \log(nm)\} n^{-1/2}$ and $C_{1k} = C_{1k}(A_0, A_U) > 0$. Observe that, conditional on $\mathcal{D}_{kn}$, $n^{-1/2} T_k^\flat$ follows a normal distribution with mean zero and variance $\widehat{\sigma}_{k,\tau_k}^2 = (1/n) \sum_{i=1}^n \{\ell'_{\tau_k}(\varepsilon_{ik})\}^2$. With $\tau_k = v_k [n/\{s + 2\log(nm)\}]^{1/3}$, an argument similar to that used to derive Lemma A.2 may be employed to show that, with probability at least $1 - (nm)^{-2}$,

$$\begin{aligned} |\widehat{\sigma}_{k,\tau_k}^2 - \sigma_k^2| &\le |\widehat{\sigma}_{k,\tau_k}^2 - \sigma_{k,\tau_k}^2| + |\sigma_{k,\tau_k}^2 - \sigma_k^2| \\ &\le C_{2k}\, v_k^2 \left\{ \frac{\log(nm)}{n} \right\}^{1/3} + \frac{v_{k,4}}{v_k^2} \left\{ \frac{s + \log(nm)}{n} \right\}^{2/3}, \end{aligned} \tag{C.4}$$

where $C_{2k} = C_{2k}(A_0) > 0$. Combining (C.3), (C.4), Lemma A.7 in Spokoiny and Zhilova (2015) and the union bound, we conclude that with probability (over $\mathcal{D}_{kn}$) at least $1 - 7/(n^2 m)$,

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\{\sqrt{n}\,(\widehat{\mu}_k^\flat - \widehat{\mu}_k) \le x | \mathcal{D}_{kn}\} - \Phi(x/\sigma_k)| \le C_{3k}\, v_k^2 \biggl\{ \frac{\log(nm)}{n} \biggr\}^{1/3} + \frac{4}{(nm)^2}$$

for all $k = 1, \ldots, m$. Combining this with (A.19), (A.20) and taking $a_n = 2\log(nm)$ in Lemma A.5, we conclude that on some event that occurs with probability at least $1 - 7/(n^2 m)$,

$$\frac{\mathbb{P}(\sqrt{n}\,|\widehat{\mu}_k^\flat - \widehat{\mu}_k| \ge z | \mathcal{D}_{kn})}{2\{1 - \Phi(z/\sigma_k)\}} = 1 + o(1) \tag{C.5}$$

uniformly in $0 \le z/\sigma_k \le o\{\min(n^{1/6}, \sqrt{n}/\log m)\}$ and $1 \le k \le m$.

Step 2 (FDP control with bootstrap calibration). For $k = 1, \ldots, m$ and $z \ge 0$, define $\widehat{T}_k = \sqrt{n}\,\widehat{\mu}_k$, $G(z) = 2\{1 - \Phi(z)\}$,

$$G_k(z) = \mathbb{P}_{H_{0k}}(|\widehat{T}_k| \ge z) \quad \text{and} \quad G_k^\flat(z) = \mathbb{P}\{\sqrt{n}\,|\widehat{\mu}_k^\flat - \widehat{\mu}_k| \ge z | \mathcal{D}_{kn}\}.$$

In this notation, we have $p_k^\flat = G_k^\flat(|\widehat{T}_k|)$ for $k = 1, \ldots, m$. As a direct consequence of Lemma 1 in Storey, Taylor and Siegmund (2004), the BH procedure with p-values $\{p_k^\flat\}_{k=1}^m$ is equivalent to Storey's procedure, that is, reject H_{0k} if and only if $p_k^\flat \le t_{\mathrm{S}}^\flat$, where

$$t_{\mathrm{S}}^\flat := \sup\biggl\{ t \in [0,1] : t \le \frac{\alpha \max\{\sum_{k=1}^m I(p_k^\flat \le t), 1\}}{m} \biggr\}.$$

By the definition of $t_{\mathrm{S}}^\flat$, we have

$$t_{\mathrm{S}}^\flat = \frac{\alpha \max\{\sum_{k=1}^m I(p_k^\flat \le t_{\mathrm{S}}^\flat), 1\}}{m}. \tag{C.6}$$

For the bootstrap p-values $p_k^\flat$ and data-driven threshold $t_{\mathrm{S}}^\flat$, we claim that, as $(n, m) \to \infty$,

$$\mathbb{P}\{t_{\mathrm{S}}^\flat \ge \alpha m_{1,\lambda_0}/m\} \to 1 \tag{C.7}$$

$$\text{and} \quad \sup_{b_m/m \le t \le 1} \biggl| \frac{\sum_{k \in \mathcal{H}_0} I(p_k^\flat \le t)}{m_0\, t} - 1 \biggr| \xrightarrow{\mathbb{P}} 0 \tag{C.8}$$

for any sequence $b_m > 0$ satisfying $b_m \to \infty$ and $b_m = o(m)$, where $m_{1,\lambda_0} = \mathrm{card}\{1 \le k \le m : |\mu_k|/\sigma_k \ge \lambda_0\sqrt{(2\log m)/n}\}$. Under condition (4.5), it follows

$$\frac{\sum_{k\in\mathcal{H}_0} I(p_k^\flat \le t_{\mathrm{S}}^\flat)}{m_0\, t_{\mathrm{S}}^\flat} \xrightarrow{\mathbb{P}} 1,$$

which, together with (C.6), proves the stated result (4.6).

It remains to verify (C.7) and (C.8). By (C.6), it is clear that $t_{\mathrm{S}}^\flat \in [\alpha/m, 1]$. Recall that $\log m = o(n^{1/3})$. Then, by (C.5),

$$G_k^\flat(\sigma_k\sqrt{2\log m}) = G(\sqrt{2\log m})\{1 + o_{\mathbb{P}}(1)\}$$

uniformly in $1 \le k \le m$ as $(n, m) \to \infty$. Note that

$$G(\sqrt{2\log m}) = 2\{1 - \Phi(\sqrt{2\log m})\} \sim \sqrt{\frac{2}{\pi}}\frac{1}{m\sqrt{2\log m}} = o(m^{-1}).$$

Combining the last two displays, we see that with probability tending to 1, $t_{\mathrm{S}}^\flat \ge G_k^\flat(\sigma_k\sqrt{2\log m})$ for all $1 \le k \le m$. It follows

$$t_{\mathrm{S}}^\flat \ge \frac{\alpha}{m}\sum_{k=1}^m I\{G_k^\flat(|\widehat{T}_k|) \le G_k^\flat(\sigma_k\sqrt{2\log m})\} = \frac{\alpha}{m}\sum_{k=1}^m I\{|\widehat{T}_k| \ge \sigma_k\sqrt{2\log m}\}.$$

Furthermore,

$$\begin{aligned}&\sum_{k=1}^m I\{|\widehat{T}_k| \ge \sigma_k\sqrt{2\log m}\}\\ &\ge \sum_{k=1}^m I\biggl\{\sqrt{n}\,\frac{|\mu_k|}{\sigma_k} \ge \sqrt{2\log m} + \sqrt{n}\max_{1\le k\le m}\frac{|\widehat{\mu}_k - \mu_k|}{\sigma_k}\biggr\}.\end{aligned}$$

For any $\epsilon > 0$, define the event

$$\mathcal{A}(\varepsilon) = \biggl\{\sqrt{n}\max_{1\le k\le m}\frac{|\widehat{\mu}_k - \mu_k|}{\sigma_k} \le (1+\epsilon)\sqrt{2\log m}\biggr\},$$

on which it holds

$$\sum_{k=1}^m I\{|\widehat{T}_k| \ge \sigma_k\sqrt{2\log m}\} \ge \sum_{k=1}^m I\biggl\{\frac{|\mu_k|}{\sigma_k} \ge (2+\epsilon)\sqrt{\frac{2\log m}{n}}\biggr\}.$$

Using Lemma A.5 and the union bound shows that, as $(n, m) \to \infty$,

$$\begin{aligned}\mathbb{P}\{\mathcal{A}(\epsilon)^{\mathrm{c}}\} &\le \sum_{k=1}^m \mathbb{P}\Bigl\{\sqrt{n}\,|\widehat{\mu}_k - \mu_k| > \sigma_k(1+\epsilon)\sqrt{2\log m}\Bigr\}\\ &\le 2m\exp\{-(1+\epsilon)^2\log m\} = 2m^{-2\epsilon-\epsilon^2} = o(1).\end{aligned} \tag{C.9}$$

Putting the above calculations together leads to the claim (C.7).

Finally we verify (C.8). By Lemma A.5 and (C.5), it is easy to see that

$$\max_{1\le k\le m}\sup_{0\le z\le\sigma_k\sqrt{2\log(nm)}}\left|\frac{G_k^\flat(z)}{G(z/\sigma_k)}-1\right|\xrightarrow{\mathbb{P}}0$$

$$\text{and}\quad \max_{1\le k\le m}\sup_{0\le z\le\sigma_k\sqrt{2\log(nm)}}\left|\frac{G_k^\flat(z)}{G_k(z)}-1\right|\xrightarrow{\mathbb{P}}0$$

as $(n,m)\to\infty$. Also, consider the event

$$\mathcal{A}_0=\left\{\max_{k\in\mathcal{H}_0}|\widehat{T}_k|/\sigma_k\le\sqrt{2\log(nm)}\right\},$$

Similarly to (C.9), it can be shown that $\mathbb{P}(\mathcal{A}_0^c)\to 0$. Consequently, there exits a sequence $\{\alpha_n\}_{n\ge1}$ of positive numbers satisfying $\alpha_n\to 0$ such that

$$\begin{aligned}&\sum_{k\in\mathcal{H}_0}I\{G(|\widehat{T}_k|/\sigma_k)\le(1-\alpha_n)t\}\\ &\le\sum_{k\in\mathcal{H}_0}I(p_k^\flat\le t)\le\sum_{k\in\mathcal{H}_0}I\{G(|\widehat{T}_k|/\sigma_k)\le(1+\alpha_n)t\}.\end{aligned}\tag{C.10}$$

Again, using Lemma A.5 gives

$$\max_{k\in\mathcal{H}_0}\sup_{0\le z\le\sigma_k\sqrt{2\log m}}\left|\frac{\mathbb{P}(|\widehat{T}_k|\ge z)}{G(z/\sigma_k)}-1\right|\to 1.\tag{C.11}$$

Note that, with $0\le z\le\sigma_k\sqrt{2\log m}$, it holds

$$\sqrt{\frac{2}{\pi}}\frac{\sqrt{2\log m}}{1+2\log m}\frac{1}{m}\le G(z/\sigma_k)\le 1.$$

In (C.11), we change the variable by $t=G(z/\sigma_k)$ to obtain

$$\max_{k\in\mathcal{H}_0}\sup_{m^{-1}\le t\le1}\left|\frac{\mathbb{P}\{G(|\widehat{T}_k|/\sigma_k)\le t\}}{t}-1\right|\to 0.$$

By an argument similar to that in the proof of Proposition B.3 in Zhou et al. (2018), it follows that for any sequence $b_m>0$ satisfying $b_m\to\infty$ and $b_m=o(m)$,

$$\sup_{b_m/m\le t\le1}\left|\frac{\sum_{k\in\mathcal{H}_0}I\{G(|\widehat{T}_k|/\sigma_k)\le t\}}{m_0\,t}-1\right|\xrightarrow{\mathbb{P}}0.$$

Together with (C.10), this proves (C.8) as desired. □

APPENDIX D: IMPLEMENTATION

Since the bootstrap Huber estimator needs to be computed repeatedly, an efficient optimization solver is critical for applications. Ideally, second order methods such as Newton's method should be adopted due to fast convergence. Denote the gradient of the weighted Huber loss in (2.12) by

$$\mathbf{g}(\boldsymbol{\theta}) = \sum_{i=1}^{n} W_i\{I(|Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}| \le \tau)(\boldsymbol{X}_i^\intercal\boldsymbol{\theta} - Y_i)\boldsymbol{X}_i + I(|Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}| > \tau)\tau \cdot \operatorname{sgn}(\boldsymbol{X}_i^\intercal\boldsymbol{\theta} - Y_i)\boldsymbol{X}_i\}, \quad \boldsymbol{\theta} \in \mathbb{R}^d. \tag{D.1}$$

Although $\mathbf{g}(\boldsymbol{\theta})$ is not differentiable everywhere with respect to $\boldsymbol{\theta}$, we can still compute a generalized Jacobian of $\mathbf{g}(\boldsymbol{\theta})$:

$$\mathbf{H}(\boldsymbol{\theta}) = \sum_{i=1}^{n} W_i I(|Y_i - \boldsymbol{X}_i^\intercal\boldsymbol{\theta}| \le \tau)\boldsymbol{X}_i\boldsymbol{X}_i^\intercal, \tag{D.2}$$

which serves as an "approximate Hessian matrix". Given (D.2), the generalized Newton method can be directly implemented via the following iterative procedure (for $t = 1, 2, \ldots$):

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t\{\mathbf{H}(\boldsymbol{\theta}^t)\}^{-1}\mathbf{g}(\boldsymbol{\theta}^t), \tag{D.3}$$

where η_t is the step-size. We note that the constraint in (2.12) is omitted here, since it is introduced mainly for theoretical analysis and will not affect the empirical performance.

Although (D.3) is easy to implement, there remains a practical issue that the Hessian matrix $\mathbf{H}(\boldsymbol{\theta}^t)$ is not always invertible. To address this issue, we adopt the damped semismooth Newton method, which is a combination of Newton's method and gradient descent. The idea is straightforward: when $\mathbf{H}(\boldsymbol{\theta}^t)$ is invertible, $\boldsymbol{\theta}^{t+1}$ is computed via the generalized Newton step in (D.3); otherwise, the gradient descent step is performed, that is,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t\,\mathbf{g}(\boldsymbol{\theta}^t). \tag{D.4}$$

The step-size η_t is determined via the backtracking-Armijo line search rule.

Now we briefly discuss the the convergence of the damped semismooth Newton method. Note that the random weights W_i may sometimes take negative values, our objective function could be non-convex, and thus we only discuss the convergence to a stationary point, i.e. some $\widehat{\boldsymbol{\theta}}$ such that $g(\widehat{\boldsymbol{\theta}}) = 0$. The following proposition from Qi and Sun (1999) and De Luca and Facchinei and Kanzow (1996) provides the local convergence rate for solving a system $g(\boldsymbol{\theta}) = 0$.

PROPOSITION D.1. Suppose that $g(\widehat{\boldsymbol{\theta}}) = 0$, where g is locally Lipschitz, and that all $V \in \partial g(\widehat{\boldsymbol{\theta}})$ are non-singular. If g is strongly semismooth at $\widehat{\boldsymbol{\theta}}$, then the method is quadratically convergent in a neighborhood of $\widehat{\boldsymbol{\theta}}$.

Now, let us verify the conditions in Proposition D.1 for the weighted Huber regression. Given the Huber loss $\ell_\tau(x)$ and its gradient $\ell'_\tau(x) = xI(|x| \le \tau) + \tau \cdot \text{sign}(x) I(|x| > \tau)$, the Clarke's generalized Jacobian of $\ell'_\tau(x)$ (Hiriart-Urruty and Lemaréchal, 2001) can be calculated as

$$\partial \ell'_\tau(x) = \begin{cases} 0 & x > \tau \text{ or } x < -\tau \\ 1 & |x| < \tau \\ [0,1] & x = \pm\tau \end{cases}, \tag{D.5}$$

The boundedness of $\ell'_\tau(x)$ implies that g in (D.1) is locally Lipschitz. Moreover, we can easily verify that $\ell'_\tau(x)$ is a strongly semismooth function. Since the semi-smoothness is preserved under linear transformation, the function g in (D.1) is also strongly semismooth. Then the remaining condition is on the non-singularity of $V \in \partial g(\widehat{\boldsymbol{\theta}})$, where ∂ denotes the Clarke's generalized Jacobian of g. According to (D.5), we have

$$\partial g(\widehat{\boldsymbol{\theta}}) = \left\{ \sum_{i=1}^n W_i \boldsymbol{X}_i \boldsymbol{X}_i^\intercal \{ I(|Y_i - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\theta}}| < \tau) + v_i I(|Y_i - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\theta}}| = \pm\tau) \} : v_i \in [0,1] \right\}.$$

We note that $\mathbf{H}$ in (D.2) is also a member of $\partial g(\widehat{\boldsymbol{\theta}})$. The non-singularity condition depends on the realization of random weights W_i and $\widehat{\boldsymbol{\theta}}$. However, since the dimension d is small as compared to n, W_i and $(Y_i, \boldsymbol{X}_i)$ are IID random variables, as long as $\widehat{\boldsymbol{\theta}}$ is not too extreme (e.g. there are at least d terms such that $|Y_i - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\theta}}| < \tau$), the non-singularity condition will be easily satisfied.

APPENDIX E: SELECTING ROBUSTIFICATION PARAMETER: A DATA-DRIVEN APPROACH

E.1. Preliminaries. Let X be a real-valued random variable with finite variance. For $z \ge 0$, define

$$G(z) = \mathbb{P}(|X| > z),\ P(z) = \mathbb{E}\{X^2 I(|X| \le z)\},\ Q(z) = \mathbb{E}\{\psi_z^2(X)\}, \tag{E.1}$$

where $\psi_z(x) = (|x| \wedge z)\text{sgn}(x)$, $x \in \mathbb{R}$. Moreover, for $z > 0$, we define

$$p(z) = z^{-2} P(z) \quad \text{and} \quad q(z) = z^{-2} Q(z). \tag{E.2}$$

It is easy to see that $Q(z) = P(z) + z^2 G(z)$ and $q(z) = p(z) + G(z)$. 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 E.1. Let functions G, Q, p and q be given in (E.1) and (E.2).

(i) The function $Q : [0, \infty) \to \mathbb{R}$ is non-decreasing with $\lim_{z\to\infty} Q(z) = \mathbb{E}(X^2)$. For any $z > 0$, we have

$$Q(z) = 2\int_0^z yG(y)\,dy, \quad q'(z) = -2\frac{p(z)}{z}, \tag{E.3}$$

and

$$q(z) = \mathbb{P}(X \neq 0) - 2\int_0^z \frac{p(y)}{y} dy. \tag{E.4}$$

(ii) The 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,

$$q(s) = \mathbb{P}(X \neq 0) \tag{E.5}$$

for all $0 \leq s \leq \Delta := \inf\{y > 0 : G(y) < \mathbb{P}(X \neq 0)\}$, and $q(s)$ decreases strictly and continuously on (Δ, ∞) with $\lim_{z\to\infty} q(z) = 0$.

PROOF OF LEMMA E.1. Note that

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

Taking expectations on both sides implies $Q(z) = \mathbb{E}(|X|\wedge z)^2 = 2\int_0^z \mathbb{P}(|X| > y)y\,dy = 2\int_0^z yG(y)dy$, as stated. It follows that $Q'(z) = 2zG(z)$ and thus Q is non-decreasing. Moreover, by the monotone convergence theorem we see that $\lim_{z\to\infty} Q(z) = \mathbb{E}(X^2)$.

Next, taking derivatives with respect to z on both sides of (E.2) gives $2zq(z) + z^2 q'(z) = 2zG(z) = 2z\{q(z) - p(z)\}$, which proves the the second equation in (E.3). To prove (E.4), note that, for any $0 < s < z$, $q(z) = q(s) - 2\int_s^z y^{-1}p(y)\,dy$. On event $\{X \neq 0\}$, it holds almost surely that

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

By the dominated convergence theorem,

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

as $s \to 0$. In the equation $q(z) = q(s) - 2\int_s^z y^{-1}p(y)\,dy$ for $0 < s < z$, letting s tend to zero proves (E.4).

Move to part (ii), by the definition of Δ, we have $\mathbb{P}(0 < |X| \le y) = 0$ and thus $p(y) = 0$ for all $0 < y < \Delta$. This, together with (E.4), implies $q(s) = \mathbb{P}(X \neq 0) > 0$ for all $0 \le s \le \Delta$. It is easy to see that $p(y) > 0$ for any $y > \Delta$, and therefore $q(\cdot)$ is strictly decreasing on (Δ, ∞). Finally, note that

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

By the dominated convergence theorem, $\lim_{z\to\infty} q(z) = 0$ as desired. □

E.2. Catoni's lower bound of sample mean. Let $X_1, \ldots, X_n$ be IID random variables from X with mean zero and variance $\sigma^2 > 0$. Let $\mathcal{A}_{\sigma^2}$ be the set of probability measures on the real line with variance bounded by σ^2. Catoni (2012) proved a lower bound for the deviations of the empirical mean $\bar{X}_n$ when the underlying distribution is the least favorable in $\mathcal{A}_{\sigma^2}$: for any $t \ge 2e$, there exists some distribution with mean zero and variance σ^2 such that the IID sample of size drawn from it satisfies

$$|\bar{X}_n| \ge \left(1 - \frac{1}{n}\right)^{(n-1)/2} \sigma\sqrt{\frac{t}{2n}}$$

with probability at least $2t^{-1}$. This shows that the worst case deviations of $\bar{X}_n$ are suboptimal with heavy-tailed data.

E.3. Proof of Proposition 3.1. For any $\tau > 0$, note that $\psi_\tau(X_i)$'s are independent random variables satisfying $|\psi_\tau(X_i)| \le \tau$ and $\mathbb{E}\psi_\tau^2(X_i) = \sigma_\tau^2$. By Bernstein's inequality,

$$|\widehat{m}_\tau - \mu_\tau| \le \sigma_\tau\sqrt{\frac{2t}{n}} + \frac{\tau t}{3n}$$

with probability at least $1 - 2e^{-t}$. Taking $\tau = \tau_t$ in the last display leads to the first inequality in (3.9), which, together with (3.6), proves the second one.

To prove (3.10), we first make 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 $\tau_t/2 \le \tau \le 3\tau_t/2$ with $\tau_t = \sigma_{\tau_t}\sqrt{n/t}$, there exists some

$1 \le k \le n$ such that $|\tau - \tau_{t,k}| \le \sigma_{\tau_t}(nt)^{-1/2}$, where $\tau_{t,k} := c_k\sigma_{\tau_t}\sqrt{n/t}$. It follows that

$$\sup_{\tau_t/2\le\tau\le 3\tau_t/2} |\widehat{m}_\tau| \le \max_{1\le k\le n} |\widehat{m}_{\tau_{t,k}}| + \frac{\sigma_{\tau_t}}{\sqrt{nt}}. \tag{E.6}$$

For every $1 \le k \le n$, we have

$$|\widehat{m}_{\tau_{t,k}} - \mu_{\tau_{t,k}}| \le \sigma_{\tau_{t,k}}\sqrt{\frac{2t}{n}} + \frac{\tau_{t,k}}{3}\frac{t}{n}$$

with probability at least $1 - 2e^{-t}$. By (3.6), $|\mu_{\tau_{t,k}}| \le (\sigma^2 - \sigma^2_{\tau_{t,k}})/\tau_{t,k}$. Apply the union bound over $1 \le k \le n$ to see that

$$\max_{1\le k\le n} |\widehat{m}_{\tau_{t,k}}| \le \max_{1\le k\le n} \left(\sigma_{\tau_{t,k}}\sqrt{2} + \frac{c_k\sigma_{\tau_t}}{3} + \frac{\sigma^2}{c_k\sigma_{\tau_t}} - \frac{\sigma^2_{\tau_{t,k}}}{c_k\sigma_{\tau_t}}\right)\sqrt{\frac{t}{n}} \tag{E.7}$$

with probability at least $1 - 2ne^{-t}$. Together, (E.6) and (E.7) prove (3.10). □

E.4. Proof of Proposition 3.2. Using the notation in Section E.1, equation (3.8) can be written as $q(\tau) = t/n$. By Lemma E.1, the function q satisfies $\max_{z\ge 0} q(z) = \lim_{z\to 0} q(z) = \mathbb{P}(|X| > 0)$, $\lim_{z\to\infty} q(z) = 0$ and is strictly decreasing on (Δ, ∞). Provided $t/n < \mathbb{P}(|X| > 0)$, equation (3.8) has a unique solution that lies in (Δ, ∞).

By definition, this unique solution τ_t satisfies

$$\tau_t^2 = \mathbb{E}(X^2 \wedge \tau_t^2)\frac{n}{t} \le \sigma^2\frac{n}{t}. \tag{E.8}$$

On the other hand, note that $\mathbb{E}(X^2 \wedge \tau^2) \ge \tau^2\mathbb{P}(|X| > \tau)$ for any $\tau > 0$. It follows that $\mathbb{P}(|X| > \tau_t) \le t/n$, which implies $\tau_t \ge q_{t/n}$. Substituting this into (E.8) gives $\tau_t^2 \ge \mathbb{E}(X^2 \wedge q^2_{t/n})(n/t)$.

To prove Part (ii), recall that $q(\tau_t) = t/n$. Since $t/n \to 0$ and $q(z)$ strictly decreases to zero as $z \to \infty$, we have $\tau_t \to \infty$ and therefore $\mathbb{E}(X^2 \wedge \tau_t^2) \to \sigma^2$ as $n \to \infty$. □

E.5. Proof of Theorem 3.3. By Proposition 3.3, $\widehat{\tau}_t$ is uniquely determined and positive on the event $\{t < \sum_{i=1}^n I(|X_i| > 0)\}$. Under the condition $\mathbb{P}(X = 0) = 0$ and when $t < n$, this event occurs with probability one. We divide the rest of the proof into four steps.

STEP 1. Define functions

$$p_n(z) = \frac{1}{n}\sum_{i=1}^n \frac{X_i^2 I(|X_i| \le z)}{z^2} \quad \text{and} \quad q_n(z) = \frac{1}{n}\sum_{i=1}^n \frac{\psi_z^2(X_i)}{z^2}, \quad z > 0.$$

Applying Lemma E.1 to p_n and q_n implies $q_n'(z) = -2z^{-1}p_n(z)$. Therefore,

$$q_n(\tau_t) - q_n(\widehat{\tau}_t) = 2\int_{\tau_t}^{\widehat{\tau}_t} \frac{p_n(z)}{z}dz = 2\int_0^{(\widehat{\tau}_t-\tau_t)/\tau_t} \frac{p_n(\tau_t+\tau_t u)}{1+u}du$$

by change of variables $u = (z-\tau_t)/\tau_t$. By definition, $q_n(\widehat{\tau}_t) = t/n = q(\tau_t)$. It then follows that

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

For any $r \in (0,1)$, it holds on the event $\{(\widehat{\tau}_t-\tau_t)/\tau_t \geq r\}$ that

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

Similarly, on the event $\{(\widehat{\tau}_t-\tau_t)/\tau_t \leq -r\}$, it holds

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

Putting the above calculations together, we arrive at

$$\begin{aligned} &\mathbb{P}(|\widehat{\tau}_t/\tau_t - 1| \geq r) \\ &\leq \mathbb{P}\{q_n(\tau_t) - q(\tau_t) \geq D_1 + R_1\} + \mathbb{P}\{q_n(\tau_t) - q(\tau_t) \leq -D_2 + R_2\}. \end{aligned} \tag{E.9}$$

Set $\zeta_i = (X_i^2 \wedge \tau_t^2)/\tau_t^2$ such that $q_n(\tau_t) - q(\tau_t) = (1/n)\sum_{i=1}^n \{\zeta_i - \mathbb{E}(\zeta_i)\}$. Note that $0 \leq \zeta_i \leq 1$ and $\mathbb{E}(\zeta_i^2) \leq \mathbb{E}(X_i^2 \wedge \tau_t^2)/\tau_t^2 = t/n$. By Bernstein's inequality, for any $u > 0$ it holds

$$\mathbb{P}\{q_n(\tau_t) - q(\tau_t) \geq u/n\} \leq \exp\{-u^2/(2t + 2u/3)\}. \tag{E.10}$$

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 $0 < u < t$,

$$\mathbb{P}\{q_n(\tau_t) - q(\tau_t) \leq -u/n\} \leq \exp\{-u^2/(2t)\}. \tag{E.11}$$

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

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

are bounded, non-negative random variables satisfying

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

In addition,

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

and

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

Recall that $q(\tau_t) = t/n$. Again, by Theorem 2.19 in de la Peña, Lai and Shao (2009) we have, for any $v > 0$,

$$\mathbb{P}(R_1 \le -2rv/n) \le \exp\{-v^2/(2t)\} \tag{E.12}$$

and

$$\mathbb{P}\{R_2 \ge 2rv/(1-r)n\} \le \exp\{-v^2/(2t)\}. \tag{E.13}$$

STEP 3 (Bounding D_1 and D_2). Starting with D_1, by Lemma E.1 we have

$$\begin{aligned} D_1 = q(\tau_t) - q(\tau_t+\tau_t r) &= 2\int_{\tau_t}^{\tau_t(1+r)} \frac{P(u)}{u^3}du \\ &\ge 2P(\tau_t)\int_{\tau_t}^{\tau_t(1+r)} \frac{1}{u^3}du = \frac{r^2+2r}{(1+r)^2}\frac{P(\tau_t)}{\tau_t^2}. \end{aligned} \tag{E.14}$$

Similarly,

$$D_2 = q(\tau_t - \tau_t r) - q(\tau_t) = 2\int_{\tau_t(1-r)}^{\tau_t} \frac{P(u)}{u^3}du \ge \frac{2r-r^2}{(1-r)^2}\frac{P(\tau_t-\tau_t r)}{\tau_t^2}. \tag{E.15}$$

STEP 4. Together, (E.9) and (E.12)–(E.15) imply that, for any $0 < r < 1$ and $v > 0$,

$$\begin{aligned}
&\mathbb{P}(|\widehat{\tau}_t/\tau_t - 1| \geq r) \\
&\leq 2\exp\{-v^2/(2t)\} + \mathbb{P}\bigg\{q_n(\tau_t) - q(\tau_t) \geq \frac{r^2+2r}{(1+r)^2}\frac{P(\tau_t)}{\tau_t^2} - \frac{2rv}{n}\bigg\} \\
&\quad + \mathbb{P}\bigg\{q_n(\tau_t) - q(\tau_t) \leq -\frac{2r-r^2}{(1-r)^2}\frac{P(\tau_t-\tau_t r)}{\tau_t^2} + \frac{2rv}{(1-r)n}\bigg\}.
\end{aligned} \tag{E.16}$$

Note that

$$\frac{r^2+2r}{(1+r)^2}\frac{P(\tau_t)}{\tau_t^2} - \frac{2rv}{n} = \bigg\{\frac{P(\tau_t)}{Q(\tau_t)}\frac{2+r}{(1+r)^2}t - 2v\bigg\}\frac{r}{n}$$

and

$$\frac{2r-r^2}{(1-r)^2}\frac{P(\tau_t-\tau_t r)}{\tau_t^2} - \frac{2rv}{(1-r)n} = \bigg\{\frac{P(\tau_t-\tau_t r)}{Q(\tau_t)}\frac{2-r}{1-r}t - 2v\bigg\}\frac{r}{(1-r)n}.$$

Taking $v = (a_1 \wedge a_2)t/2$ for a_1 and a_2 as in (3.14), the right-hand side of (E.16) can further be bounded by

$$\mathbb{P}\bigg\{q_n(\tau_t) - q(\tau_t) \geq \frac{a_1 rt}{n}\bigg\} + \mathbb{P}\bigg\{q_n(\tau_t) - q(\tau_t) \leq -\frac{a_2 rt}{n}\bigg\} + 2\exp\{-v^2/(2t)\}.$$

Combining this with (E.10), (E.11) and (E.16) proves (3.13). □

APPENDIX F: ADDITIONAL SIMULATION STUDIES

F.1. Standard deviations of estimated quantiles. In this section, we report the standard deviations of the estimated quantiles for the results in Table 1 and Table 3; see Table 6 and Table 7 below, which correspond to Table 1 and Table 3, respectively. For each setting in Tables 6 and 7, the standard deviation of the estimated quantiles for the boot-Huber is slightly smaller than that for the boot-OLS method. In a sense both the two bootstrap-based methods are rather stable, although the latter one suffers from distorted empirical coverage due to heavy-tailedness. For settings in other tables in the main text, the observations are similar and thus we omit the details.

F.2. Correlated design. In this section, we consider some more challenging cases in which the designs are highly correlated and/or non-Gaussian. Specifically, we consider the following two scenarios:

1. The covariate vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal \in \mathbb{R}^d$ follows a multivariate uniform distribution $\mathrm{Unif}([0,1]^d)$ with $\mathrm{Corr}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \le j \ne k \le d$. See Falk (1999) for the construction of a multivariate uniform distribution. Each component of $\boldsymbol{\theta}^*$ follows a Bernoulli distribution with probability 0.5, i.e. $\mathrm{Ber}(0.5)$. The results for this case are presented in Tables 8 (with $n = 100$) and 9 (with $n = 200$).
2. The covariate vector $\boldsymbol{X}$ follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le d}$ has a Toeplitz structure with $\sigma_{jk} = 0.9^{|j-k|}$. The components of $\boldsymbol{\theta}^*$ are equally spaced in $[0,1]$. The results for this case are presented in Tables 10 (with $n = 100$) and 11 (with $n = 200$).

From Tables 8–11 we find that the average coverage probabilities of the boot-Huber method are in general close to nominal levels, while the boot-OLS leads to severe under-coverage in many heavy-tailed noise settings.

F.3. Simulations on multiple testing. In this section, we evaluate the empirical performance of the proposed robust multiple testing procedure described in Algorithm 2. Recall the multi-response regression model (1.2):

$$y_{ik} = \mu_k + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_k + \varepsilon_{ik}, \quad i = 1, 2, \ldots, n, \quad k = 1, \ldots, m,$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^s$. We choose $\mu_k = \gamma\sigma\sqrt{(2\log m)/n}$ for $1 \le k \le m_1$ with $m_1 = 0.05m$ and $\mu_k = 0$ for $m_1 + 1 \le k \le m$, where $\sigma^2 = \mathrm{var}(\varepsilon_i) = 1$. The parameter γ takes the value either 1.5 (i.e. the weaker signal strength case) or 3 (i.e. the stronger signal strength case). We generate $\{\boldsymbol{x}_i\}_{i=1}^n$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_s)$ and $\boldsymbol{\beta}_k$ from the uniform distribution on $[-1,1]^s$ for $k = 1, \ldots, m$. The settings of error distributions are the same as in Section 5.1. The number of tests m is set to be 1000. The bootstrap weights $\{w_{ik}, 1 \le i \le n, 1 \le k \le m\}$ are IID from $\mathcal{N}(1,1)$. For each setup, we report the average false discovery proportion and empirical power based on 1000 simulations. The FDP nominal level takes value in $\{5\%, 10\%, 15\%, 20\%, 25\%\}$.

Tables 12 and 13 show the empirical FDPs and powers for the weaker signal case with $\mu_k = 1.5\sqrt{2(\log m)/n}$; while Tables 14 and 15 show the results for the stronger signal case with $\mu_k = 3\sqrt{2(\log m)/n}$. Moreover, Tables 12 and 14 consider different error distributions when $n = 100$ and $s = 5$. When the error is from a t-Weibull mixture distribution, Tables 13 and 15 present the results for different combinations of (s, n), revealing the influence of s on the difficulty of the problem. In particular, the combination of $s = 10$, $n = 100$ and signal strength $= 1.5\sqrt{2(\log m)/n}$ corresponds to the most challenging scenario. Increasing either the sample size or the signal strength improves both the FDP control and power performance, which is consistent with our theoretical result in Theorem 4.1. In summary,

with various types of heavy-tailed errors and across different settings, the proposed robust testing procedure performs well and steadily in terms of FDP control and power.

REFERENCES

Adamczak, R., Litvak, A. E., Pajor, A. and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constr. Approx.* **34** 61–88.

Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** 411–420.

Bentkus, V. (2005). A Lyapunov-type bound in R^d. *Theory Probab. Appl.* **49** 311–323.

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

Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans. Inform. Theory* **59** 7711–7717.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185.

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

De Luca, T., Facchinei, F. and Kanzow, C. (1996). A semismooth equation approach to the solution of nonlinear complementarity problems. *Math. Program.* **75** 407–439.

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

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Comm. Statist. Simulation Comput.* **28** 785–791.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265.

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

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

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). *Fundamentals of Convex Analysis*. Springer, Berlin.

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

Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.

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

Pugh, C. C. (2015). *Real Mathematical Analysis*, 2nd ed. Springer-Verlag, New York.

Qi, L. and Sun, D. (1999). A survey of some nonsmooth equations and smoothing Newton methods. In *Progress in Optimization* 121–146. Springer US.

TABLE 6
Standard deviations of the estimated quantiles for $(n, d) = (100, 5)$ *and nominal levels* $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$. *The weights* W_i *are generated from* $\mathcal{N}(1, 1)$

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.210	0.301	0.365	0.412	0.443
	boot-OLS	0.214	0.299	0.365	0.416	0.446
t_ν						
	boot-Huber	0.191	0.295	0.364	0.399	0.442
	boot-OLS	0.222	0.336	0.420	0.467	0.491
Gamma						
	boot-Huber	0.191	0.292	0.366	0.410	0.441
	boot-OLS	0.220	0.309	0.384	0.423	0.456
Wbl mix						
	boot-Huber	0.176	0.265	0.343	0.392	0.435
	boot-OLS	0.201	0.308	0.392	0.440	0.472
Par mix						
	boot-Huber	0.181	0.286	0.358	0.408	0.443
	boot-OLS	0.196	0.324	0.414	0.466	0.488
Logn mix						
	boot-Huber	0.176	0.268	0.331	0.392	0.425
	boot-OLS	0.189	0.327	0.416	0.461	0.488

SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909.

SPOKOINY, V. (2013). Bernstein–von Mises theorem for growing parameter dimension. Preprint. Available at arXiv:1302.3430.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675.

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rate: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205.

VERSHYNIN, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Univ. Press, Cambridge.

WAINWRIGHT, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge.

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. *Ann. Statist.* **46** 1904–1931.

TABLE 7
Standard deviations of the estimated quantiles for $(n, d) = (200, 5)$ *and nominal levels* $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$. *The weights* W_i *are generated from* $\mathcal{N}(1, 1)$

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.232	0.333	0.390	0.427	0.456
	boot-OLS	0.236	0.340	0.388	0.431	0.455
t_ν						
	boot-Huber	0.205	0.291	0.357	0.407	0.445
	boot-OLS	0.236	0.342	0.437	0.481	0.497
Gamma						
	boot-Huber	0.212	0.295	0.358	0.401	0.440
	boot-OLS	0.232	0.307	0.366	0.415	0.451
Wbl mix						
	boot-Huber	0.194	0.292	0.364	0.409	0.439
	boot-OLS	0.220	0.335	0.409	0.447	0.480
Par mix						
	boot-Huber	0.168	0.252	0.333	0.395	0.433
	boot-OLS	0.189	0.314	0.415	0.469	0.495
Logn mix						
	boot-Huber	0.232	0.314	0.379	0.418	0.447
	boot-OLS	0.245	0.363	0.452	0.491	0.500

TABLE 8
Average coverage probabilities for $(n, d) = (100, 5)$ *and nominal levels* $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$ *when* $\boldsymbol{X}_i$ *are IID from a multivariate uniform distribution. Each component of* $\boldsymbol{\theta}^*$ *follows* $Ber(0.5)$, *and* W_i *are generated from* $\mathcal{N}(1, 1)$.

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.946	0.898	0.848	0.781	0.725
	boot-OLS	0.944	0.898	0.843	0.782	0.720
t_ν						
	boot-Huber	0.968	0.919	0.877	0.825	0.777
	boot-OLS	0.954	0.881	0.803	0.729	0.642
Gamma						
	boot-Huber	0.961	0.911	0.868	0.812	0.751
	boot-OLS	0.958	0.900	0.842	0.778	0.716
Wbl mix						
	boot-Huber	0.963	0.907	0.866	0.808	0.748
	boot-OLS	0.947	0.880	0.817	0.724	0.663
Par mix						
	boot-Huber	0.974	0.928	0.882	0.842	0.775
	boot-OLS	0.963	0.897	0.815	0.715	0.634
Logn mix						
	boot-Huber	0.972	0.936	0.888	0.834	0.780
	boot-OLS	0.962	0.901	0.804	0.701	0.615

TABLE 9

Average coverage probabilities for $(n, d) = (200, 5)$ and nominal levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$ when $\boldsymbol{X}_i$ are IID from a multivariate uniform distribution. Each component of $\boldsymbol{\theta}^*$ follows Ber(0.5), and W_i are generated from $\mathcal{N}(1, 1)$.

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.953	0.893	0.844	0.799	0.743
	boot-OLS	0.955	0.893	0.846	0.798	0.744
t_ν						
	boot-Huber	0.960	0.910	0.850	0.795	0.750
	boot-OLS	0.948	0.860	0.759	0.657	0.586
Gamma						
	boot-Huber	0.949	0.896	0.836	0.782	0.729
	boot-OLS	0.947	0.886	0.817	0.765	0.704
Wbl mix						
	boot-Huber	0.957	0.906	0.861	0.811	0.766
	boot-OLS	0.941	0.879	0.805	0.722	0.656
Par mix						
	boot-Huber	0.963	0.924	0.862	0.798	0.743
	boot-OLS	0.958	0.869	0.751	0.669	0.581
Logn mix						
	boot-Huber	0.958	0.909	0.849	0.795	0.739
	boot-OLS	0.947	0.861	0.723	0.600	0.531

TABLE 10

Average coverage probabilities for $(n, d) = (100, 5)$ and nominal levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$ when $\boldsymbol{X}_i$ are IID from a multivariate normal distribution with Toplitz covariance structure and $\boldsymbol{\theta}^*$ is a vector of equally spaced points in $[0, 1]$. The weights W_i are generated from $\mathcal{N}(1, 1)$.

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.954	0.899	0.842	0.783	0.732
	boot-OLS	0.952	0.901	0.842	0.778	0.726
t_ν						
	boot-Huber	0.962	0.904	0.843	0.802	0.734
	boot-OLS	0.948	0.870	0.772	0.678	0.594
Gamma						
	boot-Huber	0.962	0.906	0.841	0.786	0.736
	boot-OLS	0.949	0.893	0.820	0.767	0.706
Wbl mix						
	boot-Huber	0.968	0.924	0.864	0.811	0.747
	boot-OLS	0.958	0.894	0.811	0.737	0.665
Par mix						
	boot-Huber	0.966	0.910	0.849	0.790	0.733
	boot-OLS	0.960	0.881	0.780	0.681	0.609
Logn mix						
	boot-Huber	0.968	0.922	0.875	0.811	0.763
	boot-OLS	0.963	0.878	0.778	0.693	0.608

TABLE 11
Average coverage probabilities for $(n, d) = (200, 5)$ and nominal levels $1 - \alpha = [0.95, 0.9, 0.85, 0.8, 0.75]$ when $\boldsymbol{X}_i$ are IID from a multivariate normal distribution with Toplitz covariance structure and $\boldsymbol{\theta}^*$ is a vector of equally spaced points in $[0, 1]$. The weights W_i are generated from $\mathcal{N}(1, 1)$.

Noise	Approach	0.95	0.9	0.85	0.8	0.75
Gaussian						
	boot-Huber	0.943	0.873	0.813	0.761	0.706
	boot-OLS	0.941	0.867	0.815	0.754	0.708
t_ν						
	boot-Huber	0.956	0.907	0.850	0.791	0.729
	boot-OLS	0.941	0.865	0.744	0.639	0.561
Gamma						
	boot-Huber	0.953	0.904	0.849	0.799	0.738
	boot-OLS	0.943	0.895	0.841	0.779	0.717
Wbl mix						
	boot-Huber	0.961	0.906	0.843	0.788	0.739
	boot-OLS	0.949	0.871	0.788	0.724	0.641
Par mix						
	boot-Huber	0.971	0.932	0.873	0.807	0.751
	boot-OLS	0.963	0.889	0.779	0.674	0.572
Logn mix						
	boot-Huber	0.943	0.889	0.826	0.775	0.725
	boot-OLS	0.936	0.844	0.715	0.597	0.516

TABLE 12
Empirical FDP and power for $(n, s) = (100, 5)$. The nominal level α takes value in $\{0.05, 0.10, 0.15, 0.20, 0.25\}$. The signal strength $\mu_k = 1.5\sqrt{2(\log m)/n}$ for $1 \le k \le m_1$.

Noise	α	0.05	0.10	0.15	0.20	0.25
Gaussian						
	FDP	0.027	0.044	0.075	0.106	0.138
	Power	0.935	0.962	0.978	0.986	0.989
t_ν						
	FDP	0.017	0.030	0.053	0.080	0.105
	Power	0.928	0.953	0.969	0.978	0.983
Gamma						
	FDP	0.048	0.076	0.119	0.159	0.197
	Power	0.957	0.981	0.993	0.996	0.999
Wbl mix						
	FDP	0.038	0.060	0.098	0.130	0.160
	Power	0.951	0.977	0.988	0.993	1.000
Par mix						
	FDP	0.033	0.056	0.094	0.129	0.165
	Power	0.998	0.999	1.000	1.000	1.000
Logn mix						
	FDP	0.029	0.067	0.108	0.149	0.184
	Power	0.999	1.000	1.000	1.000	1.000

TABLE 13
Empirical FDP and power for the Wbl mix model. The nominal level α takes value in $\{0.05, 0.10, 0.15, 0.20, 0.25\}$. The signal strength $\mu_k = 1.5\sqrt{2(\log m)/n}$ for $1 \le k \le m_1$.

s	n	α	0.05	0.10	0.15	0.20	0.25
2	100	FDP	0.061	0.098	0.143	0.186	0.232
		Power	0.981	0.991	0.995	0.997	0.998
	200	FDP	0.048	0.101	0.149	0.195	0.242
		Power	0.989	0.995	0.997	0.998	0.998
5	100	FDP	0.038	0.060	0.098	0.130	0.160
		Power	0.951	0.977	0.988	0.993	1.000
	200	FDP	0.056	0.092	0.137	0.180	0.223
		Power	0.938	0.991	0.996	0.997	0.999
10	100	FDP	0.006	0.014	0.022	0.035	0.052
		Power	0.607	0.825	0.897	0.940	0.961
	200	FDP	0.043	0.070	0.110	0.145	0.187
		Power	0.971	0.983	0.989	0.993	0.995

TABLE 14
Empirical FDP and power for $(n, s) = (100, 5)$. The nominal level α takes value in $\{0.05, 0.10, 0.15, 0.20, 0.25\}$. The signal strength $\mu_k = 3\sqrt{2(\log m)/n}$ for $1 \le k \le m_1$.

Noise	α	0.05	0.10	0.15	0.20	0.25
Gaussian						
	FDP	0.015	0.039	0.063	0.093	0.124
	Power	1.000	1.000	1.000	1.000	1.000
t_ν						
	FDP	0.009	0.027	0.046	0.072	0.098
	Power	0.999	1.000	1.000	1.000	1.000
Gamma						
	FDP	0.038	0.063	0.103	0.136	0.178
	Power	1.000	1.000	1.000	1.000	1.000
Wbl mix						
	FDP	0.038	0.049	0.089	0.120	0.156
	Power	0.999	1.000	1.000	1.000	1.000
Par mix						
	FDP	0.037	0.060	0.100	0.135	0.167
	Power	1.000	1.000	1.000	1.000	1.000
Logn mix						
	FDP	0.024	0.067	0.099	0.128	0.157
	Power	1.000	1.000	1.000	1.000	1.000

TABLE 15

Empirical FDP and power for the Wbl mix model. The nominal level α *takes value in* $\{0.05, 0.10, 0.15, 0.20, 0.25\}$. *The signal strength* $\mu_k = 3\sqrt{2(\log m)/n}$ *for* $1 \le k \le m_1$.

s	n	α	0.05	0.10	0.15	0.20	0.25
2	100	FDP	0.042	0.087	0.125	0.179	0.221
		Power	1.000	1.000	1.000	1.000	1.000
	200	FDP	0.049	0.102	0.144	0.187	0.234
		Power	1.000	1.000	1.000	1.000	1.000
5	100	FDP	0.026	0.049	0.089	0.120	0.156
		Power	0.999	1.000	1.000	1.000	1.000
	200	FDP	0.040	0.069	0.089	0.132	0.184
		Power	1.000	1.000	1.000	1.000	1.000
10	100	FDP	0.011	0.014	0.022	0.041	0.054
		Power	0.991	0.995	0.999	0.999	0.999
	200	FDP	0.040	0.069	0.102	0.131	0.166
		Power	1.000	1.000	1.000	1.000	1.000