Supplement to "Smoothed Quantile Regression with Large-Scale Inference"

A Asymptotic mean squared error

In this section, we provide details for the statements in Remark 4.3. Throughout, assume the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ satisfies (i) $|f_{\varepsilon|\boldsymbol{x}}(u)-f_{\varepsilon|\boldsymbol{x}}(v)| \le l_0(\boldsymbol{x})|u-v|$, and (ii) $|f''_{\varepsilon|\boldsymbol{x}}(u)-f''_{\varepsilon|\boldsymbol{x}}(v)| \le l_2(\boldsymbol{x})|u-v|$ for all $u, v \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$, where $\ell_0(\cdot)$ and $\ell_2(\cdot)$ are such that $\max\{\mathbb{E}\ell_0^2(\boldsymbol{x}), \mathbb{E}\ell_2^2(\boldsymbol{x})\} \le C$ for some $C > 0$. The refined Berry-Esseen bound (4.9) in Theorem 4.3 indicates that

$$n^{1/2}\frac{\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* + 0.5\kappa_2 h^2 \mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\rangle}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{J}_h^{-1}\mathbb{E}[\{\mathcal{K}_h(-\varepsilon)-\tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]\mathbf{J}_h^{-1}\boldsymbol{a}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

uniformly over $\boldsymbol{a} \in \mathbb{R}^p$ as $n \to \infty$, provided $\sqrt{(p+\log n)/n} \lesssim h \lesssim \{(p+\log n)/n\}^{1/4}$ and $p + \log n = o((nh)^{1/2})$. With dimension p fixed, the asymptotic mean squared error (AMSE) of $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)$ is thus determined by that of

$$\boldsymbol{g}_h \overset{\mathrm{d}}{=} \mathcal{N}\left(\frac{1}{2}\kappa_2 n^{1/2}h^2\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\},\ \mathbf{J}_h^{-1}\mathbb{E}[\{\mathcal{K}_h(-\varepsilon)-\tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]\mathbf{J}_h^{-1}\right).$$

To prove (4.10), it suffices to derive the asymptotic expression of $\mathbb{E}(\xi^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$, where $\xi = \tau - \mathcal{K}_h(-\varepsilon)$.

Using integration by parts and a change of variable, we have

$$\begin{aligned}\mathbb{E}(\xi^2|\boldsymbol{x}) &= \int_{-\infty}^{\infty}\{\tau - \mathcal{K}(-t/h)\}^2 \mathrm{d}F_{\varepsilon|\boldsymbol{x}}(t)\\ &= \tau^2 - \frac{2}{h}\int_{-\infty}^{\infty}\{\tau - \mathcal{K}(-t/h)\}K(t/h)F_{\varepsilon|\boldsymbol{x}}(t)\,\mathrm{d}t = \tau^2 - 2\int_{-\infty}^{\infty}\{\mathcal{K}(u)-1+\tau\}K(u)F_{\varepsilon|\boldsymbol{x}}(hu)\mathrm{d}u.\end{aligned}$$

Note that $\int_{-\infty}^{\infty} K(u)\mathcal{K}(u)\mathrm{d}u = 1/2$ and $F_{\varepsilon|\boldsymbol{x}}(0) = \tau$. Then, applying a Taylor series expansion of

$F_{\varepsilon|\boldsymbol{x}}(\cdot)$ around 0 yields

$$\begin{aligned}
\mathbb{E}(\xi^2|\boldsymbol{x}) &= \tau^2 - 2\int_{-\infty}^{\infty}\{\mathcal{K}(u) - 1 + \tau\}K(u)\left[\tau + huf_{\varepsilon|\boldsymbol{x}}(0) + hu\int_0^1\{f_{\varepsilon|\boldsymbol{x}}(huw) - f_{\varepsilon|\boldsymbol{x}}(0)\}\mathrm{d}w\right]\mathrm{d}u,\\
&= \tau^2 - 2\tau\int_{-\infty}^{\infty}\{\mathcal{K}(u) - 1 + \tau\}K(u)\mathrm{d}u - 2hf_{\varepsilon|\boldsymbol{x}}(0)\int_{-\infty}^{\infty}u\{\mathcal{K}(u) - 1 + \tau\}K(u)\mathrm{d}u\\
&\quad - 2h\int_{-\infty}^{\infty}\int_0^1 u\{\mathcal{K}(u) - 1 + \tau\}K(u)\{f_{\varepsilon|\boldsymbol{x}}(huw) - f_{\varepsilon|\boldsymbol{x}}(0)\}\mathrm{d}w\mathrm{d}u\\
&= \tau(1-\tau) - 2\bar{\kappa}_1 f_{\varepsilon|\boldsymbol{x}}(0)h - 2h\int_{-\infty}^{\infty}\int_0^1 u\{\mathcal{K}(u) - 1 + \tau\}K(u)\{f_{\varepsilon|\boldsymbol{x}}(huw) - f_{\varepsilon|\boldsymbol{x}}(0)\}\mathrm{d}w\mathrm{d}u,
\end{aligned}$$

where $\bar{\kappa}_1 := \int_{-\infty}^{\infty} uK(u)\mathcal{K}(u)\mathrm{d}u > 0$. Consequently,

$$\left\|\mathbb{E}(\xi^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) - \tau(1-\tau)\boldsymbol{\Sigma} + 2\bar{\kappa}_1 h\mathbf{J}\right\|_2 \lesssim \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\mathbb{E}\{l_0(\boldsymbol{x})\langle\boldsymbol{x},\boldsymbol{u}\rangle^2\}\cdot h^2 \lesssim \{m_4\cdot\mathbb{E}l_0^2(\boldsymbol{x})\}^{1/2}h^2.$$

Putting together the pieces we conclude that

$$\begin{aligned}
\mathbb{E}(\boldsymbol{g}_h\boldsymbol{g}_h^{\mathrm{T}}) &= \mathbf{J}_h^{-1}\mathbb{E}(\xi^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\mathbf{J}_h^{-1} + \frac{1}{4}\kappa_2^2 nh^4\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}^{\mathrm{T}}\}\mathbf{J}_h^{-1}\\
&= \mathbf{J}_h^{-1}\left[\tau(1-\tau)\boldsymbol{\Sigma} - 2\bar{\kappa}_1 h\mathbf{J} + \frac{1}{4}\kappa_2^2 nh^4\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}^{\mathrm{T}}\} + O(h^2)\right]\mathbf{J}_h^{-1}.
\end{aligned}$$

Under the prescribed bandwidth constraint, the leading term of the MSE of $\boldsymbol{g}_h$, and thus of the AMSE of $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)$, is

$$\mathbf{J}_h^{-1}\boldsymbol{\Sigma}^{1/2}\left[\tau(1-\tau)\mathbf{I}_p - 2\bar{\kappa}_1 h\mathbf{H} + \frac{1}{4}\kappa_2^2 nh^4\mathbb{E}(\boldsymbol{b})\mathbb{E}(\boldsymbol{b}^{\mathrm{T}})\right]\boldsymbol{\Sigma}^{1/2}\mathbf{J}_h^{-1},$$

where $\boldsymbol{b} = f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w} = f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ and $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$. This proves (4.10).

B One-step conquer with higher-order kernels

As noted in Section 4.1, the smoothing bias is of order h^2 when a non-negative kernel is used. The ensuing empirical loss $\boldsymbol{\beta} \mapsto (1/n)\sum_{i=1}^n(\rho_\tau * K_h)(y_i - \langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)$ is not only twice-differentiable and convex, but also (provably) strongly convex in a local vicinity of $\boldsymbol{\beta}^*$ with high probability. Kernel smoothing is ubiquitous in nonparametric statistics. The order of a kernel, ν, is defined as the order of the first non-zero moment. The order of a symmetric kernel is always even. A kernel is called *high-order* if $\nu > 2$, which inevitably has negative parts and thus is no longer a probability density. Thus far we have focused on conquer with second-order kernels, and the resulting estimator achieves

an ℓ_2-error of the order $\sqrt{p/n} + h^2$.

Let $G(\cdot)$ be a higher-order symmetric kernel with order $\nu \geq 4$, and $b > 0$ be a bandwidth. Again, via convolution smoothing, we may consider a bias-reduced estimator that minimizes the empirical loss $\boldsymbol{\beta} \mapsto \widehat{Q}_b^G(\boldsymbol{\beta}) := (1/n)\sum_{i=1}^n (\rho_\tau * G_b)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$. This, however, leads to a non-convex optimization. Without further assumptions, finding a global minimum is computationally intractable: finding an ϵ-suboptimal point for a k-times continuously differentiable loss function requires at least $\Omega\{(1/\epsilon)^{p/k}\}$ evaluations of the function and its first k derivatives, ignoring problem-dependent constants; see Section 1.6 in Nemirovski and Yudin (1983). Instead, various gradient-based methods have been developed for computing *stationary points*, which are points $\boldsymbol{\beta}$ with sufficiently small gradient $\|\nabla \widehat{Q}_b^G(\boldsymbol{\beta})\|_2 \leq \epsilon$, where $\epsilon \geq 0$ is optimization error. However, the equation $\nabla \widehat{Q}_b^G(\boldsymbol{\beta}) = \mathbf{0}$ does not necessarily have a unique solution, whose statistical guarantees remain unknown.

Motivated by the classical one-step estimator (Bickel, 1975), we further propose a one-step conquer estimator using high-order kernels, which bypasses solving a large-scale non-convex optimization. To begin with, we choose two symmetric kernel functions, $K : \mathbb{R} \mapsto [0, \infty)$ with order two and $G(\cdot)$ with order $\nu \geq 4$, and let $h, b > 0$ be two bandwidths. First, compute an initial conquer estimator $\overline{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{Q}_h^K(\boldsymbol{\beta})$, where $\widehat{Q}_h^K(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (\rho_\tau * K_h)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$. Denote by $\bar{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \overline{\boldsymbol{\beta}}\rangle$ for $i = 1, \ldots, n$ the fitted residuals. Next, with slight abuse of notation, we define the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$ as a solution to the equation $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) = -\nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$, or equivalently,

$$\left\{\frac{1}{n}\sum_{i=1}^n G_b(\bar{\varepsilon}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right\}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \{\mathcal{G}_b(\bar{\varepsilon}_i) + \tau - 1\}\boldsymbol{x}_i. \tag{B.1}$$

where $\widehat{Q}_b^G(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (\rho_\tau * G_b)(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ and $\mathcal{G}_b(u) = \int_{-\infty}^{u/b} G(v)\,\mathrm{d}v$. Provided that $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$ is positive definite, the one-step conquer estimate $\widehat{\boldsymbol{\beta}}$ essentially performs a Newton-type step based on $\overline{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}} - \{\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})\}^{-1} \nabla \widehat{Q}_b^G(\overline{\boldsymbol{\beta}}). \tag{B.2}$$

In this case, $\widehat{\boldsymbol{\beta}}$ can be computed by the conjugate gradient method (Hestenes and Stiefel, 1952).

Theoretical properties of the one-step estimator $\widehat{\boldsymbol{\beta}}$ defined in (B.1), including the Bahadur representation and asymptotic normality with explicit Berry-Esseen bound, will be provided in on-line supplementary materials. For practical implementation, we consider higher-order Gaussian-based

kernels. For $r = 1, 2, \ldots$, the $(2r)$-th order Gaussian kernels are

$$G(u; 2r) = \frac{(-1)^r \phi^{(2r-1)}(u)}{2^{r-1}(r-1)!u} = \sum_{\ell=0}^{r-1} \frac{(-1)^\ell}{2^\ell \ell!} \phi^{(2\ell)}(u);$$

see Section 2 of Wand and Schucany (1990). Integrating $G(\cdot; 2r)$ yields

$$\mathcal{G}(v; 2r) = \int_{-\infty}^{v} G(u; 2r)\, \mathrm{d}u = \sum_{\ell=0}^{r-1} \frac{(-1)^\ell}{2^\ell \ell!} \phi^{(2\ell-1)}(v).$$

In fact, both $G(\cdot; 2r)$ and $\mathcal{G}(\cdot; 2r)$ have simpler forms $G(u; 2r) = p_r(u)\phi(u)$ and $\mathcal{G}(u; 2r) = \Phi(u) + P_r(u)\phi(u)$, where $p_r(\cdot)$ and $P_r(\cdot)$ are polynomials in u. For example, $p_1(u) = 1$, $P_1(u) = 0$, $p_2(u) = (-u^2 + 3)/2$, $P_2(u) = u/2$, $p_3(u) = (u^4 - 10u^2 + 15)/8$, and $P_3(u) = (-u^3 + 7u)/8$. We refer to Oryshchenko (2020) for more details when r is large.

C Proofs

Recall that $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is such that $x_1 \equiv 1$, $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$, and $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite. In this case,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \mathbf{0}_{p-1}^{\mathrm{T}} \\ \mathbf{0}_{p-1} & \mathbf{S} \end{bmatrix} \quad \text{with} \quad \mathbf{S} = \mathbb{E}(\boldsymbol{x}_{-}\boldsymbol{x}_{-}^{\mathrm{T}}) \quad \text{and} \quad \boldsymbol{w} = \begin{bmatrix} 1 \\ \mathbf{S}^{-1/2}\boldsymbol{x}_{-} \end{bmatrix},$$

where $\mathbf{0}_k$ is the zero vector in $\mathbb{R}^k$ $(k \geq 2)$. For every $r \geq 0$, define the ellipse $\Theta(r) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} \leq r\}$ and its boundary $\partial\Theta(r) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_{\boldsymbol{\Sigma}} = r\}$.

C.1 Proof of Proposition 4.1

To begin with, define $\boldsymbol{\delta}_h = \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \in \mathbb{R}^p$ and $\delta_h = \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}$. By the convexity of $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ and the first-order optimality condition $\nabla Q_h(\boldsymbol{\beta}_h^*) = \mathbf{0}$, we have

$$0 \leq \langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \rangle = \langle -\nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\delta}_h \rangle \leq \|\boldsymbol{\Sigma}^{-1/2} \nabla Q_h(\boldsymbol{\beta}^*)\|_2 \cdot \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}, \tag{C.1}$$

where the last step follows from Hölder's inequality. Note that $\nabla Q_h(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathcal{K}(-\varepsilon/h) - \tau\}\boldsymbol{x}$. By integration by parts and a Taylor series expansion,

$$\begin{aligned}\mathbb{E}\{\mathcal{K}(-\varepsilon/h)|\boldsymbol{x}\} &= \int_{-\infty}^{\infty} \mathcal{K}(-t/h)\,\mathrm{d}F_{\varepsilon|\boldsymbol{x}}(t) \\ &= -\frac{1}{h}\int_{-\infty}^{\infty} K(-t/h)F_{\varepsilon|\boldsymbol{x}}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} K(u)F_{\varepsilon|\boldsymbol{x}}(-hu)\,\mathrm{d}u \\ &= \tau + \int_{-\infty}^{\infty} K(u)\int_0^{-hu}\{f_{\varepsilon|\boldsymbol{x}}(t) - f_{\varepsilon|\boldsymbol{x}}(0)\}\,\mathrm{d}t\,\mathrm{d}u,\end{aligned}$$

from which it follows that $|\mathbb{E}\{\mathcal{K}(-\varepsilon/h)|\boldsymbol{x}\} - \tau| \le 0.5 l_0\kappa_2 h^2$. Consequently,

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\{\mathcal{K}(-\varepsilon/h) - \tau\}\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle \le 0.5 l_0\kappa_2 h^2. \tag{C.2}$$

Turning to the left-hand side of (C.1), applying the mean value theorem for vector-valued functions implies

$$\nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*) = \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta}_h)\,\mathrm{d}t\,\boldsymbol{\delta}_h, \tag{C.3}$$

where $\nabla^2 Q_h(\boldsymbol{\beta}) = \mathbb{E}\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$. With $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, note that

$$\mathbb{E}\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)|\boldsymbol{x}\} = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{u - \langle \boldsymbol{x}, \boldsymbol{\delta}\rangle}{h}\right) f_{\varepsilon|\boldsymbol{x}}(u)\,\mathrm{d}u = \int_{-\infty}^{\infty} K(v) f_{\varepsilon|\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle + hv)\,\mathrm{d}v.$$

By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$,

$$\mathbb{E}\{K_h(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)|\boldsymbol{x}\} = f_{\varepsilon|\boldsymbol{x}}(0) + R_h(\boldsymbol{\delta}) \tag{C.4}$$

with $R_h(\boldsymbol{\delta})$ satisfying $|R_h(\boldsymbol{\delta})| \le l_0(|\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle| + \kappa_1 h)$. Together, (C.3), (C.4) and the assumption $f_{\varepsilon|\boldsymbol{x}}(0) \ge \underline{f} > 0$ (almost surely) yield

$$\begin{aligned}&\langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*),\ \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\rangle \\ &\ge \underline{f}\cdot\|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}^2 - 0.5 l_0\mathbb{E}|\langle \boldsymbol{x}, \boldsymbol{\delta}_h\rangle|^3 - l_0\kappa_1 h\cdot\|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}^2 \ge \underline{f}\cdot\delta_h^2 - 0.5 l_0 m_3\cdot\delta_h^3 - l_0\kappa_1 h\cdot\delta_h^2.\end{aligned} \tag{C.5}$$

Combining (C.1) with the upper and lower bounds (C.2) and (C.5), we find that $\delta_h \ge 0$ satisfies $0.5 l_0 m_3\cdot\delta_h^2 - (\underline{f} - l_0\kappa_1 h)\delta_h + 0.5 l_0\kappa_2 h^2 \ge 0$. Provided that $l_0\{\kappa_1 + (m_3\kappa_2)^{1/2}\}h > \underline{f}$, solving this

inequality yields

$$\delta_h \le \frac{l_0\kappa_2 h^2}{\underline{f} - l_0\kappa_1 h + \Delta_h^{1/2}} \quad \text{or} \quad \delta_h \ge \frac{\underline{f} - l_0\kappa_1 h + \Delta_h^{1/2}}{l_0 m_3}. \tag{C.6}$$

where $\Delta_h := (\underline{f} - l_0\kappa_1 h)^2 - l_0^2 m_3\kappa_2 h^2 > 0$. It remains to rule out the second bound in (C.6).

Assume δ_h satisfies the second bound in (C.6), so that $\delta_h > l_0(m_3\kappa_2)^{1/2}h/(l_0 m_3) = (\kappa_2/m_3)^{1/2}h =: r_0$. Then, there exists some $\eta \in (0,1)$ such that $\widetilde{\boldsymbol{\beta}} := (1-\eta)\boldsymbol{\beta}^* + \eta\boldsymbol{\beta}_h^*$ satisfies $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \eta\delta_h = r_0$. By the convexity of $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ and Lemma C.1 in the supplementary material of Sun, Zhou and Fan (2020),

$$\langle \nabla Q_h(\widetilde{\boldsymbol{\beta}}) - \nabla Q_h(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \le \eta \cdot \langle \nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^* \rangle = \langle -\nabla Q_h(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle.$$

Repeating the above analysis, we find that the right-hand side of the above inequality $\le 0.5 l_0\kappa_2 h^2 \cdot r_0$, and the left-hand side $\ge \underline{f} \cdot r_0^2 - 0.5 l_0 m_3 \cdot r_0^3 - l_0\kappa_1 h \cdot r_0^2 = \{\underline{f} - l_0\kappa_1 h - 0.5 l_0 (m_3\kappa_2)^{1/2} h\} r_0^2$. Canceling out the common factor r_0 from both sides, we obtain

$$r_0 \le \frac{0.5 l_0\kappa_2 h^2}{\underline{f} - l_0\kappa_1 h - 0.5 l_0 (m_3\kappa_2)^{1/2}} < \frac{0.5 l_0\kappa_2 h^2}{0.5 l_0 (m_3\kappa_2)^{1/2} h} = (\kappa_2/m_3)^{1/2} h = r_0,$$

which leads to a contradiction. Consequently, δ_h must satisfy the first bound in (C.6), which in turn implies the claimed result.

Next, to investigate the leading term in the bias, define

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1/2}\{\nabla Q_h(\boldsymbol{\beta}_h^*) - \nabla Q_h(\boldsymbol{\beta}^*) - \mathbf{J}(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*)\} \quad \text{and} \quad \mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\},$$

where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$. Again, by the mean value theorem for vector-valued functions,

$$\boldsymbol{\Delta} = \left\{\boldsymbol{\Sigma}^{-1/2} \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta}_h)\,\mathrm{d}t\, \boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\right\}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_h \quad \text{with} \quad \boldsymbol{\delta}_h = \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*. \tag{C.7}$$

The Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ ensures that

$$
\begin{aligned}
&\left\|\boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta}_h)\,\mathrm{d}t\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\right\|_2 \\
&= \left\|\mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(u)\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\delta}_h\rangle - hu) - f_{\varepsilon|\boldsymbol{x}}(0)\}\,\mathrm{d}u\,\mathrm{d}t\,\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2 \\
&\le l_0 \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\int_0^1\int_{-\infty}^{\infty} K(u)(t|\langle \boldsymbol{x}, \boldsymbol{\delta}_h\rangle| + h|u|)\,\mathrm{d}u\,\mathrm{d}t\,\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2 \\
&\le 0.5 l_0 \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}(|\langle \boldsymbol{x}, \boldsymbol{\delta}_h\rangle|\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2) + l_0\kappa_1 h \\
&\le 0.5 l_0 m_3 \|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}} + l_0\kappa_1 h.
\end{aligned}
$$

This bound, together with (C.7), implies

$$
\|\boldsymbol{\Delta}\|_2 \le l_0(0.5 m_3\|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}} + \kappa_1 h)\|\boldsymbol{\delta}_h\|_{\boldsymbol{\Sigma}}. \tag{C.8}
$$

Moreover, applying a second-order Taylor series expansion to $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ yields

$$
\begin{aligned}
&\mathbb{E}\{\mathcal{K}(-\varepsilon/h)|\boldsymbol{x}\} - \tau \\
&= \int_{-\infty}^{\infty} K(u)\int_0^{-hu}\{f_{\varepsilon|\boldsymbol{x}}(t) - f_{\varepsilon|\boldsymbol{x}}(0)\}\,\mathrm{d}t\,\mathrm{d}u \\
&= 0.5\kappa_2 h^2 \cdot f'_{\varepsilon|\boldsymbol{x}}(0) + \int_{-\infty}^{\infty}\int_0^{-hu}\int_0^t K(u)\{f'_{\varepsilon|\boldsymbol{x}}(v) - f'_{\varepsilon|\boldsymbol{x}}(0)\}\,\mathrm{d}v\,\mathrm{d}t\,\mathrm{d}u.
\end{aligned}
$$

For $\nabla Q_h(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathcal{K}(-\varepsilon/h) - \tau\}\boldsymbol{x}$, it follows that

$$
\left\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*) - \frac{1}{2}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\right\|_2 \le \frac{1}{6} l_1\kappa_3 h^3. \tag{C.9}
$$

Combining (C.8) and (C.9) completes the proof of (4.4). □

C.2 Proof of Theorem 4.1

For every $\boldsymbol{\delta} \in \mathbb{R}^p$, define $\widehat{D}_h(\boldsymbol{\delta}) = \widehat{Q}_h(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \widehat{Q}_h(\boldsymbol{\beta}^*)$, $D_h(\boldsymbol{\delta}) = Q_h(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - Q_h(\boldsymbol{\beta}^*)$, as well as first-order Taylor series remainder terms $\widehat{R}_h(\boldsymbol{\delta}) = \widehat{D}_h(\boldsymbol{\delta}) - \langle \nabla\widehat{Q}_h(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle$ and $R_h(\boldsymbol{\delta}) = D_h(\boldsymbol{\delta}) -$

$\langle \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle$. With these notations, we have

$$\begin{aligned}
\widehat{D}_h(\boldsymbol{\delta}) &= \langle \nabla Q_h(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle + R_h(\boldsymbol{\delta}) + \{\widehat{D}_h(\boldsymbol{\delta}) - D_h(\boldsymbol{\delta})\} \\
&\geq R_h(\boldsymbol{\delta}) - \|\boldsymbol{\Sigma}^{-1/2} \nabla Q_h(\boldsymbol{\beta}^*)\|_2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \\
&\geq R_h(\boldsymbol{\delta}) - 0.5 l_0 \kappa_2 h^2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - \underbrace{\{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\}}_{\text{sampling error}},
\end{aligned} \tag{C.10}$$

where we used (C.2) in the last step. Following the same argument that leads to (C.5), it can be shown that

$$R_h(\boldsymbol{\delta}) \geq \frac{1}{2}(\underline{f} - l_0 \kappa_1 h - 0.5 l_0 m_3 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}) \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \text{ for all } \boldsymbol{\delta} \in \mathbb{R}^p. \tag{C.11}$$

Take $r_0 = (2\kappa_2/m_3)^{1/2} h$ as an intermediate convergence radius. For any $\boldsymbol{\delta} \in \partial\Theta(r_0)$, i.e., $\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} = r_0$, the last two displays imply

$$R_h(\boldsymbol{\delta}) - 0.5 l_0 \kappa_2 h^2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \geq \frac{1}{2}\{\underline{f} - l_0 \kappa_1 h - l_0 (2\kappa_2 m_3)^{1/2} h\} r_0^2 \text{ for all } \boldsymbol{\delta} \in \partial\Theta(r_0). \tag{C.12}$$

To control the last term on the right-hand side of (C.10), the following lemma provides some type of uniform law of large numbers for the zero-mean stochastic process $\{\widehat{D}_h(\boldsymbol{\delta}) - D_h(\boldsymbol{\delta}), \boldsymbol{\delta} \in \Theta(r)\}$, $r > 0$. This implies a form of restricted strong convexity (RSC) of the empirical loss $\widehat{Q}_h(\cdot)$.

Lemma C.1. Given any $r \geq 0$, the bound

$$\sup_{\boldsymbol{\delta} \in \Theta(r)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \leq 3\bar{\tau} \upsilon_0 r \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n} \right) \tag{C.13}$$

holds with probability at least $1 - e^{4p-u}$ for any $u \geq 0$, where $\bar{\tau} = \max(\tau, 1 - \tau)$. In addition, given any $r_u > r_l > 0$, with probability at least $1 - \lceil e \log(\frac{r_u}{r_l}) \rceil e^{4p-u}$ for any $u \geq 0$,

$$D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) \leq 4.25 \bar{\tau} \upsilon_0 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \left(\sqrt{\frac{u}{n}} + \frac{u}{n} \right) \text{ holds for all } \boldsymbol{\delta} \text{ satisfying } r_l \leq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r_u. \tag{C.14}$$

First, applying (C.13) with $r = r_0$ and $u = 4p + t$ yields that, with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\delta} \in \Theta(r_0)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \leq 3\bar{\tau} \upsilon_0 r_0 \left(\sqrt{\frac{4p + t}{n}} + \frac{4p + t}{n} \right). \tag{C.15}$$

Combining this bound with (C.10) and (C.12) implies that with probability at least $1 - e^{-t}$, $\widehat{D}_h(\boldsymbol{\delta}) > 0$

for all $\boldsymbol{\delta} \in \partial\Theta(r_0)$ as long as the bandwidth is subject to $\underline{f}^{-1} m_3^{1/2} \upsilon_0 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. On the other hand, by the optimality of $\widehat{\boldsymbol{\beta}}_h$, $\widehat{\boldsymbol{\delta}} := \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$ satisfies $\widehat{D}_h(\widehat{\boldsymbol{\delta}}) \le 0$. Consequently, the convexity of $\widehat{Q}_h(\cdot)$ ensures that $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le r_0$. See, e.g., Lemma 9.21 in Wainwright (2019).

Next, we refine the convergence rate of $\widehat{\boldsymbol{\beta}}_h$ from r_0 to the claimed one under the above event. Consider the ring-shaped set $\Theta(r_l, r_0) = \{\boldsymbol{\delta} \in \mathbb{R}^p : r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_0\}$ with $r_l = r_0 h$. If $\widehat{\boldsymbol{\delta}} \notin \Theta(r_l, r_0)$, we must have $\widehat{\boldsymbol{\delta}} \in \Theta(r_0 h)$, and thus the claimed bound follows immediately. Hereinafter, we assume $\widehat{\boldsymbol{\delta}} \in \Theta(r_l, r_0)$. Using the second inequality in Lemma C.1 with $(r_l, r_u) = (r_0 h, r_0)$ and $u = \sqrt{\log(e \log h^{-1}) + 4p + t}$, we find that with probability at least $1 - e^{-t}$,

$$D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \cdot \underbrace{4.25\bar{\tau}\upsilon_0 \left\{ \sqrt{\frac{\log(e \log h^{-1}) + 4p + t}{n}} + \frac{\log(e \log h^{-1}) + 4p + t}{n} \right\}}_{=:r_1} \tag{C.16}$$

holds for all $\boldsymbol{\delta} \in \Theta(r_l, r_0)$, hence including $\widehat{\boldsymbol{\delta}}$. Applying this along with the earlier bounds (C.10), (C.11) and the fact $\widehat{D}_h(\widehat{\boldsymbol{\delta}}) \le 0$, we obtain that

$$(\underline{f} - l_0 \kappa_1 h)\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 \le (2r_1 + l_0 \kappa_2 h^2)\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} + 0.5 l_0 m_3 \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^3 \le 2(r_1 + l_0 \kappa_2 h^2)\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}.$$

Canceling out $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}$ proves the claimed bound. □

C.2.1 Proof of Lemma C.1

For each sample $\boldsymbol{z}_i = (\boldsymbol{x}_i, \varepsilon_i)$, define the loss difference $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = \ell_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) - \ell_h(\varepsilon_i)$, so that $\widehat{D}_h(\boldsymbol{\delta}) = (1/n) \sum_{i=1}^n d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$. By the Lipschitz continuity of $u \mapsto \ell_h(u)$, $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle$. That is, for any $\boldsymbol{z}_i$ and $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathbb{R}^p$, $|d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) - d_h(\boldsymbol{\delta}'; \boldsymbol{z}_i)| \le \bar{\tau} |\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle - \langle \boldsymbol{x}_i, \boldsymbol{\delta}' \rangle|$.

For any given $r > 0$ and some $\epsilon \in (0, 1)$ to be determined, define the random variable $\Delta_\epsilon(r) = n(1-\epsilon) \sup_{\boldsymbol{\delta} \in \Theta(r)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\}/(2\bar{\tau} r)$, where $D_h(\boldsymbol{\delta}) = \mathbb{E}\widehat{D}_h(\boldsymbol{\delta})$. By Chernoff's inequality, for any $u \ge 0$,

$$\mathbb{P}\{\Delta_\epsilon(r) \ge u\} \le \exp\left[- \sup_{\lambda \ge 0} \{\lambda u - \log \mathbb{E} e^{\lambda \Delta_\epsilon(r)}\} \right]. \tag{C.17}$$

To control the moment generating function $\mathbb{E} e^{\lambda \Delta_\epsilon(r)}$, by Rademacher symmetrization we have

$$\mathbb{E} e^{\lambda \Delta_\epsilon(r)} \le \mathbb{E} \exp\left\{ 2\lambda(1-\epsilon) \sup_{\boldsymbol{\delta} \in \Theta(r)} \frac{1}{2\bar{\tau} r} \sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Recall that $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$ is $\bar{\tau}$-Lipschitz continuous in $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle$, and $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = 0$ if $\langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle = 0$. Applying the Ledoux-Talagrand contraction inequality (see Theorem 4.12 and inequality (4.10) in Ledoux and Talagrand (1991)) yields

$$\begin{aligned}
&\mathbb{E} \exp\left\{ 2\lambda(1-\epsilon) \sup_{\boldsymbol{\delta} \in \Theta(r)} \frac{1}{2\bar{\tau} r} \sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) \right\} \\
&\leq \mathbb{E} \exp\left\{ \frac{\lambda}{r}(1-\epsilon) \sup_{\boldsymbol{\delta} \in \Theta(r)} \sum_{i=1}^n e_i \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle \right\} \leq \mathbb{E} \exp\left\{ \lambda(1-\epsilon) \left\| \sum_{i=1}^n e_i \boldsymbol{w}_i \right\|_2 \right\},
\end{aligned}$$

where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}_i$. For this $\epsilon \in (0, 1)$, there exists an ϵ-net $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{N_\epsilon}\}$ of $\mathbb{S}^{p-1}$ with cardinality $N_\epsilon \leq (1 + 2/\epsilon)^p$ such that $\| \sum_{i=1}^n e_i \boldsymbol{w}_i \|_2 \leq (1-\epsilon)^{-1} \max_{1 \leq j \leq N_\epsilon} \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i$. This implies

$$\mathbb{E} \exp\left\{ \lambda(1-\epsilon) \left\| \sum_{i=1}^n e_i \boldsymbol{w}_i \right\|_2 \right\} \leq \sum_{j=1}^{N_\epsilon} \mathbb{E} \exp\left\{ \lambda \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i \right\}.$$

Write $S_j = \sum_{i=1}^n e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i$, which is a sum of zero-mean random variables. Note that $e_i \in \{-1, 1\}$ is symmetric, and Condition 4.4 ensures that for any $k \geq 3$, $\mathbb{E}|\boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i|^k \leq \upsilon_0^k k \int_0^\infty t^{k-1} e^{-t} \mathrm{d}t = \upsilon_0^k k!$. Hence, for every $0 \leq c < 1/\upsilon_0$,

$$\begin{aligned}
\mathbb{E} e^{c e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i} &= 1 + \frac{c^2}{2} \mathbb{E}(e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i)^2 + \sum_{\ell=3}^\infty \frac{c^\ell}{\ell!} \mathbb{E}(e_i \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{w}_i)^\ell \\
&\leq 1 + \frac{c^2}{2} + \sum_{\ell=2}^\infty \frac{c^{2\ell}}{(2\ell)!} \upsilon_0^{2\ell} (2\ell)! \leq 1 + \frac{c^2}{2} + \sum_{\ell=2}^\infty (c^2 \upsilon_0^2)^\ell \leq 1 + \frac{\upsilon_0^2}{2} \sum_{\ell \geq 2} c^\ell (\sqrt{2} \upsilon_0)^{\ell-2}.
\end{aligned}$$

It follows that for every $0 < \lambda < 1/(\sqrt{2}\upsilon_0)$ and $j = 1, \ldots, N_\epsilon$,

$$\log \mathbb{E} e^{\lambda S_j} \leq \frac{n \upsilon_0^2 \lambda^2}{2(1 - \sqrt{2}\upsilon_0 \lambda)} \quad \text{and thus} \quad \log \mathbb{E} e^{\lambda \Delta_\epsilon(r)} \leq \log N_\epsilon + \frac{n \upsilon_0^2 \lambda^2}{2(1 - \sqrt{2}\upsilon_0 \lambda)}.$$

For any $u \geq 0$, note that

$$\sup_{\lambda \geq 0} \{ \lambda u - \log \mathbb{E} e^{\lambda \Delta_\epsilon(r)} \} \geq -\log N_\epsilon + \sup_{\lambda \in (0, (\sqrt{2}\upsilon_0)^{-1})} \left\{ \lambda u - \frac{n \upsilon_0^2 \lambda^2}{2(1 - \sqrt{2}\upsilon_0 \lambda)} \right\}$$

Substituting this into (C.17), and following the proof of Bernstein's inequality (see, e.g., Theorem 2.10 in Boucheron, Lugosi and Massart (2013)), it can be shown that with probability at least

$1 - \exp\{p\log(1+2/\epsilon) - u\}$,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \le \frac{2\sqrt{2}}{1-\epsilon}\bar{\tau}\upsilon_0 r \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right). \tag{C.18}$$

This proves (C.13) by taking $\epsilon = 2/(e^4 - 1)$.

Next, we prove the uniform bound (C.14), which holds for all $\boldsymbol{\delta} \in \Theta(r_l, r_u) := \{\boldsymbol{v} \in \mathbb{R}^p : r_l \le \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}} \le r_u\}$, via a peeling argument. For some $\gamma > 1$ (to be specified) and positive integers $k = 1, \ldots, N := \lceil \log(\frac{r_u}{r_l})/\log(\gamma)\rceil$, define the sets $\Theta_k = \{\boldsymbol{v} \in \mathbb{R}^p : \gamma^{k-1} r_l \le \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}} \le \gamma^k r_l\}$, so that $\Theta(r_l, r_u) \subseteq \cup_{k=1}^N \Theta_k$. Then,

$$\begin{aligned}
&\mathbb{P}\left\{\exists \boldsymbol{\delta} \in \Theta(r_l, r_u) \text{ s.t. } D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}\gamma}{1-\epsilon}\bar{\tau}\upsilon_0 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right)\right\} \\
&\le \sum_{k=1}^N \mathbb{P}\left\{\exists \boldsymbol{\delta} \in \Theta_k \text{ s.t. } D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}\gamma}{1-\epsilon}\bar{\tau}\upsilon_0 \gamma^{k-1} r_l \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right)\right\} \\
&\le \sum_{k=1}^N \mathbb{P}\left\{\sup_{\boldsymbol{\delta}\in\Theta(\gamma^k r_l)} D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) > \frac{2\sqrt{2}}{1-\epsilon}\bar{\tau}\upsilon_0 \gamma^k r_l \cdot \left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right)\right\} \\
&\overset{\text{(i)}}{\le} \sum_{k=1}^N \exp\{p\log(1+2/\epsilon) - u\} \le \lceil \log(\tfrac{r_u}{r_l})/\log(\gamma)\rceil \exp\{p\log(1+2/\epsilon) - u\},
\end{aligned}$$

where inequality (i) is obtained by repeatedly using (C.18) with $r = \gamma^k r_l$ for $k = 1, \ldots, N$. Taking $\epsilon = 2/(e^4 - 1)$ and $\gamma = e^{1/e}$ yields the claimed bound (C.14). □

C.3 An alternative proof to Theorem 4.1

From the proof of Theorem 4.1 in Section C.2, we see that the use of peeling argument will create an additional term $\log_2(1/h)$ in the upper bound, although it is further bounded by $\log(\log n)$ (under the prescribed constraint on h), a very slowly growing function of n. In this section, we argue that this extra term is an artifact of the proof technique, and can be avoided through a more careful analysis regarding the (local) restricted strong convexity of the empirical loss $\widehat{Q}_h(\cdot)$.

The key is to refine the convergence rate of $\widehat{\boldsymbol{\beta}}_h$ from $r_0 \asymp h$ to the claimed one, conditioned on $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \in \Theta(r_0)$. The first-order optimality condition implies $\nabla \widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) = 0$, and hence

$$\begin{aligned}
&\langle \nabla \widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\delta}}\rangle \\
&= \langle -\nabla \widehat{Q}_h(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\delta}}\rangle \le \Big(\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 + \underbrace{\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_h(\boldsymbol{\beta}^*)\|_2}_{\le 0.5 l_0 \kappa_2 \cdot h^2}\Big)\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}.
\end{aligned} \tag{C.19}$$

Define the symmetrized Bregman divergence associated with the convex function $\widehat{Q}_h(\cdot)$:

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \langle \nabla \widehat{Q}_h(\boldsymbol{\beta}_1) - \nabla \widehat{Q}_h(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \geq 0, \quad \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p. \tag{C.20}$$

Then the left-hand side of (C.19) reads $D(\boldsymbol{\beta}^* + \widehat{\boldsymbol{\delta}}, \boldsymbol{\beta}^*)$. Starting from (C.19), we need to bound $\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2$ from above, and derive a lower bound for $D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*)$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^p$ in a local neighborhood of the origin. The following two lemmas serve for this purpose.

Lemma C.2. Assume Conditions 4.1–4.4 hold. For any $t \geq 0$,

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 \leq 1.46 \upsilon_0 \left(C_\tau \sqrt{\frac{4p + 2t}{n}} + \bar{\tau} \frac{4p + 2t}{n} \right) \tag{C.21}$$

with probability at least $1 - e^{-t}$, where $C_\tau^2 = \tau(1 - \tau) + (1 + \tau) l_0 \kappa_2 h^2$.

In addition to Condition 4.2, assume $f_{\varepsilon|\boldsymbol{x}}(0) \leq \bar{f}$ for some $\bar{f} \geq \underline{f} > 0$. Then, for all $0 < h \leq \underline{f}/(4 l_0)$,

$$\frac{7}{8} \underline{f} \leq \inf_{|u| \leq h/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \sup_{|u| \leq h/2} f_{\varepsilon|\boldsymbol{x}}(u) \leq \frac{9}{8} \bar{f}, \tag{C.22}$$

almost surely (over $\boldsymbol{x}$). Moreover, for $\delta \in (0, 1]$, define $\iota_\delta \geq 0$ as

$$\iota_\delta = \inf\{\iota > 0 : \mathbb{E}\{\langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 \mathbb{1}(|\langle \boldsymbol{u}, \boldsymbol{w} \rangle| > \iota)\} \leq \delta \ \text{ for all } \boldsymbol{u} \in \mathbb{S}^{p-1}\}, \tag{C.23}$$

where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x}$ is the standardized covariate vector satisfying $\mathbb{E}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) = \mathbf{I}_p$, and hence $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 = 1$ for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$. It can be shown that ι_δ depends only on δ and υ_0 in Condition 4.4, and the map $\delta \mapsto \iota_\delta$ is non-increasing with $\iota_\delta \downarrow 0$ as $\delta \uparrow 1$ and $\iota_1 = 0$. By Markov's inequality, for any $\iota > 0$ it holds $\sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}\{\langle \boldsymbol{u}, \boldsymbol{w} \rangle^2 \mathbb{1}(|\langle \boldsymbol{u}, \boldsymbol{w} \rangle| > \iota)\} \leq \iota^{-2} m_4$. Hence, a rather crude upper bound for ι_δ is $\iota_\delta \leq (m_4/\delta)^{1/2}$.

Lemma C.3. Assume the kernel $K(\cdot)$ is such that $\kappa_l = \min_{|u| \leq 1} K(u) > 0$, and let the bandwidth satisfy $0 < h \leq \underline{f}/(4 l_0)$. Given any $0 < r \leq h/(4 \iota_{1/4})$ with $\iota_{1/4}$ defined in (C.23),

$$\inf_{\boldsymbol{\delta} \in \Theta(r)} \frac{D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*)}{\kappa_l \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2} \geq c_0 \underline{f} - c_1 \left\{ \bar{f}^{1/2} \frac{1}{r} \sqrt{\frac{ph}{n}} + (\bar{f} m_4)^{1/2} \sqrt{\frac{t}{nh}} + \frac{ht}{r^2 n} \right\} \tag{C.24}$$

with probability at least $1 - e^{-t}$ for any $t \geq 0$, where $c_1 > 0$ is an absolute constant, and $c_0 = 21/32$.

Let $\mathcal{G}(t)$ be the "good" event that the bounds (C.21) and (C.24) are satisfied. Together, Lemma C.2,

Lemma C.3 and (C.19) imply that, conditioned on $\{\widehat{\boldsymbol{\delta}} \in \Theta(r_0)\} \cap \mathcal{G}(t)$ with $r_0 \le h/(4\iota_{1/4})$,

$$c_0\kappa_l \underline{f} \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}^2 \le \left\{1.46\upsilon_0\left(C_\tau\sqrt{\frac{4p+2t}{n}} + \bar{\tau}\frac{4p+2t}{n}\right) + 0.5 l_0\kappa_2 h^2\right\}\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}$$
$$+ c_1\kappa_l\left\{\bar{f}^{1/2}\frac{1}{r_0}\sqrt{\frac{ph}{n}} + (\bar{f}m_4)^{1/2}\sqrt{\frac{t}{nh}} + \frac{ht}{r_0^2 n}\right\} r_0 \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}}.$$

Consequently, we obtain the bound

$$c_0\underline{f} \cdot \|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le 1.46\kappa_l^{-1}\upsilon_0\left(C_\tau\sqrt{\frac{4p+2t}{n}} + \bar{\tau}\frac{4p+2t}{n}\right) + 0.5\kappa_l^{-1} l_0\kappa_2 h^2$$
$$+ c_1\left\{\bar{f}^{1/2}\sqrt{\frac{ph}{n}} + (\bar{f}m_4)^{1/2} r_0\sqrt{\frac{t}{nh}} + \frac{ht}{r_0 n}\right\}$$

without having the additional $\log_2(1/h)$ term. For example, we may take $r_0 = h/(8m_4^{1/2})$. □

C.3.1 Proof of Lemma C.2

Define $\xi_i = \mathcal{K}(-\varepsilon_i/h) - \tau$ for $i = 1, \ldots, n$, so that $\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\} = (1/n)\sum_{i=1}^n\{\xi_i\boldsymbol{w}_i - \mathbb{E}(\xi_i\boldsymbol{w}_i)\} \in \mathbb{R}^p$, where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. Using a covering argument, for any $\epsilon \in (0,1)$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of the unit sphere with cardinality $|\mathcal{N}_\epsilon| \le (1+2/\epsilon)^p$ such that

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 \le (1-\epsilon)^{-1}\max_{\boldsymbol{u}\in\mathcal{N}_\epsilon}\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\rangle.$$

For each unit vector $\boldsymbol{u} \in \mathcal{N}_\epsilon$, define centered random variables $\gamma_{\boldsymbol{u},i} = \langle \boldsymbol{u}, \xi_i\boldsymbol{w}_i - \mathbb{E}(\xi_i\boldsymbol{w}_i)\rangle$. It can be shown that $|\xi_i| \le \bar{\tau} := \max(1-\tau, \tau)$ and $\mathbb{E}(\xi_i^2|\boldsymbol{x}_i) \le C_\tau = \tau(1-\tau) + (1+\tau)l_0\kappa_2 h^2$ (see the proof of Theorem 4.3 below). Hence, for $k = 2, 3, \ldots$,

$$\begin{aligned}
\mathbb{E}(|\langle \boldsymbol{u}, \xi_i\boldsymbol{w}_i\rangle|^k) &\le \bar{\tau}^{k-2}\,\mathbb{E}\{|\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle|^k \cdot \mathbb{E}(\xi_i^2|\boldsymbol{x}_i)\} \\
&\le C_\tau^2\bar{\tau}^{k-2}\upsilon_0^k\int_0^\infty \mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle| \ge \upsilon_0 t)kt^{k-1}\,\mathrm{d}t \\
&\le C_\tau^2\bar{\tau}^{k-2}\upsilon_0^k k\int_0^\infty t^{k-1}e^{-t}\,\mathrm{d}t \\
&= k! \cdot C_\tau^2\bar{\tau}^{k-2} \cdot \upsilon_0^k \\
&\le \frac{k!}{2} \cdot (C_\tau\upsilon_0)^2 \cdot (2\bar{\tau}\upsilon_0)^{k-2}.
\end{aligned}$$

Consequently, it follows from Bernstein's inequality that for every $u \geq 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{u},i} \leq \upsilon_0\left(C_\tau\sqrt{\frac{2u}{n}} + \frac{2\bar{\tau}u}{n}\right)$$

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

Finally, applying a union bound over $\boldsymbol{u} \in \mathcal{N}_\epsilon$ yields

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}^*) - \nabla Q_h(\boldsymbol{\beta}^*)\}\|_2 \leq \frac{\upsilon_0}{1-\epsilon}\left(C_\tau\sqrt{\frac{2u}{n}} + \frac{2\bar{\tau}u}{n}\right)$$

with probability at least $1 - e^{\log(1+2/\epsilon)p-u}$. Taking $\epsilon = 2/(e^2-1)$ and $u = 2p + t$ $(t \geq 0)$ proves the claimed result. □

C.3.2 Proof of Lemma C.3

Recall that the empirical loss $\widehat{Q}_h(\cdot)$ in (2.6) is convex and twice continuously differentiable with $\nabla\widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n\{\mathcal{K}_h(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle - y_i) - \tau\}\boldsymbol{x}_i$ and $\nabla^2\widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n K_h(\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle - y_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. For the symmetrized Bregman divergence $D : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ defined in (C.20), we have

$$D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{K}\left(\frac{\langle\boldsymbol{w}_i,\boldsymbol{v}\rangle - \varepsilon_i}{h}\right) - \mathcal{K}\left(\frac{-\varepsilon_i}{h}\right)\right\}\langle\boldsymbol{w}_i,\boldsymbol{v}\rangle, \tag{C.25}$$

where $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ and $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$. Define the events $\mathcal{E}_i = \{|\varepsilon_i| \leq h/2\} \cap \{|\langle\boldsymbol{w}_i,\boldsymbol{v}\rangle| \leq \|\boldsymbol{v}\|_2 \cdot h/(2r)\}$ for $i = 1,\ldots,n$. For any $\boldsymbol{v} \in \mathbb{B}^p(r)$, note that $|\varepsilon_i - \langle\boldsymbol{w}_i,\boldsymbol{v}\rangle| \leq h$ on $\mathcal{E}_i$, implying

$$D(\boldsymbol{\beta}^* + \boldsymbol{\delta}, \boldsymbol{\beta}^*) \geq \frac{\kappa_l}{nh}\sum_{i=1}^{n}\langle\boldsymbol{w}_i,\boldsymbol{v}\rangle^2\mathbb{1}_{\mathcal{E}_i}, \tag{C.26}$$

where $\mathbb{1}_{\mathcal{E}_i}$ is the indicator function of $\mathcal{E}_i$ and $\kappa_l = \min_{|u|\leq 1} K(u)$. It then suffices to bound the right-hand side of the above inequality from below uniformly over $\boldsymbol{v} \in \mathbb{B}^p(r)$.

For $R > 0$, define the function $\varphi_R(u) = u^2\mathbb{1}(|u| \leq R/2) + \{u\,\mathrm{sign}(u) - R\}^2\mathbb{1}(R/2 < |u| \leq R)$, which is R-Lipschitz continuous and satisfies

$$u^2\mathbb{1}(|u| \leq R/2) \leq \varphi_R(u) \leq u^2\mathbb{1}(|u| \leq R). \tag{C.27}$$

Moreover, note that $\varphi_{cR}(cu) = c^2\varphi_R(u)$ for any $c > 0$ and $\varphi_0(u) = 0$. Hence,

$$\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle^2 \mathbb{1}_{\mathcal{E}_i} \geq \varphi_{\|\boldsymbol{v}\|_2 h/(2r)}(\|\boldsymbol{v}\|_2\langle \boldsymbol{w}_i, \boldsymbol{v}/\|\boldsymbol{v}\|_2\rangle) \cdot \omega_i = \|\boldsymbol{v}\|_2^2 \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}/\|\boldsymbol{v}\|_2\rangle) \cdot \omega_i,$$

where $\omega_i := \mathbb{1}(|\varepsilon_i| \leq h/2)$. By a change of variable, the problem is reduced to bounding

$$D_0(\boldsymbol{v}) := \frac{1}{nh}\sum_{i=1}^n \omega_i \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \tag{C.28}$$

from below uniformly over $\boldsymbol{v} \in \mathbb{S}^{p-1}$.

In the following, we bound the expectation $\mathbb{E}\{D_0(\boldsymbol{v})\}$ and the random fluctuation $D_0(\boldsymbol{v}) - \mathbb{E}\{D_0(\boldsymbol{v})\}$, separately, starting with the former. By (C.22),

$$\frac{7}{8}\underline{f}h \leq \mathbb{E}(\omega_i|\boldsymbol{x}_i) = \int_{-h/2}^{h/2} f_{\varepsilon_i|\boldsymbol{x}_i}(u)\,\mathrm{d}u \leq \frac{9}{8}\bar{f}h. \tag{C.29}$$

Moreover, define $\xi_{\boldsymbol{v}} = \langle \boldsymbol{w}, \boldsymbol{v}\rangle$ such that $\mathbb{E}(\xi_{\boldsymbol{v}}^2) = 1$. By (C.27) and (C.29),

$$\mathbb{E}\{\omega_i \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle)\} \geq \frac{7}{8}\underline{f}h \cdot \mathbb{E}\varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \geq \frac{7}{8}\underline{f}h \cdot \left[1 - \mathbb{E}\xi_{\boldsymbol{v}}^2\mathbb{1}\{|\xi_{\boldsymbol{v}}| > h/(4r)\}\right].$$

With $r \leq h/(4\iota_{1/4})$ and $\iota_{1/4}$ defined in (C.23), it follows that

$$\inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\{D_0(\boldsymbol{v})\} \geq \frac{7}{8}\underline{f} \cdot \left\{1 - \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2\mathbb{1}(|\langle \boldsymbol{w}, \boldsymbol{u}\rangle| \geq \iota_{1/4})\right\} \geq \frac{21}{32}\underline{f}. \tag{C.30}$$

Turning to the random fluctuation, we will use Theorem 7.3 in Bousquet (2003) (a refined Talagrand's inequality) to bound

$$\Delta = \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \{D_0^-(\boldsymbol{v}) - \mathbb{E}D_0^-(\boldsymbol{v})\}, \tag{C.31}$$

where $D_0^-(\boldsymbol{v}) := -D_0(\boldsymbol{v})$. Note that $0 \leq \varphi_R(u) \leq (R/2)^2$ for all $u \in \mathbb{R}$ and $\omega_i \in \{0, 1\}$. Hence, $\chi_i := (\omega_i/h) \cdot \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) \geq 0$ is bounded by $h/(4r)^2$. Moreover, it follows from (C.29) that $\mathbb{E}(\chi_i^2) \leq 9\bar{f}m_4/(8h)$. We then apply Theorem 7.3 in Bousquet (2003) and obtain that, for any $t > 0$,

$$\Delta \leq \mathbb{E}(\Delta) + \{\mathbb{E}(\Delta)\}^{1/2}\frac{1}{2r}\sqrt{\frac{ht}{n}} + \frac{3}{2}(\bar{f}m_4)^{1/2}\sqrt{\frac{t}{nh}} + \frac{h}{(4r)^2}\frac{t}{3n} \tag{C.32}$$

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

It remains to bound the expected value $\mathbb{E}(\Delta)$. Recall that $\omega_i = \mathbb{1}(|\varepsilon_i| \le h/2) \in \{0, 1\}$, and hence $\omega_i \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \omega_i^2 \varphi_{h/(2r)}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \varphi_{\omega_i h/(2r)}(\langle \omega_i \boldsymbol{w}_i, \boldsymbol{v}\rangle) = \varphi_{h/(2r)}(\langle \omega_i \boldsymbol{w}_i, \boldsymbol{v}\rangle)$. Define

$$\bar{\boldsymbol{w}}_i = \omega_i \boldsymbol{w}_i \quad \text{and} \quad \mathcal{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) = \varphi_{h/(2r)}(\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle), \quad \boldsymbol{v} \in \mathbb{S}^{p-1}.$$

By Rademacher symmetrization,

$$\mathbb{E}(\Delta) \le 2\mathbb{E}\left\{ \sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \frac{1}{nh} \sum_{i=1}^n e_i \cdot \mathcal{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) \right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since $\varphi_R(\cdot)$ is R-Lipschitz, $\mathcal{E}(\boldsymbol{v}; \boldsymbol{z}_i)$ is an $(h/2r)$-Lipschitz function in $\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle$, i.e., for any $\bar{\boldsymbol{w}}_i$ and parameters $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{S}^{p-1}$,

$$\left|\mathcal{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) - \mathcal{E}(\boldsymbol{v}'; \bar{\boldsymbol{w}}_i)\right| \le \frac{h}{2r} \left|\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle - \langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}'\rangle\right|. \tag{C.33}$$

Moreover, observe that $\mathcal{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) = 0$ for any $\boldsymbol{v}$ such that $\langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle = 0$. With the above preparations, we are ready to use Talagrand's contraction principle to bound $\mathbb{E}(\Delta)$. Define the subset $T \subseteq \mathbb{R}^n$ as

$$T = \{\boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}} : t_i = \langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle, i = 1, \ldots, n, \ \boldsymbol{v} \in \mathbb{S}^{p-1}\},$$

and contractions $\phi_i : \mathbb{R} \to \mathbb{R}$ as $\phi_i(t) = (2r/h) \cdot \varphi_{h/(2r)}(t)$. By (C.33), $|\phi(t) - \phi(s)| \le |t - s|$ for all $t, s \in \mathbb{R}$. Applying Talagrand's contraction principle (see, e.g., Theorem 4.12 and (4.20) in Ledoux and Talagrand (1991)), we have

$$\begin{aligned}
\mathbb{E}(\Delta) &\le 2\mathbb{E}\left\{ \sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \frac{1}{nh} \sum_{i=1}^n e_i \cdot \mathcal{E}(\boldsymbol{v}; \bar{\boldsymbol{w}}_i) \right\} = \mathbb{E}\left\{ \sup_{\boldsymbol{t} \in T} \frac{1}{nr} \sum_{i=1}^n e_i \phi_i(t_i) \right\} \le \mathbb{E}\left(\sup_{\boldsymbol{t} \in T} \frac{1}{nr} \sum_{i=1}^n e_i t_i \right) \\
&= \mathbb{E}\left(\sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \frac{1}{nr} \sum_{i=1}^n e_i \langle \bar{\boldsymbol{w}}_i, \boldsymbol{v}\rangle \right) \le \mathbb{E}\left\| \frac{1}{nr} \sum_{i=1}^n e_i \bar{\boldsymbol{w}}_i \right\|_2 \le \bar{f}^{1/2} \frac{3}{2r} \sqrt{\frac{ph}{2n}}.
\end{aligned}$$

This, combined with (C.31) and (C.32), yields that

$$\Delta \le \frac{3}{2} \bar{f}^{1/2} \left(1.25 \frac{1}{r} \sqrt{\frac{ph}{2n}} + m_4^{1/2} \sqrt{\frac{t}{nh}} \right) + (1 + 1/48) \frac{ht}{r^2 n}$$

with probability at least $1 - e^{-t}$. Combining this with (C.25), (C.28) and (C.30) proves (C.24). □

C.4 Proof of Theorem 4.2

We keep the notation used in the proof of Theorem 4.1, and for any $t \geq 0$, let $r = r(n, p, t) \asymp \sqrt{(p+t)/n} + h^2 > 0$ be such that $\mathbb{P}\{\widehat{\boldsymbol{\beta}}_h \in \boldsymbol{\beta}^* + \Theta(r)\} \geq 1 - 2e^{-t}$, provided $\sqrt{(p+t)/n} \lesssim h \lesssim 1$. Define the vector-valued random process

$$\Delta(\boldsymbol{\delta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_h(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^*) - \mathbf{J}_h \boldsymbol{\delta}\}, \quad \boldsymbol{\delta} \in \mathbb{R}^p, \tag{C.34}$$

where $\mathbf{J}_h = \nabla^2 Q_h(\boldsymbol{\beta}^*)$ is the population Hessian at $\boldsymbol{\beta}^*$. Since $\widehat{\boldsymbol{\beta}}_h$ falls in a local neighborhood of $\boldsymbol{\beta}^*$ with high probability, it suffices to bound the local fluctuation $\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2$. By the triangle inequality,

$$\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2 \leq \sup_{\boldsymbol{\delta} \in \Theta(r)} \|\mathbb{E}\Delta(\boldsymbol{\delta})\|_2 + \sup_{\boldsymbol{\delta} \in \Theta(r)} \|\Delta(\boldsymbol{\delta}) - \mathbb{E}\Delta(\boldsymbol{\delta})\|_2 := I_1 + I_2. \tag{C.35}$$

We now provide upper bounds for I_1 and I_2, respectively.

Upper bound for I_1: By the mean value theorem for vector-valued functions,

$$\begin{aligned}
\mathbb{E}\Delta(\boldsymbol{\delta}) &= \boldsymbol{\Sigma}^{-1/2}\left\langle \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\mathrm{d}t, \boldsymbol{\delta} \right\rangle - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h\boldsymbol{\delta} \\
&= \left\langle \boldsymbol{\Sigma}^{-1/2} \int_0^1 \nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\mathrm{d}t\, \boldsymbol{\Sigma}^{-1/2} - \mathbf{H}_h, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \right\rangle,
\end{aligned}$$

where $\mathbf{H}_h := \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{K_h(\varepsilon)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$. By law of iterative expectation and a change of variable,

$$\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{K_h(t\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle - \varepsilon)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\} = \mathbb{E}\left\{\int_{-\infty}^{\infty} K(u) f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{x}, \boldsymbol{\delta}\rangle - hu)\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\}.$$

Write $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ for $\boldsymbol{\delta} \in \Theta(r)$, so that $\|\boldsymbol{v}\|_2 \leq r$ and

$$\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\left\{\int_{-\infty}^{\infty} K(u) f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{w}, \boldsymbol{v}\rangle - hu)\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\}.$$

By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$,

$$\begin{aligned}
&\left\|\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_h(\boldsymbol{\beta}^* + t\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}_h\right\|_2 \\
&= \left\|\mathbb{E}\int K(u)\{f_{\varepsilon|\boldsymbol{x}}(t\langle \boldsymbol{w}, \boldsymbol{v}\rangle - hu) - f_{\varepsilon|\boldsymbol{x}}(-hu)\}\,\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2 \\
&\le l_0 t \sup_{\|\boldsymbol{u}\|_2=1} \mathbb{E}(\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2 |\langle \boldsymbol{w}, \boldsymbol{v}\rangle|) \le l_0 m_3 r t,
\end{aligned}$$

where the third inequality holds by the Cauchy-Schwarz inequality. Consequently,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\mathbb{E}\Delta(\boldsymbol{\delta})\|_2 \le 0.5 l_0 m_3 r^2. \tag{C.36}$$

Upper bound for I_2: Next, we provide an upper bound for the supremum of the zero-mean stochastic process $\Delta(\boldsymbol{\delta}) - \mathbb{E}\Delta(\boldsymbol{\delta})$ under ℓ_2-norm. Define the centered gradient process $G(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h(\boldsymbol{\beta}) - \nabla Q_h(\boldsymbol{\beta})\}$, so that $\Delta(\boldsymbol{\delta}) - \mathbb{E}\Delta(\boldsymbol{\delta}) = G(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G(\boldsymbol{\beta}^*)$. Again, by a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$,

$$\begin{aligned}
\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\beta}) - \mathbb{E}\Delta(\boldsymbol{\beta})\|_2 &\le \sup_{\boldsymbol{\delta}\in\Theta(r)} \|G(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G(\boldsymbol{\beta}^*)\|_2 \\
&= \sup_{\|\boldsymbol{v}\|_2\le r} \|\underbrace{G(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - G(\boldsymbol{\beta}^*)}_{=:\Delta_0(\boldsymbol{v})}\|_2.
\end{aligned}$$

We will employ Theorem A.3 in Spokoiny (2013) to bound the supremum $\sup_{\|\boldsymbol{v}\|_2\le r} \|\Delta_0(\boldsymbol{v})\|_2$, where $\Delta_0(\cdot)$ defined above satisfies $\Delta_0(\mathbf{0}) = \mathbf{0}$ and $\mathbb{E}\{\Delta_0(\boldsymbol{v})\} = \mathbf{0}$. Taking the gradient with respect to $\boldsymbol{v}$ yields

$$\nabla\Delta_0(\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^n \{K_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}} - \mathbb{E}(K_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}})\},$$

where $K_{i,\boldsymbol{v}} := K_h(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - \varepsilon_i)$ satisfies $0 \le K_{i,\boldsymbol{v}} \le \kappa_u h^{-1}$. For any $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{S}^{p-1}$ and $\lambda \in \mathbb{R}$, using the elementary inequality $|e^u - 1 - u| \le u^2 e^{|u|}/2$, we obtain

$$\begin{aligned}
&\mathbb{E}\exp\{\lambda n^{1/2}\langle \boldsymbol{u}, \nabla\Delta_0(\boldsymbol{v})\boldsymbol{u}'\rangle/\upsilon_1^2\} \\
&\le \prod_{i=1}^n \left\{1 + \frac{\lambda^2}{2\upsilon_1^4 n} e^{\frac{\bar{f}|\lambda|}{\upsilon_1^2\sqrt{n}}\mathbb{E}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|} \mathbb{E}\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)\}^2 e^{\frac{\kappa_u|\lambda|}{h\sqrt{n}}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/\upsilon_1^2}\right\} \\
&\le \prod_{i=1}^n \left\{1 + \frac{\lambda^2}{2\upsilon_1^4 n} e^{\frac{\bar{f}|\lambda|}{\sqrt{n}}} \mathbb{E}\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)\}^2 e^{\frac{\kappa_u|\lambda|}{h\sqrt{n}}|\langle \boldsymbol{w}_i,\boldsymbol{u}\rangle\langle \boldsymbol{w}_i,\boldsymbol{u}'\rangle|/\upsilon_1^2}\right\},
\end{aligned} \tag{C.37}$$

where we used the bound $\mathbb{E}|\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle| \le (\mathbb{E}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle^2)^{1/2}(\mathbb{E}\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle^2)^{1/2} = 1$ in the second inequality. Moreover, the first and second conditional moments of $K_{i,\boldsymbol{v}}$ can be rewritten as follows:

$$\mathbb{E}(K_{i,\boldsymbol{v}}|\boldsymbol{x}_i) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} K(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - hu)\,\mathrm{d}u;$$
$$\mathbb{E}(K_{i,\boldsymbol{v}}^2|\boldsymbol{x}_i) = \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \frac{1}{h}\int_{-\infty}^{\infty} K^2(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - hu)\,\mathrm{d}u,$$

from which it follows that $|\mathbb{E}(K_{i,\boldsymbol{v}}|\boldsymbol{x}_i)| \le \bar{f}$ and $\mathbb{E}(K_{i,\boldsymbol{v}}^2|\boldsymbol{x}_i) \le \kappa_u \bar{f} h^{-1}$ almost surely.

By the Cauchy-Schwarz inequality and the inequality $ab \le a^2/2 + b^2/2$, $a, b \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}(\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)^2 e^{t|\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle|} \\
&\le \mathbb{E}(\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)^2 e^{\frac{t}{2}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle^2 + \frac{t}{2}\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle^2} \\
&\le \left(\mathbb{E}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle^4 e^{t\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle^2}\right)^{1/2}\left(\mathbb{E}\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle^4 e^{t\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle^2}\right)^{1/2}, \quad \text{valid for any } t > 0.
\end{aligned}$$

Given a unit vector $\boldsymbol{u}$, let $\chi = \langle \boldsymbol{w}, \boldsymbol{u}\rangle^2/(2\upsilon_1)^2$ so that under Condition 4.5, $\mathbb{P}(\chi \ge u) \le 2e^{-2u}$ for any $u \ge 0$. It follows that $\mathbb{E}(e^{\chi}) = 1 + \int_0^{\infty} e^u \mathbb{P}(\chi \ge u)\mathrm{d}u \le 1 + 2\int_0^{\infty} e^{-u}\mathrm{d}u = 3$, and

$$\mathbb{E}(\chi^2 e^{\chi}) = \int_0^{\infty} (u^2 + 2u)e^u \mathbb{P}(\chi \ge u)\mathrm{d}u \le 2\int_0^{\infty} (u^2 + 2u)e^{-u}\mathrm{d}u = 8.$$

Taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{p-1}$, we have

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}e^{\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2/(2\upsilon_1)^2} \le 3 \quad \text{and} \quad \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{u}\rangle^4 e^{\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2/(2\upsilon_1)^2} \le 8(2\upsilon_1)^4.$$

Substituting the above bounds into (C.37) yields that, for any $|\lambda| \le \min\{h/(4\kappa_u), 1/\bar{f}\}n^{1/2}$,

$$\begin{aligned}
&\mathbb{E}\exp\{\lambda n^{1/2}\langle \boldsymbol{u}, \nabla\Delta_0(\boldsymbol{v})\boldsymbol{u}'\rangle/\upsilon_1^2\} \\
&\le \prod_{i=1}^{n}\left[1 + \frac{e\lambda^2}{2\upsilon_1^4 n}\mathbb{E}\{K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle - \mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)\}^2 e^{|\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle|/(4\upsilon_1^2)}\right] \\
&\le \prod_{i=1}^{n}\left[1 + \frac{e\lambda^2}{\upsilon_1^4 n}\mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)^2 e^{|\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle|/(4\upsilon_1^2)}\right. \\
&\qquad\qquad \left. + \frac{e\lambda^2}{\upsilon_1^4 n}\{\mathbb{E}(K_{i,\boldsymbol{v}}\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle)\}^2 \mathbb{E}e^{|\langle \boldsymbol{w}_i, \boldsymbol{u}\rangle\langle \boldsymbol{w}_i, \boldsymbol{u}'\rangle|/(4\upsilon_1^2)}\right] \\
&\le \prod_{i=1}^{n}\left(1 + C_0^2\frac{\lambda^2}{2nh}\right) \le \exp\{C_0^2\lambda^2/(2h)\},
\end{aligned}$$

where $C_0 > 0$ depends only on $(\kappa_u, \bar{f})$. We have thus verified condition (A.4) in Spokoiny (2013) with $\mathrm{g} = \min\{h/(4\kappa_u), 1/\bar{f}\}(n/2)^{1/2}$ and $\nu_0 = C_0 h^{-1/2}$. Applying Theorem A.3 therein to the process $\{\Delta_0(\boldsymbol{r})/\upsilon_1^2, \boldsymbol{v} \in \mathbb{B}^p(r)\}$, we obtain that with probability at least $1 - e^{-t}$,

$$\sup_{\|\boldsymbol{v}\|_2 \le r} \|\Delta_0(\boldsymbol{v})\|_2 \le 6C_0\upsilon_1^2 r\sqrt{\frac{4p+2t}{nh}}$$

as long as $h \ge 8\kappa_u\sqrt{(2p+t)/n}$ and $n \ge 4\bar{f}^2(2p+t)$.

Joint with (C.35) and (C.36), this implies that with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2 \le 6C_0\upsilon_1^2 r\sqrt{\frac{4p+2t}{nh}} + 0.5 l_0 m_3 r^2. \tag{C.38}$$

Recall from the beginning of the proof that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* \in \Theta(r)$ with probability at least $1 - 2e^{-t}$ with $r = r(n,p,t) \asymp \sqrt{(p+t)/n} + h^2$. Combined with (C.38), we conclude that with probability at least $1 - 3e^{-t}$, $\|\Delta(\widehat{\boldsymbol{\delta}})\|_2 \lesssim (p+t)/(h^{1/2}n) + h^{3/2}\sqrt{(p+t)/n} + h^4$, as claimed. □

C.5 Proof of Theorem 4.3

Let $\boldsymbol{a} \in \mathbb{R}^p$ be an arbitrary vector defining a linear functional of interest. Given $h = h_n > 0$, define $S_n = n^{-1/2}\sum_{i=1}^n \gamma_i\xi_i$ and its centered version $S_n^0 = S_n - \mathbb{E}(S_n)$, where $\xi_i = \tau - \mathcal{K}(-\varepsilon_i/h)$ and $\gamma_i = \langle \mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}_i\rangle$. By the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ and the fundamental theorem of calculus, it can be shown that $|\mathbb{E}(\xi_i|\boldsymbol{x}_i)| \le 0.5 l_0\kappa_2 h^2$, from which it follows by the law of iterated expectation that $|\mathbb{E}(\gamma_i\xi_i)| \le 0.5 l_0\kappa_2\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot h^2$.

Let $\eta_n = (p + \log n)/n$. Then, applying (C.34) and (C.38) with $t = \log n$ and the triangle inequality, we obtain that under the constraint $\eta_n^{1/2} \lesssim h \lesssim 1$,

$$\begin{aligned}
&\left|n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle - S_n^0\right| \\
&= n^{1/2}\left|\left\langle \boldsymbol{\Sigma}^{1/2}\mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) - \boldsymbol{\Sigma}^{-1/2}\frac{1}{n}\sum_{i=1}^n\{\tau - \mathcal{K}(-\varepsilon_i/h)\}\boldsymbol{x}_i\right\rangle\right| + |\mathbb{E}(S_n)| \\
&\le c_1\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}(h^{-1/2}\delta_n + h^2)
\end{aligned} \tag{C.39}$$

with probability at least $1 - 3n^{-1}$ for some constant $c_1 > 0$.

For the centered partial sum $S_n^0 = S_n - \mathbb{E}(S_n) = n^{-1/2}\sum_{i=1}^n(1 - \mathbb{E})\gamma_i\xi_i$, we have $\mathrm{var}(S_n^0) = \mathrm{var}(S_n) = \mathbb{E}(\gamma\xi)^2 - \{\mathbb{E}(\gamma\xi)\}^2$, where $\gamma = \langle \mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}\rangle$ and $\xi = \tau - \mathcal{K}(-\varepsilon/h)$. By the Berry-Esseen

inequality (see, e.g., Tyurin (2011)),

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\{S_n^0 \le \mathrm{var}(S_n)^{1/2}x\} - \Phi(x)\right| \le \frac{\mathbb{E}\{|\gamma\xi - \mathbb{E}(\gamma\xi)|^3\}}{2[\mathbb{E}(\gamma\xi)^2 - \{\mathbb{E}(\gamma\xi)\}^2]^{3/2}\sqrt{n}}. \tag{C.40}$$

We have shown that $|\mathbb{E}(\gamma\xi)| \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot h^2$. To bound $\mathbb{E}(\xi^2|\boldsymbol{x})$, by a change of variable and integration by parts,

$$\begin{aligned}
\mathbb{E}\{\mathcal{K}^2(-\varepsilon/h)|\boldsymbol{x}\} &= 2\int_{-\infty}^{\infty} K(v)\mathcal{K}(v)F_{\varepsilon|\boldsymbol{x}}(-vh)\mathrm{d}v \\
&= 2\tau \underbrace{\int_{-\infty}^{\infty} K(v)\mathcal{K}(v)\mathrm{d}v}_{=1/2} - 2hf_{\varepsilon|\boldsymbol{x}}(0) \underbrace{\int_{-\infty}^{\infty} vK(v)\mathcal{K}(v)\mathrm{d}v}_{=\int_0^\infty \mathcal{K}(v)\{1-\mathcal{K}(v)\}\mathrm{d}v>0} \\
&\quad + 2\int_{-\infty}^{\infty}\int_0^{-vh} \{f_{\varepsilon|\boldsymbol{x}}(t) - f_{\varepsilon|\boldsymbol{x}}(0)\}K(v)\mathcal{K}(v)\mathrm{d}t\mathrm{d}v \\
&\le \tau + l_0\kappa_2 h^2,
\end{aligned}$$

where κ_2 and l_0 are the constants from Conditions 4.1 and 4.2. It then follows that

$$\tau(1-\tau) - Ch - (1+\tau)l_0\kappa_2 h^2 \le \mathbb{E}(\xi^2|\boldsymbol{x}_i) \le \tau(1-\tau) + (1+\tau)l_0\kappa_2 h^2,$$

where $C = 2\bar{f}\cdot\int_0^\infty \mathcal{K}(v)\{1-\mathcal{K}(v)\}\mathrm{d}v$. For all sufficiently small h, $\mathrm{var}(S_n) = \{\tau(1-\tau)+O(h)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$. On the other hand, $\mathbb{E}(|\gamma\xi|^3) \le \max(\tau, 1-\tau)\mathbb{E}(\xi^2|\gamma|^3) \le m_3\{\tau(1-\tau) + O(h^2)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^3$. Substituting these bounds into (C.40) yields

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\{S_n^0 \le \mathrm{var}(S_n)^{1/2}x\} - \Phi(x)\right| \le c_2 n^{-1/2} \tag{C.41}$$

for some constant $c_2 > 0$.

Recall that $\sigma_h^2 = \mathbb{E}\{\mathcal{K}_h(-\varepsilon) - \tau\}^2\langle\mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}\rangle^2 = \mathbb{E}(\gamma\xi)^2$, and thus $|\mathrm{var}(S_n) - \sigma_h^2| = (\mathbb{E}\gamma\xi)^2 \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2 \cdot h^4$. By an application of Lemma A.7 in the supplement of Spokoiny and Zhilova (2015), for sufficiently small h, we have

$$\sup_{x\in\mathbb{R}}\left|\Phi(x/\mathrm{var}(S_n)^{1/2}) - \Phi(x/\sigma_h)\right| \le c_3 h^4. \tag{C.42}$$

Before proceeding, we note that the constants c_1–c_3 appeared above are all independent of $\boldsymbol{a}$ and (n, p). Let $G \sim N(0, 1)$. Putting together the above derivations, for any $x \in \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{R}^p$, we

obtain

$$
\begin{aligned}
&\mathbb{P}(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le x) \\
&\le \mathbb{P}\{S_n^0 \le x + c_1\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}(h^{-1/2}\eta_n + h^2)\} + 3n^{-1} \\
&\le \mathbb{P}\{\mathrm{var}(S_n)^{1/2}G \le x + c_1\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}(h^{-1/2}\eta_n + h^2)\} + c_2 n^{-1/2} + 3n^{-1} \\
&\le \mathbb{P}\{\sigma_h G \le x + c_1\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}(h^{-1/2}\eta_n + h^2)\} + c_2 n^{-1/2} + c_3 h^4 + 3n^{-1} \\
&\le \mathbb{P}(\sigma_h G \le x) + c_1(2\pi)^{-1/2}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2}(h^{-1/2}\eta_n + h^2)/\sigma_h + c_2 n^{-1/2} + c_3 h^4 + 3n^{-1},
\end{aligned}
$$

where the first, second, and third inequalities holds by (C.39), (C.41), and (C.42), respectively, and the last inequality follows from the fact that for any $a \le b$ and $\sigma > 0$, $\Phi(b/\sigma) - \Phi(a/\sigma) \le (2\pi)^{-1/2}(b-a)/\sigma$. Recall that $\sigma_h^2 = \{\tau(1-\tau) + O(h)\}\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$, $\|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}/\sigma_h \lesssim \{\tau(1-\tau)\}^{-1/2}$ for all sufficiently small h. A similar argument leads to a series of reverse inequalities. The above bounds are independent of x and $\boldsymbol{a}$, and therefore hold uniformly over all x and $\boldsymbol{a}$.

Putting together the pieces, we conclude that under the bandwidth requirement $\eta_n^{1/2} \lesssim h \lesssim 1$,

$$
\sup_{x\in\mathbb{R},\, \boldsymbol{a}\in\mathbb{R}^p} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le \sigma_h x) - \Phi(x)\right| \lesssim \frac{p + \log n}{(nh)^{1/2}} + n^{1/2}h^2,
$$

as claimed.

Under the additional smoothness condition on $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, using a higher-order Taylor series expansion of $F_{\varepsilon|\boldsymbol{x}}(\cdot)$ gives

$$
\begin{aligned}
&\mathbb{E}\{\mathcal{K}(-\varepsilon_i/h)|\boldsymbol{x}_i\} = \int_{-\infty}^{\infty} K(u)F_{\varepsilon|\boldsymbol{x}}(hu)\,\mathrm{d}u \\
&= \int_{-\infty}^{\infty} K(u)\Bigg[F_{\varepsilon|\boldsymbol{x}}(0) + huf_{\varepsilon|\boldsymbol{x}}(0) + \frac{h^2}{2}u^2 f'_{\varepsilon|\boldsymbol{x}}(0) + \frac{h^3}{3!}u^3 f''_{\varepsilon|\boldsymbol{x}}(0) \\
&\qquad\qquad + \frac{h^3}{2!}u^3\int_0^1 (1-w)^2\{f''_{\varepsilon|\boldsymbol{x}}(huw) - f''_{\varepsilon|\boldsymbol{x}}(0)\}\mathrm{d}w\Bigg]\mathrm{d}u \\
&= \tau + \frac{\kappa_2}{2}f'_{\varepsilon|\boldsymbol{x}}(0)h^2 + O(h^4)\kappa_4 l_2(\boldsymbol{x}),
\end{aligned}
$$

from which it follows that

$$
\left|\mathbb{E}(\gamma_i\xi_i) + \frac{\kappa_2}{2}\langle \mathbf{J}_h^{-1}\boldsymbol{a}, \mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}_i\}\rangle \cdot h^2\right| \lesssim \kappa_4\mathbb{E}|\langle \mathbf{J}_h^{-1}\boldsymbol{a}, \boldsymbol{x}_i\rangle l_2(\boldsymbol{x})| \cdot h^4 \lesssim \kappa_4\|\mathbf{J}_n^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}\{\mathbb{E}l_2^2(\boldsymbol{x})\}^{1/2}h^4.
$$

Together, the above bound and (C.38) with $t = \log n$ imply

$$\left| n^{1/2} \left\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* + \frac{\kappa_2}{2} \mathbf{J}_h^{-1} \mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}_i\} \right\rangle - S_n^0 \right| \lesssim \|\mathbf{J}_h^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}} \cdot n^{1/2} \left(\frac{p + \log n}{nh^{1/2}} + h^{3/2} \sqrt{\frac{p + \log n}{n}} + h^4 \right) \tag{C.43}$$

with probability at least $1 - 3n^{-1}$. Repeating the above analysis, with (C.39) replaced by (C.43), proves the refined Berry-Esseen bound (4.9). □

C.6 Proof of Theorem 4.4

Keep the notation used in the proof of Theorem 4.1. With non-negative multipliers w_i's, the weighted loss $\widehat{Q}_h^\flat(\cdot)$ given in (2.9) is also convex. For $\boldsymbol{\delta} \in \mathbb{R}^p$, define the bootstrap counterpart of $\widehat{D}_h(\cdot)$ as $\widehat{D}_h^\flat(\boldsymbol{\delta}) = \widehat{Q}_h^\flat(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \widehat{Q}_h^\flat(\boldsymbol{\beta}^*)$. It is easy to see that $\mathbb{E}^*\{\widehat{D}_h^\flat(\boldsymbol{\delta})\} = \widehat{D}_h(\boldsymbol{\delta})$. Similarly to (C.10), we have

$$\begin{aligned}
\widehat{D}_h^\flat(\boldsymbol{\delta}) &= \widehat{D}_h(\boldsymbol{\delta}) + \{\widehat{D}_h^\flat(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} \\
&\geq R_h(\boldsymbol{\delta}) - 0.5 l_0 \kappa_2 h^2 \cdot \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} - \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\} \\
&\geq \frac{1}{2}(\underline{f} - l_0\kappa_1 h - 0.5 l_0 m_3 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} - l_0\kappa_2 h^2/\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}})\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}^2 \\
&\qquad - \underbrace{\{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\}}_{\text{sampling error}} - \underbrace{\{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}}_{\text{bootstrap error}}, \quad \boldsymbol{\delta} \in \mathbb{R}^p.
\end{aligned} \tag{C.44}$$

As before, let $r_0 = (2\kappa_2/m_3)^{1/2} h$ be an intermediate convergence radius, and set $r_l = r_0 h$. For any $\boldsymbol{\delta} \in \partial\Theta(r_0)$, it follows that

$$\widehat{D}_h^\flat(\boldsymbol{\delta}) \geq \frac{1}{2}\{\underline{f} - l_0\kappa_1 h - l_0(2\kappa_2/m_3)^{1/2} h\} r_0^2 - \{D_h(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta})\} - \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}. \tag{C.45}$$

As a bootstrap counterpart of Lemma C.4, the following lemma provides high probability bounds for the bootstrap error $\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})$ uniformly over $\boldsymbol{\delta} \in \mathbb{R}^p$ in some compact subset.

Lemma C.4. For each $t \geq 0$, there exists an event $\mathcal{E}_1(t)$ with $\mathbb{P}\{\mathcal{E}_1(t)\} \geq 1 - e^{-t}$ such that conditioned on $\mathcal{E}_1(t)$,

$$\mathbb{P}^* \left\{ \sup_{\boldsymbol{\delta} \in \Theta(r)} \{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\} \geq C\bar{\tau}\upsilon_1 r \sqrt{\frac{u}{n}} \right\} \leq e^{2p-u}$$

for any $u \geq 0$ as long as $n \gtrsim p + t$. Moreover, for any $r_u > r_l > 0$, with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$)

at least $1 - \lceil e \log(\frac{r_u}{r_l}) \rceil e^{2p-u}$ conditioned on $\mathcal{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \le C' \bar{\tau} \upsilon_1 \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \sqrt{\frac{u}{n}}$$

holds for all $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying $r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_u$ as long as $n \gtrsim p + t$. Here both $C, C' > 0$ are absolute constants.

Let $\mathcal{E}_1(t)$ be the event in Lemma C.4 that occurs with probability at least $1 - e^{-t}$. Applying the first inequality with $r = r_0$ and $u = 2p + t$, we obtain that conditioned on $\mathcal{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \le C \bar{\tau} \upsilon_1 r_0 \sqrt{\frac{2p + t}{n}}.$$

Let $\mathcal{E}_2(t)$ be the event that the bounds (C.15) and (C.16) hold. Then, the "good event" $\mathcal{E}(t) := \mathcal{E}_1(t) \cap \mathcal{E}_2(t)$ satisfies $\mathbb{P}\{\mathcal{E}(t)\} \ge 1 - 3e^{-t}$. By (C.45) and the above bound, we find that with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ conditioned on $\mathcal{E}(t)$, $\widehat{D}_h^\flat(\boldsymbol{\delta}) > 0$ for all $\boldsymbol{\delta} \in \partial\Theta(r_0)$ as long as the bandwidth satisfies $\underline{f}^{-1} m_3^{1/2} \upsilon_1 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. Let $\widehat{\boldsymbol{\delta}}^\flat = \widehat{\boldsymbol{\beta}}_h^\flat - \boldsymbol{\beta}^*$. Then, $\widehat{D}_h^\flat(\widehat{\boldsymbol{\delta}}^\flat) \le 0$ by the optimality of $\widehat{\boldsymbol{\beta}}_h^\flat$. This, combined with the convexity of $\widehat{Q}_h^\flat(\cdot)$, ensures that $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_0)$.

Next, consider the decomposition $\Theta(r_0) = \Theta(r_l) \cap \Theta(r_l, r_0)$, where $\Theta(r_l, r_0) = \{\boldsymbol{\delta} \in \mathbb{R}^d : r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_0\}$ and $r_l = r_0 h = (2\kappa_2/m_3)^{1/2} h^2$. If $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_l)$, the claimed bound holds trivially. Throughout the rest, we assume $\widehat{\boldsymbol{\delta}}^\flat \in \Theta(r_l, r_0)$. Taking $u = \log(e \log h^{-1}) + 2p + t$ in the second inequality in Lemma C.4 yields that, with $\mathbb{P}^*$-probability at least $1 - e^{-t}$ conditioned on $\mathcal{E}(t)$,

$$\widehat{D}_h(\widehat{\boldsymbol{\delta}}^\flat) - \widehat{D}_h^\flat(\widehat{\boldsymbol{\delta}}^\flat) \le \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}} \cdot \underbrace{C' \bar{\tau} \upsilon_1 \sqrt{\frac{\log(e \log h^{-1}) + 2p + t}{n}}}_{=: r_1^\flat}.$$

Substituting the above bound and (C.16) into (C.44), we conclude that

$$(\underline{f} - l_0 \kappa_1 h) \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}^2 \le (2r_1 + 2r_1^\flat + l_0 \kappa_2 h^2) \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}} + 0.5 l_0 m_3 \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}^3 \le 2(r_1 + r_1^\flat + l_0 \kappa_2 h^2) \|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}.$$

Canceling out a factor of $\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}}$ from both sides yields the claimed bound. □

C.6.1 Proof of Lemma C.4

We will use a similar argument as in the proof of Lemma C.1. Consider the bootstrap loss difference $\widehat{D}_h^\flat(\boldsymbol{\delta}) - \widehat{D}_h(\boldsymbol{\delta}) = (1/n) \sum_{i=1}^n e_i d_h(\boldsymbol{\delta}; \boldsymbol{z}_i)$, where $d_h(\boldsymbol{\delta}; \boldsymbol{z}_i) = \ell_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta} \rangle) - \ell_h(\varepsilon_i)$ with $\boldsymbol{z}_i = (\boldsymbol{x}_i, \varepsilon_i)$,

and e_i's are independent Rademacher random variables that are independent of $\{\boldsymbol{z}_i\}_{i=1}^n$.

For any $r > 0$ and any $\epsilon \in (0,1)$, applying a conditional version of Chernoff's inequality to $\Delta_\epsilon^\flat(r) := n(1-\epsilon)\sup_{\boldsymbol{\delta}\in\Theta(r)}\{\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta})\}/(\bar{\tau}r)$ yields

$$\mathbb{P}^*\{\Delta_\epsilon^\flat(r) \geq u\} \leq \exp\bigg[-\sup_{\lambda\geq 0}\{\lambda u - \log \mathbb{E}^* e^{\lambda\Delta_\epsilon^\flat(r)}\}\bigg].$$

Again, by the Ledoux-Talagrand contraction principle and discretization via an ϵ-net,

$$\begin{aligned}
\mathbb{E}^* e^{\lambda\Delta_\epsilon^\flat(r)} &\leq \mathbb{E}^* \exp\bigg\{\frac{\lambda}{r}(1-\epsilon)\sup_{\boldsymbol{\delta}\in\Theta(r)}\sum_{i=1}^n e_i\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle\bigg\} \\
&\leq \mathbb{E}^* \exp\bigg\{\lambda(1-\epsilon)\bigg\|\sum_{i=1}^n e_i\boldsymbol{w}_i\bigg\|_2\bigg\} \leq \mathbb{E}^* \exp\bigg\{\lambda \max_{1\leq j\leq N_\epsilon}\sum_{i=1}^n e_i\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{w}_i\bigg\} \\
&\leq \sum_{j=1}^{N_\epsilon}\prod_{i=1}^n \mathbb{E}^* e^{\lambda e_i\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{w}_i} \leq \sum_{j=1}^{N_\epsilon} e^{(\lambda^2/2)\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2},
\end{aligned}$$

where we used the Rademacher moment bound $\mathbb{E}e^{\lambda e_i} \leq e^{\lambda^2/2}$ for any $\lambda \in \mathbb{R}$. Here $\{\boldsymbol{u}_j\}_{j=1}^{N_\epsilon}$ are unit vectors, and $N_\epsilon \leq (1+2/\epsilon)^p$. Consequently,

$$\log \mathbb{E}^* e^{\lambda\Delta_\epsilon^\flat(r)} \leq \log N_\epsilon + \frac{\lambda^2}{2}\max_{1\leq j\leq N_\epsilon}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2,$$

so that for any $u \geq 0$,

$$\begin{aligned}
\sup_{\lambda\geq 0}\{\lambda u - \log \mathbb{E}^* e^{\lambda\Delta_\epsilon^\flat(r)}\} &\geq -\log N_\epsilon + \sup_{\lambda\geq 0}\bigg(\lambda u - \frac{1}{2}\lambda^2\max_{1\leq j\leq N_\epsilon}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2\bigg) \\
&= -\log N_\epsilon + \frac{u^2}{2\max_{1\leq j\leq N_\epsilon}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}.
\end{aligned}$$

Substituting this into the earlier Chernoff's inequality, and by a change of variable, we obtain that for any $u \geq 0$,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)}\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^\flat(\boldsymbol{\delta}) \leq \frac{\bar{\tau}}{1-\epsilon}\sqrt{\max_{1\leq j\leq N_\epsilon}\frac{1}{n}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}\cdot r\sqrt{\frac{2u}{n}}$$

holds with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - e^{p\log(1+2/\epsilon)-u}$.

The above bound holds for any given $r > 0$. Proceed via a peeling argument, we obtain that for

any prespecified $\gamma > 1$ and radii $r_u > r_l > 0$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^{\flat}(\boldsymbol{\delta}) \le \frac{\gamma\bar{\tau}}{1-\epsilon}\sqrt{\max_{1\le j\le N_\epsilon}\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2}\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{2u}{n}} \text{ holds for all } r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_u \quad \text{(C.46)}$$

with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - \lceil \log(\frac{r_u}{r_l})/\log(\gamma)\rceil e^{p\log(1+2/\epsilon)-u}$.

Next, we bound the data-dependent quantity $\max_{1\le j\le N_\epsilon}(1/n)\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2$ with $N_\epsilon \le (1 + 2/\epsilon)^p$ and $\boldsymbol{u}_j \in \mathbb{S}^{p-1}$. Note that $\mathbb{E}\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2 = 1$. By the sub-Gaussianity of $\boldsymbol{w}_i \in \mathbb{R}^p$ ensured by Condition 4.5, for integers $k = 2, 3, \ldots$,

$$\begin{aligned}
\mathbb{E}(\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2)^k &\le \upsilon_1^{2k}\cdot 2k\int_0^\infty \mathbb{P}(|\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle| \ge \upsilon_1 u)u^{2k-1}\,\mathrm{d}u \\
&\le \upsilon_1^{2k}\cdot 4k\int_0^\infty u^{2k-1}e^{-u^2/2}\,\mathrm{d}u \\
&= 2^k\upsilon_1^{2k}\cdot 2k\int_0^\infty v^{k-1}e^{-v}\,\mathrm{d}v = 2^{k+1}k!\cdot\upsilon_1^{2k}.
\end{aligned}$$

In particular, $\mathbb{E}(\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle^4) \le 16\upsilon_1^4$ and $\mathbb{E}(\langle \boldsymbol{u}, \boldsymbol{w}_i\rangle^2)^k \le \frac{k!}{2}\cdot(4\upsilon_1^2)^2\cdot(2\upsilon_1^2)^{k-2}$ for $k \ge 3$. With the above calculations, applying Bernstein's inequality (see, e.g., Theorem 2.10 in Boucheron, Lugosi and Massart (2013)), and taking the union bound over $j = 1, \ldots, N_\epsilon$, we obtain that for any $v \ge 0$,

$$\mathbb{P}\left(\max_{1\le j\le N_\epsilon}\frac{1}{n}\sum_{i=1}^n\langle \boldsymbol{u}_j, \boldsymbol{w}_i\rangle^2 \ge 1 + 4\upsilon_1^2\sqrt{\frac{2v}{n}} + 2\upsilon_1^2\frac{v}{n}\right) \le \exp\{p\log(1+2/\epsilon) - v\}. \quad \text{(C.47)}$$

Finally, we take $\epsilon = 2/(e^2-1)$, $\gamma = e^{1/e}$ and $v = 2p + t$ ($t \ge 0$) in (C.46) and (C.47). Let $\mathcal{E}_1(t)$ be the event that (C.47) holds. Then, $\mathbb{P}\{\mathcal{E}_1(t)\} \ge 1 - e^{-t}$, and with $\mathbb{P}^*$-probability (over $\{e_i\}_{i=1}^n$) at least $1 - \lceil e\log(\frac{r_u}{r_l})\rceil e^{2p-u}$ conditioned on $\mathcal{E}_1(t)$,

$$\widehat{D}_h(\boldsymbol{\delta}) - \widehat{D}_h^{\flat}(\boldsymbol{\delta}) \le 3\bar{\tau}\sqrt{1 + 4\upsilon_1^2\sqrt{(4p+2t)/n} + \upsilon_1^2(4p+2t)/n}\cdot\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\sqrt{\frac{u}{n}}$$

holds for all $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying $r_l \le \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \le r_u$. This proves the claimed bound under the scaling $n \gtrsim p + t$. □

C.7 Proof of Theorem 4.5

The proof is based on an argument similar to that used in the proof of Theorem 4.2. To begin with, define the random process

$$\Delta^{\flat}(\boldsymbol{\delta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h^{\flat}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\widehat{Q}_h^{\flat}(\boldsymbol{\beta}^*) - \mathbf{J}_h\boldsymbol{\delta}\}, \quad \boldsymbol{\delta} \in \mathbb{R}^p. \tag{C.48}$$

For a prespecified $r > 0$, a key step is to bound the local fluctuation $\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta^{\flat}(\boldsymbol{\delta})\|_2$. Since $\mathbb{E}(w_i) = \mathbb{E}(1 + e_i) = 1$, we have $\mathbb{E}^*\{\nabla\widehat{Q}_h^{\flat}(\boldsymbol{\beta})\} = \nabla\widehat{Q}_h(\boldsymbol{\beta})$. Define the (conditionally) centered process

$$G^{\flat}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_h^{\flat}(\boldsymbol{\beta}) - \nabla\widehat{Q}_h(\boldsymbol{\beta})\} = \frac{1}{n}\sum_{i=1}^n \{\mathcal{K}_h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i) - \tau\}e_i\boldsymbol{w}_i,$$

so that $\Delta^{\flat}(\boldsymbol{\delta})$ be be written as

$$\Delta^{\flat}(\boldsymbol{\delta}) = \underbrace{\{G^{\flat}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G^{\flat}(\boldsymbol{\beta}^*)\}}_{\text{bootstrap error}} + \underbrace{\Delta(\boldsymbol{\delta})}_{\text{sampling error}},$$

where $\Delta(\boldsymbol{\delta})$ is given in (C.34). By the triangle inequality,

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta^{\flat}(\boldsymbol{\delta})\|_2 \le \sup_{\boldsymbol{\delta}\in\Theta(r)} \|G^{\flat}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G^{\flat}(\boldsymbol{\beta}^*)\|_2 + \sup_{\boldsymbol{\delta}\in\Theta(r)} \|\Delta(\boldsymbol{\delta})\|_2. \tag{C.49}$$

Let $\mathcal{E}_3(t)$ denote the event that the bound (C.38) holds. It suffices to deal with the first term on the right-hand side of (C.49). Using a change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \in \mathbb{B}^p(r)$ for $\boldsymbol{\delta} \in \Theta(r)$, we have $y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}$ and

$$\sup_{\boldsymbol{\delta}\in\Theta(r)} \|G^{\flat}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - G^{\flat}(\boldsymbol{\beta}^*)\|_2 \le \sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} \|\underbrace{G^{\flat}(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - G^{\flat}(\boldsymbol{\beta}^*)}_{=:\,\Delta_0^{\flat}(\boldsymbol{v})}\|_2, \tag{C.50}$$

where $\Delta_0^{\flat}(\boldsymbol{v}) = (1/n)\sum_{i=1}^n e_i\boldsymbol{w}_i\{\mathcal{K}_h(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v} - \varepsilon_i) - \mathcal{K}_h(-\varepsilon_i)\}$ satisfies $\Delta_0^{\flat}(\mathbf{0}) = \mathbf{0}$ and $\mathbb{E}^*\{\Delta_0^{\flat}(\boldsymbol{v})\} = \mathbf{0}$. Note that $\nabla\Delta_0^{\flat}(\boldsymbol{v}) = (1/n)\sum_{i=1}^n e_iK_{i,\boldsymbol{v}}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}$, where $K_{i,\boldsymbol{v}} = K_h(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v} - \varepsilon_i)$. For any $\lambda \in \mathbb{R}$ and $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{S}^{p-1}$, we have

$$\begin{aligned}
\mathbb{E}^* \exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\Delta_0^{\flat}(\boldsymbol{v})\boldsymbol{u}'\} &= \prod_{i=1}^n \mathbb{E}^* \exp\{\lambda n^{-1/2}e_iK_{i,\boldsymbol{v}}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'\} \\
&\le \prod_{i=1}^n \exp\left\{\frac{\lambda^2}{2n}K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2\right\} = \exp\left\{\frac{\lambda^2}{2n}\sum_{i=1}^n K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2\right\}.
\end{aligned}$$

Note that $K_{i,\boldsymbol{v}} \le \kappa_u h^{-1}$, and by Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^{n} K_{i,\boldsymbol{v}}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}\cdot\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^2 \le \frac{\kappa_u}{h}\cdot\left\{\frac{1}{n}\sum_{i=1}^{n} K_{i,\boldsymbol{v}}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} K_{i,\boldsymbol{v}}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}')^4\right\}^{1/2}.$$

Define the function $\Lambda_{r,h}(\cdot,\cdot):\mathbb{R}^p\times\mathbb{R}^p\mapsto[0,\infty)$

$$\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}-\varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \ \text{ for } \boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}, \tag{C.51}$$

and write $\|\Lambda_{r,h}\|_\infty = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})$. With this notation, it follows that for any $\lambda\in\mathbb{R}$ and $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$,

$$\sup_{\boldsymbol{\delta}\in\mathbb{B}^p(r)}\mathbb{E}^*\exp\{\lambda n^{1/2}\boldsymbol{u}^{\mathrm{T}}\Delta_0^{\flat}(\boldsymbol{\delta})\boldsymbol{v}\}\le\exp\left\{\frac{\lambda^2}{2}\kappa_u h^{-1}\|\Lambda_{r,h}\|_\infty\right\}.$$

Thus, applying a conditional version of Theorem A.3 in Spokoiny (2013) yields

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)}\|\Delta_0^{\flat}(\boldsymbol{v})\|_2 \le 6\kappa_u^{1/2}\|\Lambda_{r,h}\|_\infty^{1/2}\cdot r\sqrt{\frac{4p+2t}{nh}} \tag{C.52}$$

with $\mathbb{P}^*$-probability at least $1-e^{-t}$.

Next, we bound the data-dependent quantity $\|\Lambda_{r,h}\|_\infty$. For any $\epsilon_1,\epsilon_2\in(0,1)$, there exist ϵ_1- and ϵ_2-nets $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{d_1}\}$ and $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{d_2}\}$ of $\mathbb{S}^{p-1}$ with $d_1\le(1+2/\epsilon_1)^p$ and $d_2\le(1+2/\epsilon_2)^p$. Given $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$, there exist some $1\le j\le d_1$ and $1\le k\le d_2$ such that $\|\boldsymbol{u}-\boldsymbol{u}_j\|_2\le\epsilon_1$ and $\|\boldsymbol{v}-\boldsymbol{v}_k\|_2\le\epsilon_2$. At $(\boldsymbol{u},\boldsymbol{v})$, consider the decomposition

$$\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}) = \Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}) - \Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k) + \Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k).$$

For $\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})-\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)$, the Lipschitz continuity of $K(\cdot)$ ensures that

$$\begin{aligned}&|\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v})-\Lambda_{r,h}(\boldsymbol{u},\boldsymbol{v}_k)|\\ &\le\frac{l_K r}{nh^2}\sum_{i=1}^{n}|\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{v}_k)|(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \le \frac{l_K r\epsilon_2}{h^2}\cdot\max_{1\le i\le n}\|\boldsymbol{w}_i\|_2\cdot\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4.\end{aligned} \tag{C.53}$$

For $\Lambda_{r,h}(\boldsymbol{u}, \boldsymbol{v}_k)$, by the triangle inequality for the ℓ_4-norm we have

$$\begin{aligned}
\Lambda_{r,h}^{1/4}(\boldsymbol{u}, \boldsymbol{v}_k) &= \left\{\frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4\right\}^{1/4} \\
&\le \left\{\frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\right\}^{1/4} + \left\{\frac{1}{n}\sum_{i=1}^{n} K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)\langle \boldsymbol{w}_i, \boldsymbol{u} - \boldsymbol{u}_j\rangle^4\right\}^{1/4} \\
&\le \Lambda_{r,h}^{1/4}(\boldsymbol{u}_j, \boldsymbol{v}_k) + \epsilon_1 \cdot \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \Lambda_{r,h}^{1/4}(\boldsymbol{u}, \boldsymbol{v}_k),
\end{aligned}$$

which in turn implies

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \Lambda_{r,h}(\boldsymbol{u}, \boldsymbol{v}_k) \le (1-\epsilon_1)^{-4} \max_{1\le j\le d_1} \Lambda_{r,h}(\boldsymbol{u}_j, \boldsymbol{v}_k). \tag{C.54}$$

In view of (C.53) and (C.54), it suffices to bound $\max_{1\le i\le n} \|\boldsymbol{w}_i\|_2$, $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} (1/n)\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4$ and $\max_{(j,k)\in[d_1]\times[d_2]} \Lambda_{r,h}(\boldsymbol{u}_j, \boldsymbol{v}_k)$.

Lemma C.5. For any $t \ge 0$, $\max_{1\le i\le n} \|\boldsymbol{w}_i\|_2^2 \le C_1 \upsilon_1^2 (p + \log n + t)$ with probability at least $1 - e^{-t}$, where $C_1 > 0$ is an absolute constant.

For the supremum $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} (1/n)\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4$, similarly to (C.54) it can be shown that

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \le (1-\epsilon_1)^{-4} \max_{1\le j\le d_1} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4.$$

Fix j and k, Condition 4.5 implies

$$\begin{aligned}
\mathbb{E}e^{\{(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4/(36\upsilon_1^4)\}^{1/2}} &= \mathbb{E}e^{(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^2/(6\upsilon_1^2)} = 1 + \int_0^{\infty} e^u \mathbb{P}\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j| \ge \upsilon_1 (6u)^{1/2}\}\mathrm{d}u \\
&\le 1 + 2\int_0^{\infty} e^{u-3u}\mathrm{d}u = 1 + 1 = 2.
\end{aligned}$$

Therefore, $\|(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\|_{\psi_{1/2}} \le 36\upsilon_1^4$, where $\|\cdot\|_{\psi_r}$ denotes the ψ_r-norm ($r > 0$); see Definition 2.1 in Adamczak et al. (2011). Since $0 \le K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i) \le \kappa_u h^{-1}$, it is easy to see that $\|K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\|_{\psi_{1/2}} \le 36\kappa_u \upsilon_1^4 h^{-1}$. Moreover, note that $\mathbb{E}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4 \le m_4$ and

$$\mathbb{E}\{K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\} = \mathbb{E}[\mathbb{E}\{K_h(r\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_k - \varepsilon_i)|\boldsymbol{x}_i\}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4] \le \bar{f} m_4.$$

Hence, for any $u, v \ge 3$, applying inequality (3.6) (and those above it) with $s = 1/2$ in Adamczak et

al. (2011) and the union bound, we obtain that

$$\mathbb{P}\left\{\max_{1\le j\le d_1}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4 \ge m_4 + C_2\upsilon_1^4\left(\sqrt{\frac{u}{n}}+\frac{u^2}{n}\right)\right\} \le d_1 e^{-u}$$

and

$$\mathbb{P}\left\{\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k) \ge \bar{f}m_4 + C_2\kappa_u\upsilon_1^4\left(\sqrt{\frac{v}{nh^2}}+\frac{v^2}{nh}\right)\right\} \le d_1 d_2 e^{-v}.$$

Taking $\epsilon_1 = 1-2^{-1/4}$, $\epsilon_2 = n^{-2}$, $u = p\log(1+2/\epsilon_1)+t$ and $v = p\log\{(1+2/\epsilon_1)(1+2/\epsilon_2)\}+t$ in the above bounds, it follows that with probability at least $1-2e^{-t}$,

$$\max_{1\le j\le d_1}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4 \le m_4 + C_3\upsilon_1^4\left\{\sqrt{\frac{p+t}{n}}+\frac{(p+t)^2}{n}\right\}$$

and

$$\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k) \le \bar{f}m_4 + C_4\kappa_u\upsilon_1^4\left\{\sqrt{\frac{p\log n+t}{nh^2}}+\frac{(p\log n+t)^2}{nh}\right\}.$$

Denote by $\mathcal{E}_3(t)$ the event that the above two bounds and the bound in Lemma C.5 are satisfied. Then $\mathbb{P}\{\mathcal{E}_3(t)\} \ge 1-3e^{-t}$. Conditioned on $\mathcal{E}_3(t)$ with $t=\log n$, we have

$$\begin{aligned}\|\Lambda_{r,h}\|_\infty &\le 2\max_{(j,k)\in[d_1]\times[d_2]}\Lambda(\boldsymbol{u}_j,\boldsymbol{v}_k) + \frac{2l_K r}{n^2h^2}\cdot\max_{1\le i\le n}\|\boldsymbol{w}_i\|_2\cdot\sup_{1\le j\le d_1}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}_j)^4\\ &\le 2\bar{f}m_4 + C_5\frac{\upsilon_1^4}{h}\left\{\sqrt{\frac{p\log n}{n}}+\frac{(p\log n)^2}{n}\right\} + C_6\frac{\upsilon_1^5 r}{(nh)^2}(p+\log n)^{1/2}. \qquad \text{(C.55)}\end{aligned}$$

With the above preparations, we are ready to prove the Bahadur representation for the bootstrap estimate. Let $\mathcal{E}(t) = \mathcal{E}_1(t)\cap\mathcal{E}_2(t)$ be the event from Theorem 4.4 and its proof. In the rest of the proof, we take $t=\log n$, and set the bandwidth $h \asymp (q/n)^{2/5}$ with $q = p+\log n$. Recall that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\delta}}^\flat = \widehat{\boldsymbol{\beta}}_h^\flat - \boldsymbol{\beta}^*$. Conditioned on $\mathcal{E}_1(t)\cap\mathcal{E}_2(t)$, $\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le r_{\mathrm{est}} \asymp \sqrt{q/n}$, and $\|\widehat{\boldsymbol{\delta}}^\flat\|_{\boldsymbol{\Sigma}} \le r_{\mathrm{est}}^\flat \asymp \sqrt{q/n}$ with $\mathbb{P}^*$-probability at least $1-2n^{-1}$. Conditioned further on $\mathcal{E}_3(t)$,

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h-\boldsymbol{\beta}^*)+\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_h(\boldsymbol{\beta}^*)\|_2 = \|\Delta(\widehat{\boldsymbol{\delta}})\|_2 \le \sup_{\boldsymbol{\delta}\in\Theta(r_{\mathrm{est}})}\|\Delta(\boldsymbol{\delta})\|_2 \lesssim \left(\frac{q}{n}\right)^{4/5},$$

and with $\mathbb{P}^*$-probability at least $1 - 3n^{-1}$,

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h^\flat - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_h^\flat(\boldsymbol{\beta}^*)\|_2 = \|\Delta^\flat(\widehat{\boldsymbol{\delta}}^\flat)\|_2 \\
&\le \sup_{\boldsymbol{\delta}\in\Theta(r_{\mathrm{est}}^\flat)} \|\Delta^\flat(\boldsymbol{\delta})\|_2 \lesssim \left(\frac{q}{n}\right)^{4/5} \bigvee \left(\frac{q}{n}\right)^{3/5}\left(\frac{p\log n}{n}\right)^{1/4} \bigvee \left(\frac{q}{n}\right)^{3/5}\frac{p\log n}{n^{1/2}}.
\end{aligned}$$

Together, the above two bounds proves the claimed result. □

C.7.1 Proof of Lemma C.5

Note that, after centering, $\boldsymbol{w}_i = (1, \boldsymbol{w}_{i,-}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{w}_{i,-} \in \mathbb{R}^{p-1}$ is a zero-mean sub-Gaussian random vector. Under Condition 4.5, there exists some constant $\upsilon_2 \asymp \upsilon_1$ depending only on υ_1 such that $\mathbb{E}\exp\{\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{e}_1)\} \le \exp(\|\boldsymbol{\alpha}\|_2^2\upsilon_2^2/2)$ for all $\boldsymbol{\alpha} \in \mathbb{R}^p$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$. Then, applying Theorem 2.1 in Hsu, Kakade and Zhang (2012) with $\Sigma = A = \mathbf{I}_p$ yields that, for any $u \ge 0$,

$$\|\boldsymbol{w}_i\|_2^2 \le \upsilon_2^2(p + 2\sqrt{pu} + 2u) + 1 + 2(u/p)^{1/2} \le \upsilon_2^2(2p + 3u) + 1 + 2(u/p)^{1/2}$$

holds with probability at least $1 - e^{-u}$. Taking the union bound over $i = 1, \ldots, n$ and setting $u = \log n + t$ prove the claimed bound. □

C.8 Proof of Proposition 4.2

Consider the change of variable $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$, so that $\boldsymbol{\delta} \in \Theta(r)$ is equivalent to $\boldsymbol{v} \in \mathbb{B}^p(r)$. Write $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i \in \mathbb{R}^p$, which are isotropic random vectors, and define

$$\widehat{\mathbf{H}}_h(\boldsymbol{v}) = \boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{J}}_h(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1/2} = \frac{1}{n}\sum_{i=1}^n K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v})\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}, \quad \mathbf{H}_h(\boldsymbol{v}) = \mathbb{E}\{\widehat{\mathbf{H}}_h(\boldsymbol{v})\}. \tag{C.56}$$

For any $\epsilon_1 \in (0, r)$, there exists an ϵ-net $\mathcal{N}_1 := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d_1}\} \subseteq \mathbb{B}^p(r)$ with $d_1 \le (1+2r/\epsilon_1)^p$ satisfying that, for every $\boldsymbol{v} \in \mathbb{B}^p(r)$, there exists some $1 \le j \le d_1$ such that $\|\boldsymbol{v} - \boldsymbol{v}_j\|_2 \le \epsilon_1$. Hence,

$$\begin{aligned}
&\|\widehat{\mathbf{H}}_h(\boldsymbol{v}) - \mathbf{H}_h(\mathbf{0})\|_2 \\
&\le \|\widehat{\mathbf{H}}_h(\boldsymbol{v}) - \widehat{\mathbf{H}}_h(\boldsymbol{v}_j)\|_2 + \|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 + \|\mathbf{H}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\mathbf{0})\|_2 \\
&=: I_1(\boldsymbol{v}) + I_{2,j} + I_{3,j}.
\end{aligned}$$

For $I_1(\boldsymbol{v})$, note that $K_h(u) = (1/h)K(u/h)$ is Lipschitz continuous, i.e. $|K_h(u) - K_h(v)| \le l_K h^{-2}|u - v|$ for all $u, v \in \mathbb{R}$. It follows that

$$\begin{aligned} I_1(\boldsymbol{v}) &\le \sup_{\boldsymbol{u},\boldsymbol{u}' \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}) - K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)| \cdot |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \\ &\le l_K h^{-2} \sup_{\boldsymbol{u},\boldsymbol{u}' \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{v}_j) \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u} \cdot \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \\ &\le l_K h^{-2}\epsilon_1 \underbrace{\sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}|^3}_{=:M_{n,3}}. \end{aligned} \tag{C.57}$$

Next, we use the standard covering argument to bound $M_{n,3}$. Given $\epsilon_2 \in (0, 1)$, let $\mathcal{N}_2$ be an ϵ_2-net of the unit sphere $\mathbb{S}^{p-1}$ with $d_2 := |\mathcal{N}_2| \le (1 + 2/\epsilon_2)^p$ such that for every $\boldsymbol{u} \in \mathbb{S}^{p-1}$, there exists some $\boldsymbol{u}' \in \mathcal{N}_2$ satisfying $\|\boldsymbol{u} - \boldsymbol{u}'\|_2 \le \epsilon_2$. Define the (standardized) design matrix $\mathbf{W}_n = n^{-1/3}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, so that $M_{n,3} = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\mathbf{W}_n\boldsymbol{u}\|_3^3$. By the triangle inequality,

$$\begin{aligned} \|\mathbf{W}_n\boldsymbol{u}\|_3 &\le \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \|\mathbf{W}_n(\boldsymbol{u} - \boldsymbol{u}')\|_3 \\ &= \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}(\boldsymbol{u} - \boldsymbol{u}')|^3\right)^{1/3} \le \|\mathbf{W}_n\boldsymbol{u}'\|_3 + \epsilon_2 \cdot M_{n,3}^{1/3}. \end{aligned}$$

Taking the maximum over $\boldsymbol{u}' \in \mathcal{N}_2$, and then taking the supremum over $\boldsymbol{u} \in \mathbb{S}^{p-1}$, we arrive at

$$M_{n,3} \le (1 - \epsilon_2)^{-3} \cdot N_{n,3} := (1 - \epsilon_2)^{-3} \cdot \max_{\boldsymbol{u}'\in\mathcal{N}_2} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3. \tag{C.58}$$

For every $\boldsymbol{u}' \in \mathcal{N}_2$, note that

$$\mathbb{E}e^{\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3/(6^{3/2}\upsilon_1^3)\}^{2/3}} = 1 + \int_0^\infty e^u \mathbb{P}\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'| \ge \upsilon_1(6u)^{1/2}\}\mathrm{d}u \le 1 + 2\int_0^\infty e^{-2u}\mathrm{d}u = 2,$$

implying $\||\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3\|_{\psi_{2/3}} \le 6^{3/2}\upsilon_1^3$. Hence, by inequality (3.6) in Adamczak et al. (2011) with $s = 2/3$, we obtain that for any $z \ge 3$,

$$\frac{1}{n}\sum_{i=1}^n |\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}'|^3 \le \mathbb{E}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}'|^3 + C_1\upsilon_1^3\left(\sqrt{\frac{z}{n}} + \frac{z^{3/2}}{n}\right)$$

with probability at least $1 - e^{-z}$. Taking the union bound over all vectors $\boldsymbol{u}'$ in $\mathcal{N}_2$ yields that, with

probability at least $1 - d_2 e^{-z} \geq 1 - e^{p\log(1+2/\epsilon_2)-z}$,

$$N_{n,3} \leq m_3 + C_1 \upsilon_1^3 \left(\sqrt{\frac{z}{n}} + \frac{z^{3/2}}{n} \right)$$

where $m_3 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}|^3$. Reorganizing the terms, we get

$$N_{n,3} \leq m_3 + C_1 \upsilon_1^3 \left[\sqrt{\frac{p\log(1+2/\epsilon_2) + \log 2 + t}{n}} + \frac{\{p\log(1+2/\epsilon_2) + \log 2 + t\}^{3/2}}{n} \right] \tag{C.59}$$

with probability at least $1 - e^{-t}/2$. Taking $\epsilon_2 = 1/8$ in (C.58) and (C.59) implies

$$M_{n,3} \leq 1.5 m_3 + 1.5 C_1 \upsilon_1^3 \left\{ \sqrt{\frac{3p+1+t}{n}} + \frac{(3p+1+t)^{3/2}}{n} \right\}.$$

Under the sample size scaling $n \gtrsim p + t$, plugging the above bound into (C.57) yields

$$\sup_{\boldsymbol{v}\in\mathbb{B}^p(r)} I_1(\boldsymbol{v}) \lesssim \upsilon_1^3 (p+t)^{1/2} h^{-2} \epsilon_1 \tag{C.60}$$

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

Turning to $I_{2,j}$, note that $\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j) = (1/n)\sum_{i=1}^n (1-\mathbb{E})\phi_{ij}\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}$, where $\phi_{ij} = K_h(\varepsilon_i - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)$ satisfy $|\phi_{ij}| \leq \kappa_u h^{-1}$ and

$$\mathbb{E}(\phi_{ij}^2|\boldsymbol{x}_i) = \frac{1}{h^2}\int_{-\infty}^{\infty} K^2\left(\frac{\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle - t}{h}\right) f_{\varepsilon_i|\boldsymbol{x}_i}(t)\,\mathrm{d}t = \frac{1}{h}\int_{-\infty}^{\infty} K^2(u) f_{\varepsilon_i|\boldsymbol{x}_i}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v} - bu)\,\mathrm{d}u \leq \bar{f}\kappa_u h^{-1}$$

almost surely. Given $\epsilon_3 \in (0, 1/2)$, there exits an ϵ_3-net $\mathcal{N}_3$ of the sphere $\mathbb{S}^{p-1}$ with $|\mathcal{N}_3| \leq (1+2/\epsilon_3)^p$ such that $\|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 \leq (1-2\epsilon_3)^{-1}\max_{\boldsymbol{u}\in\mathcal{N}_3} |\boldsymbol{u}^{\mathrm{T}}\{\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\}\boldsymbol{u}|$. Given $\boldsymbol{u} \in \mathcal{N}_3$ and $k = 2, 3, \ldots$, we bound the higher order moments of $\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2$ by

$$\begin{aligned}
\mathbb{E}|\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2|^k &\leq \bar{f}\kappa_u h^{-1} \cdot (\kappa_u h^{-1})^{k-2} \upsilon_1^{2k} \cdot 2k \int_0^{\infty} \mathbb{P}(|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u}| \geq \upsilon_1 u) u^{2k-1}\mathrm{d}u \\
&\leq \bar{f}\kappa_u h^{-1} \cdot (\kappa_u h^{-1})^{k-2} \upsilon_1^{2k} \cdot 4k \int_0^{\infty} u^{2k-1} e^{-u^2/2}\mathrm{d}u \\
&\leq \bar{f}\kappa_u h^{-1} \cdot (\kappa_u h^{-1})^{k-2} \upsilon_1^{2k} \cdot 2^{k+1} k!.
\end{aligned}$$

In particular, $\mathbb{E}\phi_{ij}^2(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^4 \leq (4\upsilon_1^2)^2 \bar{f}\kappa_u h^{-1}$, and $\mathbb{E}|\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{u})^2|^k \leq \frac{k!}{2} \cdot (4\upsilon_1^2)^2 \bar{f}\kappa_u h^{-1} \cdot (2\upsilon_1^2\kappa_u h^{-1})^{k-2}$ for

$k \geq 3$. Applying Bernstein's inequality and the union bound, we find that for any $u \geq 0$,

$$\begin{aligned}
&\|\widehat{\mathbf{H}}_h(\boldsymbol{v}_j) - \mathbf{H}_h(\boldsymbol{v}_j)\|_2 \\
&\leq \frac{1}{1-2\epsilon_3} \max_{\boldsymbol{u} \in \mathcal{N}_3} \left| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E})\phi_{ij}(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{u})^2 \right| \leq \frac{2\upsilon_1^2}{1-2\epsilon_3} \left(2\sqrt{2\bar{f}\kappa_u \frac{u}{nh}} + \kappa_u \frac{u}{nh} \right)
\end{aligned}$$

with probability at least $1 - 2(1+2/\epsilon_3)^p e^{-u} = 1 - (1/2)e^{\log(4)+p\log(1+2/\epsilon_3)-u}$. Setting $\epsilon_3 = 2/(e^3-1)$ and $u = \log(4) + 3p + v$, it follows that with probability at least $1 - e^{-v}/2$,

$$I_{2,j} \lesssim \upsilon_1^2 \left(\sqrt{\frac{p+v}{nh}} + \frac{p+v}{nh} \right).$$

Once again, taking the union bound over $j = 1, \ldots, d_1$ and setting $v = p\log(1+2r/\epsilon_1) + t$, we obtain that with probability at least $1 - d_1 e^{-v} \geq 1 - e^{-t}/2$,

$$\max_{1 \leq j \leq d_1} I_{2,j} \lesssim \sqrt{\frac{p\log(3er/\epsilon_1) + t}{nh}} + \frac{p\log(3er/\epsilon_1) + t}{nh}. \tag{C.61}$$

For $I_{3,j}$, following the proof of (C.8) it can similarly shown that $I_{3,j} \leq 0.5 l_0 m_3 r$. Combining this with (C.60) and (C.61), and taking $\epsilon_1 = r/n \in (0, r)$ in the beginning of the proof, we conclude that with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p\log n + t}{nh}} + \frac{p\log n + t}{nh} + \frac{(p+t)^{1/2} r}{nh^2} + r$$

as long as $n \gtrsim p + t$. This leads to (4.18) under the prescribed bandwidth constraint.

To derive the same bound for $\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h$, notice that $u \mapsto \{\mathcal{K}_h(u) - \tau\}^2$ is a $(2\bar{\tau}\kappa_u/h)$-Lipschitz continuous function, where $\bar{\tau} = \max(\tau, 1-\tau)$. Moreover, for every $\boldsymbol{\delta} \in \mathbb{R}^p$, the random variable $\{\mathcal{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle - \varepsilon_i) - \tau\}^2$ takes values in $[0, \bar{\tau}^2]$. We can thus apply the same argument to bound $\sup_{\boldsymbol{\delta} \in \Theta(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2$. □

D Theoretical properties of one-step conquer

In this section, we provide theoretical properties of the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$, defined in Section B.1. The key message is that, when higher-order kernels are used (and if the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ has enough derivatives), the asymptotic normality of the one-step estimator holds under weaker growth conditions on p. For example, the scaling condition $p = o(n^{3/8})$ that is

required for the conquer estimator can be reduced to roughly $p = o(n^{7/16})$ for the one-step conquer estimator using a kernel of order 4.

Condition 1. Let $G(\cdot)$ be a symmetric kernel of order $\nu > 2$, that is, $\int_{-\infty}^{\infty} u^k G(u)\,\mathrm{d}u = 0$ for $k = 1, \ldots, \nu - 1$ and $\int_{-\infty}^{\infty} u^\nu G(u)\,\mathrm{d}u \neq 0$. Moreover, $g_k = \int_{-\infty}^{\infty} |u^k G(u)|\,\mathrm{d}u < \infty$ for $1 \leq k \leq \nu$, G is uniformly bounded with $g_u = \sup_{u\in\mathbb{R}} |G(u)| < \infty$ and is l_G-Lipschitz continuous for some $l_G > 0$.

As before, we write $\mathcal{G}_b(u) = \mathcal{G}(u/b)$ and $\mathcal{G}(u) = \int_{-\infty}^{u} G(v)\,\mathrm{d}v$ for $u \in \mathbb{R}$ and $b > 0$. The use of a higher-order kernel does not necessarily reduce bias unless the conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ of ε given $\boldsymbol{x}$ is sufficiently smooth. Therefore, we further impose the following smoothness conditions on $f_{\varepsilon|\boldsymbol{x}}(\cdot)$.

Condition 2. Let $\nu \geq 4$ be the integer in Condition 1. The conditional density $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is $(\nu - 1)$-times differentiable, and satisfies $|f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(u) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)| \leq l_{\nu-2}|u|$ for all $u \in \mathbb{R}$ almost surely (over the random vector $\boldsymbol{x}$), where $l_{\nu-2} > 0$ is a constant. Also, there exists some constant $C_G > 0$ such that $\int_{-\infty}^{\infty} |u^{\nu-1}G(u)| \cdot \sup_{|t|\leq|u|} |f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(t) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)|\,\mathrm{d}u \leq C_G$ almost surely.

Notably, we have

$$\nabla Q_b^G(\boldsymbol{\beta}) = \mathbb{E}\{\mathcal{G}_b(\langle \boldsymbol{x}, \boldsymbol{\beta}\rangle - y) - \tau\}\boldsymbol{x} \quad \text{and} \quad \nabla^2 Q_b^G(\boldsymbol{\beta}) = \mathbb{E}\{G_b(y - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}, \tag{D.1}$$

representing the population score and Hessian of $Q_b^G(\cdot) = \mathbb{E}\widehat{Q}_b^G(\cdot)$. As $b \to 0$, we expect $\nabla Q_b^G(\boldsymbol{\beta}^*)$ and $\nabla^2 Q_b^G(\boldsymbol{\beta}^*)$ to converge to $\mathbf{0}$ (zero vector in $\mathbb{R}^p$) and $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, respectively. The following proposition validates this claim by providing explicit error bounds.

Proposition D.1. Let $b \in (0, 1)$ be a bandwidth. Under Conditions 1 and 2, we have

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 \leq l_{\nu-2} g_\nu\, b^\nu/\nu! \quad \text{and} \quad \|\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_b^G(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\|_2 \leq C_G\, b^{\nu-1}/(\nu-1)!,$$

where $\mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$ with $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$.

Proposition D.2 shows that when a higher-order kernel is used, the bias is significantly reduced in the sense that $\|\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 = O(b^\nu)$ and $\|\nabla^2 Q_b^G(\boldsymbol{\beta}^*) - \mathbf{J}\|_2 = O(b^{\nu-1})$, where $\nu \geq 4$ is an even integer. Notably, even if the kernel G has negative parts, the population Hessian $\nabla^2 Q_b^G(\boldsymbol{\beta}^*)$ preserves the positive definiteness of $\mathbf{J}$ as long as the bandwidth b is sufficiently small.

To construct the one-step conquer estimator, two key quantities are the sample Hessian $\nabla^2\widehat{Q}_b^G(\cdot)$ and sample gradient $\nabla\widehat{Q}_b^G(\cdot)$, both evaluated at $\overline{\boldsymbol{\beta}}$, a consistent initial estimate. In the next two

propositions, we establish uniform convergence results of the Hessian and gradient of the empirical smoothed loss to their population counterparts. As a direct consequence, $\nabla^2 \widehat{Q}_b^G(\overline{\boldsymbol{\beta}})$ is positive definite with high probability, provided that $\overline{\boldsymbol{\beta}}$ is consistent (i.e., in a local vicinity of $\boldsymbol{\beta}^*$). To be more specific, for $r > 0$, we define the local neighborhood

$$\Theta^*(r) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq r\}. \tag{D.2}$$

Proposition D.2. Conditions 1, 2 and 4.5 ensure that, with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 \widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla^2 Q_b^G(\boldsymbol{\beta})\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p \log n + t}{nb}} + \frac{p \log n + t}{nb} + \frac{(p+t)^{1/2} r}{nb^2}$$

as long as $n \gtrsim p + t$.

Proposition D.3. Conditions 1, 2 and 4.5 ensure that, with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2 \lesssim r\left(\sqrt{\frac{p+t}{nb}} + r + b^{\nu-1}\right) \tag{D.3}$$

as long as $\sqrt{(p+t)/n} \lesssim b$.

With the above preparations, we are ready to present the Bahadur representation for the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$.

Theorem D.1. Assume Conditions 4.1, 4.2 and 4.5 in the main text and Conditions 1 and 2 hold. For any $t > 0$, let the sample size n, dimension p and the bandwidths $h, b > 0$ satisfy $n \gtrsim p(\log n)^2 + t$, $\sqrt{(p+t)/n} \lesssim h \lesssim \{(p+t)/n\}^{1/4}$ and $\sqrt{(p+t)/n} \lesssim b \lesssim \{(p+t)/n\}^{1/(2\nu)}$. Then, the one-step conquer estimator $\widehat{\boldsymbol{\beta}}$ satisfies the bound

$$\left\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n \{\tau - \mathcal{G}_b(-\varepsilon_i)\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\right\|_2 \lesssim \left\{\underbrace{\sqrt{(p \log n + t)/(nb)}}_{\text{variance term}} + \underbrace{b^{\nu-1}}_{\text{bias term}}\right\}\sqrt{\frac{p+t}{n}} \tag{D.4}$$

with probability at least $1 - 5e^{-t}$.

Theorem D.1 shows that using a higher-order kernel ($\nu \geq 4$) allows one to choose larger bandwidth, thereby reducing the "variance" and the total Bahadur linearization error. Similarly to Theorem 4.3 in the main text, the following asymptotic normal approximation result for linear projections of one-step conquer is a direct consequence of Theorem D.1.

Theorem D.2. Assume Conditions 4.1, 4.2 and 4.5 in the main text and Conditions 1 and 2 hold. Let the bandwidths satisfy $(q/n)^{1/2} \lesssim h \lesssim (q/n)^{1/4}$ and $(q/n)^{1/2} \lesssim b \lesssim (q/n)^{1/(2\nu)}$, where $q := p + \log n$. Then,

$$\sup_{x\in\mathbb{R},\, \boldsymbol{a}\in\mathbb{R}^p} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle/\sigma_0 \le x) - \Phi(x)\right| \lesssim \sqrt{\frac{(p+\log n)p\log n}{nb}} + n^{1/2}b^{\nu}, \tag{D.5}$$

where $\sigma_0^2 = \sigma_0^2(\boldsymbol{a}) = \tau(1-\tau)\|\mathbf{J}^{-1}\boldsymbol{a}\|_{\boldsymbol{\Sigma}}^2$. In particular, with a choice of bandwidth $b \asymp (q/n)^{2/(2\nu+1)}$,

$$\sup_{x\in\mathbb{R},\, \boldsymbol{a}\in\mathbb{R}^p} \left|\mathbb{P}(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \le \sigma_0 x) - \Phi(x)\right| \to 0$$

as $n \to \infty$ under the scaling $p^{4\nu/(2\nu-1)}(\log n)^{(2\nu+1)/(2\nu-1)} = o(n)$.

Let $G(\cdot)$ be a kernel of order $\nu = 4$. In view of Theorem D.2, we take $h \asymp \{(p+\log n)/n\}^{2/5}$ as in the main text and $b = \{(p+\log n)/n\}^{2/9}$, thereby obtaining that $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle$, for an arbitrary $\boldsymbol{a} \in \mathbb{R}^p$, is asymptotically normally distributed as long as $p(\log n)^{9/16} = o(n^{7/16})$ as $n \to \infty$.

D.1 Proof of Proposition D.1

We start from the gradient $\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*) = \mathbb{E}\{\mathcal{G}_b(-\varepsilon) - \tau\}\boldsymbol{w}$ with $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$. Let $\mathbb{E}_{\boldsymbol{x}}$ be the conditional expectation given $\boldsymbol{x}$. By integration by parts,

$$\mathbb{E}_{\boldsymbol{x}}\mathcal{G}_b(-\varepsilon) = \int_{-\infty}^{\infty} \mathcal{G}(-t/b)\,\mathrm{d}F_{\varepsilon|\boldsymbol{x}}(t) = \int_{-\infty}^{\infty} G(u)F_{\varepsilon|\boldsymbol{x}}(-bu)\,\mathrm{d}u. \tag{D.6}$$

Applying a Taylor series expansion with integral remainder on $F_{\varepsilon|\boldsymbol{x}}(-bu)$ yields

$$\begin{aligned} F_{\varepsilon|\boldsymbol{x}}(-bu) &= F_{\varepsilon|\boldsymbol{x}}(0) + \sum_{\ell=1}^{\nu-1} F_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^{\ell}}{\ell!} + \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\{F_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - F_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\}\mathrm{d}w \\ &= \tau + \sum_{\ell=0}^{\nu-2} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^{\ell+1}}{(\ell+1)!} + \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)\}\mathrm{d}w. \end{aligned}$$

Recall that G is a kernel of order $\nu \ge 4$ (an even integer) and $g_\nu = \int_{-\infty}^{\infty} |u^\nu G(u)|\,\mathrm{d}u < \infty$. Substituting the above expansion into (D.6), we obtain

$$\mathbb{E}_{\boldsymbol{x}}\mathcal{G}_b(-\varepsilon) = \tau - \frac{b^{\nu-1}}{(\nu-2)!}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(0)\}\mathrm{d}w\mathrm{d}u.$$

Furthermore, by the Lipschitz continuity of $f_{\varepsilon|\boldsymbol{x}}^{(\nu-2)}(\cdot)$ around 0,

$$
\begin{aligned}
|\mathbb{E}_{\boldsymbol{x}}\mathcal{G}_b(-\varepsilon) - \tau| &\le \frac{l_{\nu-2}b^{\nu}}{(\nu-2)!}\int_{-\infty}^{\infty}\int_0^1 |u^{\nu}G(u)|(1-w)^{\nu-2}w\,\mathrm{d}w\mathrm{d}u \\
&= B(2,\nu-1)l_{\nu-2}g_{\nu}\,b^{\nu}/(\nu-2)!,
\end{aligned}
$$

where $B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$ denotes the beta function. In particular, $B(2,\nu-1) = \Gamma(2)\Gamma(\nu-1)/\Gamma(\nu+1) = (\nu-2)!/\nu!$. Putting together the pieces yields

$$
\|\boldsymbol{\Sigma}^{-1/2}\nabla Q_b^G(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\,\mathbb{E}_{\boldsymbol{x}}\{\mathcal{G}_b(-\varepsilon)-\tau\}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u} \le l_{\nu-2}g_{\nu}\,b^{\nu}/\nu!.
$$

Turning to the Hessian, note that

$$
\|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 Q_b^G(\boldsymbol{\beta}^*) - \mathbf{J}\}\boldsymbol{\Sigma}^{-1/2}\|_2 = \left\|\mathbb{E}\int_{-\infty}^{\infty} G(u)\{f_{\varepsilon|\boldsymbol{x}}(-bu) - f_{\varepsilon|\boldsymbol{x}}(0)\}\mathrm{d}u\,\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2.
$$

Applying a similar Taylor series expansion as above, we have

$$
f_{\varepsilon|\boldsymbol{x}}(t) = f_{\varepsilon|\boldsymbol{x}}(0) + \sum_{\ell=1}^{\nu-1} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{t^{\ell}}{\ell!} + \frac{t^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(tw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\}\mathrm{d}w. \tag{D.7}
$$

Under Conditions 1 and 2, it follows that

$$
\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2 Q_b^G(\boldsymbol{\beta}^*) - \mathbf{J}\}\boldsymbol{\Sigma}^{-1/2}\|_{\boldsymbol{\Omega}} \\
&\le \frac{b^{\nu-1}}{(\nu-2)!}\sup_{\boldsymbol{u},\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \mathbb{E}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\}\mathrm{d}w\,\mathrm{d}u\,\langle\boldsymbol{w},\boldsymbol{u}\rangle\langle\boldsymbol{w},\boldsymbol{\delta}\rangle \\
&\le \frac{b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u},\boldsymbol{\delta}\in\mathbb{S}^{p-1}} \mathbb{E}\int_{-\infty}^{\infty}|u^{\nu-1}G(u)|\sup_{|t|\le b|u|}\left|f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(t) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\right|\mathrm{d}u\cdot|\langle\boldsymbol{w},\boldsymbol{u}\rangle\langle\boldsymbol{w},\boldsymbol{\delta}\rangle| \\
&\le \frac{C_G b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle\boldsymbol{w},\boldsymbol{u}\rangle^2 = \frac{C_G}{(\nu-1)!}b^{\nu-1}.
\end{aligned}
$$

This completes the proof. □

D.2 Proof of Proposition D.2

The proof is almost identical to that of Proposition 4.2, and thus is omitted. □

D.3 Proof of Proposition D.3

Define the stochastic process $\Delta_b(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_b^G(\boldsymbol{\beta}) - \nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}$. By the triangle inequality,

$$\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta})\|_2 \le \sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2 + \sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2$$

Recall that $\mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$. For the first term on the right-hand side, using the mean value theorem for vector-valued functions yields

$$\mathbb{E}\Delta_b(\boldsymbol{\beta}) = \left\{\boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2 Q_b^G((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta})\,\mathrm{d}s\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{J}_0\right\}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

By a change of variable $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$,

$$\nabla^2 Q_b^G((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta}) = \mathbb{E}\int_{-\infty}^{\infty} G(u) f_{\varepsilon|\boldsymbol{x}}(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta} - bu)\,\mathrm{d}u \cdot \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.$$

For every $s \in [0,1]$ and $u \in \mathbb{R}$, it ensures from $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ being Lipschitz that

$$|f_{\varepsilon|\boldsymbol{x}}(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta} - bu) - f_{\varepsilon|\boldsymbol{x}}(-bu)| \le l_0 s \cdot |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}|.$$

Moreover, by the Taylor series expansion (D.7),

$$f_{\varepsilon|\boldsymbol{x}}(-bu) = f_{\varepsilon|\boldsymbol{x}}(0) + \sum_{\ell=1}^{\nu-1} f_{\varepsilon|\boldsymbol{x}}^{(\ell)}(0)\frac{(-bu)^\ell}{\ell!} + \frac{(-bu)^{\nu-1}}{(\nu-2)!}\int_0^1 (1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\}\,\mathrm{d}w.$$

Consequently,

$$\begin{aligned}
&\left\|\boldsymbol{\Sigma}^{-1/2}\int_0^1 \nabla^2 Q_b^G((1-s)\boldsymbol{\beta}^* + s\boldsymbol{\beta})\,\mathrm{d}s\,\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\right\|_2 \\
&\le \left\|\frac{b^{\nu-1}}{(\nu-2)!}\mathbb{E}\int_{-\infty}^{\infty}\int_0^1 u^{\nu-1}G(u)(1-w)^{\nu-2}\{f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(-buw) - f_{\varepsilon|\boldsymbol{x}}^{(\nu-1)}(0)\}\,\mathrm{d}w\mathrm{d}u \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2 \\
&\quad + 0.5 l_0 \cdot \left\|\mathbb{E}\,|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}| \cdot \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\right\|_2 \\
&\le \frac{C_G b^{\nu-1}}{(\nu-1)!}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2 + \frac{l_0}{2}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathbb{E}(|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}|\langle \boldsymbol{w}, \boldsymbol{u}\rangle^2) \\
&\le \left(\frac{C_G}{(\nu-1)!}b^{\nu-1} + 0.5 l_0 m_3 \|\boldsymbol{\delta}\|_2\right).
\end{aligned}$$

Taking the supremum over $\boldsymbol{\beta} \in \Theta^*(r)$, or equivalently $\boldsymbol{\delta} \in \mathbb{B}^p(r)$, yields

$$\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\mathbb{E}\Delta_b(\boldsymbol{\beta})\| \le \left(\frac{C_G}{(\nu-1)!}b^{\nu-1} + 0.5 l_0 m_3 r\right) r \lesssim (b^{\nu-1} + r) r.$$

For the stochastic term $\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2$, following the proof of Theorem 4.2, it can be similarly shown that with probability at least $1 - e^{-t}$,

$$\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\Delta_b(\boldsymbol{\beta}) - \mathbb{E}\Delta_b(\boldsymbol{\beta})\|_2 \lesssim r\sqrt{\frac{p+t}{nb}}$$

as long as $\sqrt{(p+t)/n} \lesssim b$.

Combining the last two displays completes the proof of (D.3). □

D.4 Proof of Theorem D.1

Step 1 (Consistency of the initial estimate). First, note that the consistency of the initial estimator $\bar{\boldsymbol{\beta}}$—namely, $\bar{\boldsymbol{\beta}}$ lies in a local neighborhood of $\boldsymbol{\beta}^*$ with high probability, is a direct consequence of Theorem 4.1. Given a non-negative kernel $K(\cdot)$ and for any $t > 0$, the initial estimator $\bar{\boldsymbol{\beta}}$ satisfies

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le r_1 \asymp \sqrt{\frac{p+t}{n}} \tag{D.8}$$

with probability at least $1 - 2e^{-t}$ as long as $(\frac{p+t}{n})^{1/2} \lesssim h \lesssim (\frac{p+t}{n})^{1/4}$. Let $\mathcal{E}_{\text{init}}(t)$ be the event that (D.8) holds. Provided that the sample Hessian $\nabla^2 \widehat{Q}_b^G(\bar{\boldsymbol{\beta}})$ is invertible, we have

$$\begin{aligned}
\mathbf{J}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) &= -\mathbf{J}\{\nabla^2 \widehat{Q}_b^G(\bar{\boldsymbol{\beta}})\}^{-1} \nabla \widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) \\
&= -\mathbf{J}\{\nabla^2 \widehat{Q}_b^G(\bar{\boldsymbol{\beta}})\}^{-1} \boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} \\
&\quad - \mathbf{J}\{\nabla^2 \widehat{Q}_b^G(\bar{\boldsymbol{\beta}})\}^{-1} \boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{-1/2}\{\mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\},
\end{aligned}$$

or equivalently,

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) &= -\mathbf{H}\widehat{\mathbf{H}}_b^{-1} \cdot \boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} \\
&\quad + (\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1})\mathbf{H} \cdot \boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \mathbf{H}\widehat{\mathbf{H}}_b^{-1} \cdot \boldsymbol{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*),
\end{aligned}$$

where $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2}$ and $\widehat{\mathbf{H}}_b := \boldsymbol{\Sigma}^{-1/2}\nabla^2\widehat{Q}_b^G(\bar{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2}$. It follows that

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \\
&\le \|(\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1})\mathbf{H}\|_2 \cdot \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} + \|\mathbf{I}_p - \mathbf{H}\mathbf{H}_b^{-1}\|_2 \cdot \|\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}} \\
&\quad + \|\mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2 \cdot \|\nabla\widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}}.
\end{aligned} \tag{D.9}$$

In view of (D.9), it remains to bound the following three quantities:

$$\|\mathbf{I}_p - \mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2, \quad \|\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}} \text{ and } \|\nabla\widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*) - \mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{\boldsymbol{\Omega}}.$$

Step 2 (Consistency of the sample Hessian $\nabla^2\widehat{Q}_b^G(\bar{\boldsymbol{\beta}})$). Recall that $\widehat{\mathbf{H}}_b = \boldsymbol{\Sigma}^{-1/2}\nabla^2\widehat{Q}_b^G(\bar{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{H} = \boldsymbol{\Sigma}^{-1/2}\mathbf{J}\boldsymbol{\Sigma}^{-1/2}$. By the triangle inequality,

$$\|\widehat{\mathbf{H}}_b - \mathbf{H}\|_2 \le \|\boldsymbol{\Sigma}^{-1/2}\{\nabla^2\widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla^2 Q_b^G(\boldsymbol{\beta}^*)\}\boldsymbol{\Sigma}^{-1/2}\|_2 + \|\boldsymbol{\Sigma}^{-1/2}\nabla^2 Q_b^G(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}^{-1/2} - \mathbf{H}\|_2.$$

Let the bandwidth b satisfy $\max\{\frac{p\log n + t}{n}, n^{-1/2}\} \lesssim b \lesssim 1$. Conditioned on $\mathcal{E}_{\mathrm{init}}(t)$, applying Propositions D.1 and D.2 with $r = r_1$ yields that, with probability $1 - e^{-t}$,

$$\|\widehat{\mathbf{H}}_b - \mathbf{H}\|_2 \le r_2 \asymp \sqrt{\frac{p\log n + t}{nb}} + b^{\nu-1}.$$

Under Condition 4.2, $0 < \underline{f} \le \lambda_{\min}(\mathbf{H}) \le \lambda_{\max}(\mathbf{H}) \le \bar{f}$, so that $\|\mathbf{H}^{-1}\|_2 \le \underline{f}^{-1}$. For sufficiently large n and small b, this implies $\|\widehat{\mathbf{H}}_b\mathbf{H}^{-1} - \mathbf{I}_p\|_2 \le \underline{f}^{-1}r_2 < 1$, and hence

$$\|\mathbf{H}\widehat{\mathbf{H}}_b^{-1} - \mathbf{I}_p\|_2 \le \frac{r_2}{\underline{f} - r_2}, \quad \|\mathbf{H}\widehat{\mathbf{H}}_b^{-1}\|_2 \le \frac{\underline{f}}{\underline{f} - r_2}. \tag{D.10}$$

Step 3 (Controlling the scores). For $\|\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2$, it follows from Lemma C.2 and Proposition D.1 that with probability at least $1 - e^{-t}$,

$$\|\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \lesssim \sqrt{\frac{p+t}{n}} + b^{\nu}. \tag{D.11}$$

Turning to $\|\boldsymbol{\Sigma}^{-1/2}\{\nabla\widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla\widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$, applying the concentration bound (D.8)

and Proposition D.3 we obtain that, with probability at least $1 - e^{-t}$ conditioned on $\mathcal{E}_{\text{init}}(t)$,

$$\|\boldsymbol{\Sigma}^{-1/2}\{\nabla \widehat{Q}_b^G(\bar{\boldsymbol{\beta}}) - \nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\} - \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim \frac{p+t}{nb^{1/2}} + b^{\nu-1}\sqrt{\frac{p+t}{n}} \tag{D.12}$$

as long as $(\frac{p+t}{n})^{1/2} \lesssim b \lesssim 1$.

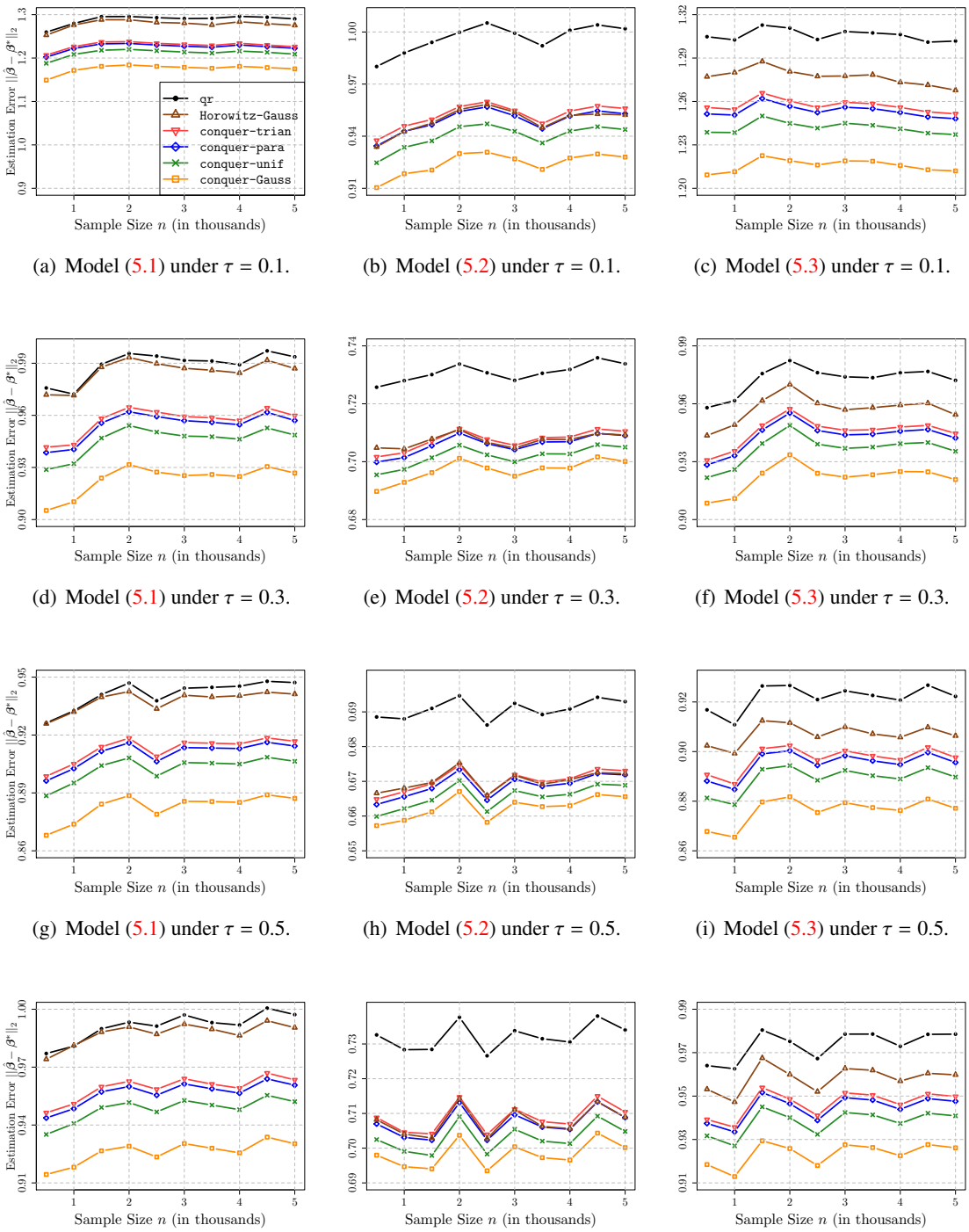
Finally, combining the bounds (D.8)–(D.12), we conclude that with probability at least $1 - 5e^{-t}$,

$$\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla \widehat{Q}_b^G(\boldsymbol{\beta}^*)\|_2 \lesssim \left(\sqrt{\frac{p\log n + t}{nb}} + b^{\nu-1}\right)\left(\sqrt{\frac{p+t}{n}} + b^{\nu}\right),$$

provided that $\max\{\frac{p\log n+t}{n}, (\frac{p+t}{n})^{1/2}\} \lesssim b \lesssim 1$. Under the sample size scaling $n \gtrsim p(\log n)^2 + t$ and bandwidth constraint $b \lesssim (\frac{p+t}{n})^{1/(2\nu)}$, this leads to the claimed bound (D.4). □

E Additional simulation results

In this section, we present additional results for the numerical studies as described in Section 5 in Figures E.1–E.10.



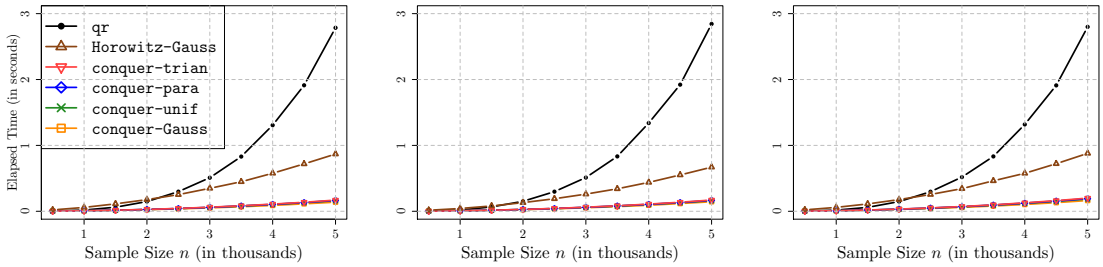
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.2) under $\tau = 0.1$. (c) Model (5.3) under $\tau = 0.1$.

(d) Model (5.1) under $\tau = 0.3$. (e) Model (5.2) under $\tau = 0.3$. (f) Model (5.3) under $\tau = 0.3$.

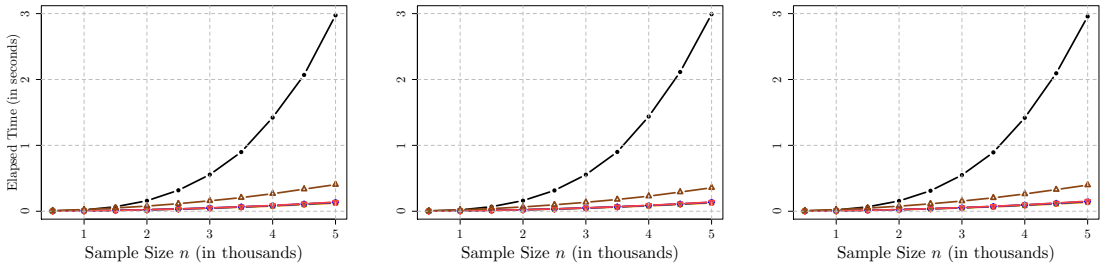
(g) Model (5.1) under $\tau = 0.5$. (h) Model (5.2) under $\tau = 0.5$. (i) Model (5.3) under $\tau = 0.5$.

(j) Model (5.1) under $\tau = 0.7$. (k) Model (5.2) under $\tau = 0.7$. (l) Model (5.3) under $\tau = 0.7$.

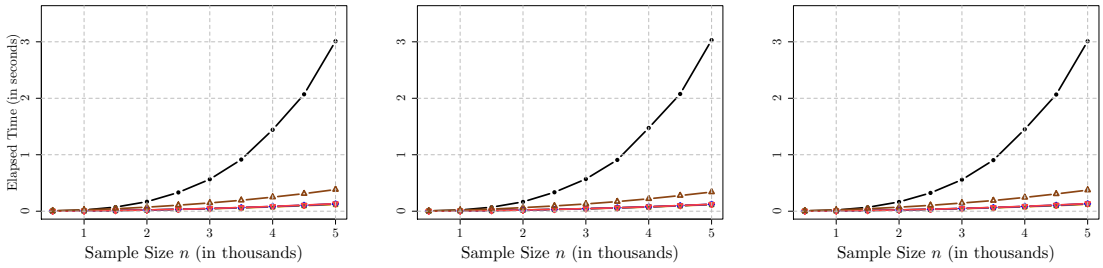
Figure E.1: Results under models (5.1)–(5.3) in Section 5 with $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ and $\mathcal{N}(0, 4)$ noise, averaged over 500 simulations. This figure extends the first row of Figure 3 to other quantile levels.



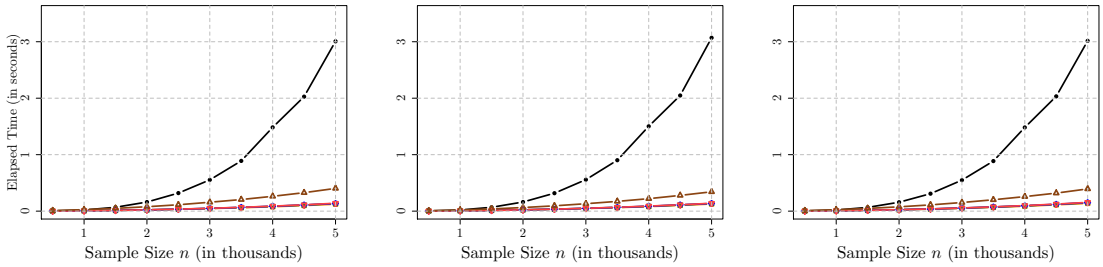
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.2) under $\tau = 0.1$. (c) Model (5.3) under $\tau = 0.1$.



(d) Model (5.1) under $\tau = 0.3$. (e) Model (5.2) under $\tau = 0.3$. (f) Model (5.3) under $\tau = 0.3$.

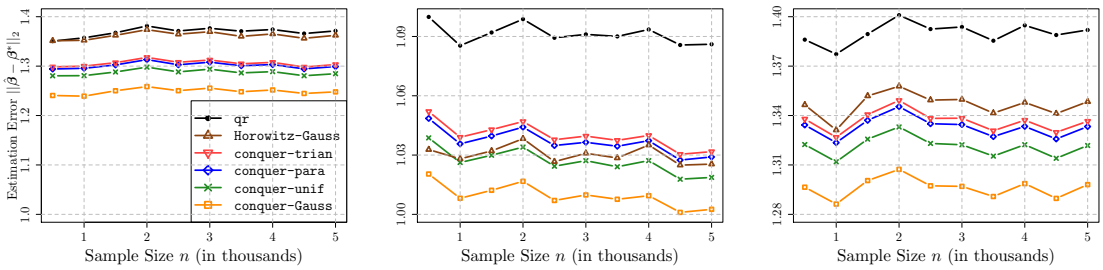


(g) Model (5.1) under $\tau = 0.5$. (h) Model (5.2) under $\tau = 0.5$. (i) Model (5.3) under $\tau = 0.5$.

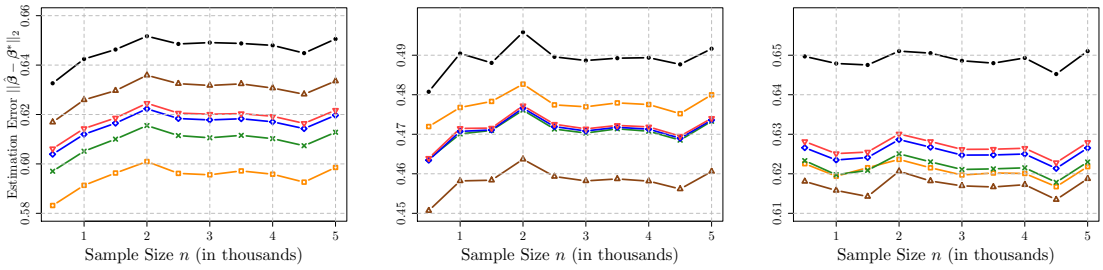


(j) Model (5.1) under $\tau = 0.7$. (k) Model (5.2) under $\tau = 0.7$. (l) Model (5.3) under $\tau = 0.7$.

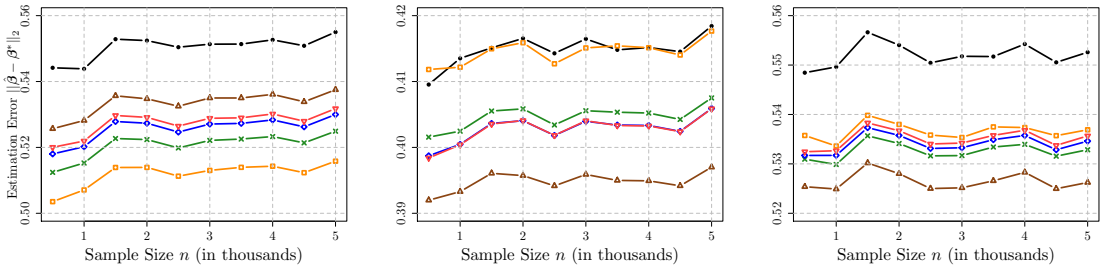
Figure E.2: Elapsed time of standard QR, Horowitz's smoothing and conquer under $\mathcal{N}(0, 4)$ noise setting when $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. This figure extends the first row of Figure 4 to other quantile levels.



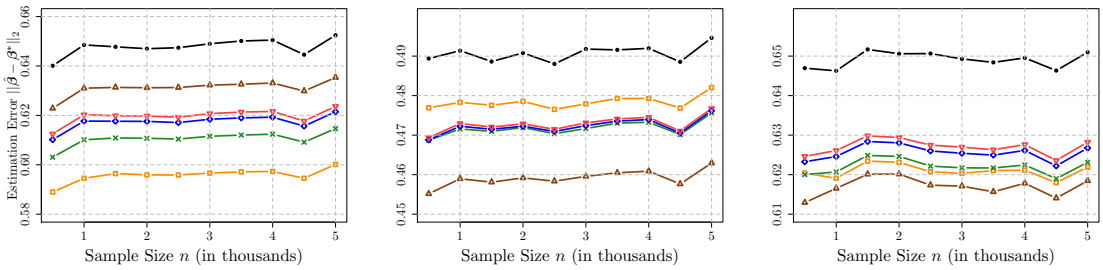
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.2) under $\tau = 0.1$. (c) Model (5.3) under $\tau = 0.1$.



(d) Model (5.1) under $\tau = 0.3$. (e) Model (5.2) under $\tau = 0.3$. (f) Model (5.3) under $\tau = 0.3$.

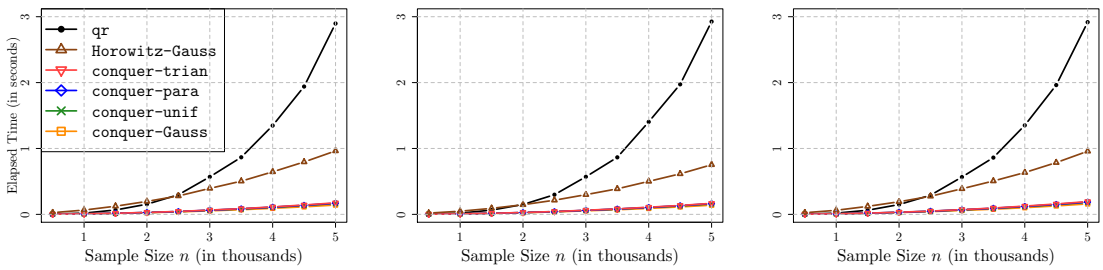


(g) Model (5.1) under $\tau = 0.5$. (h) Model (5.2) under $\tau = 0.5$. (i) Model (5.3) under $\tau = 0.5$.

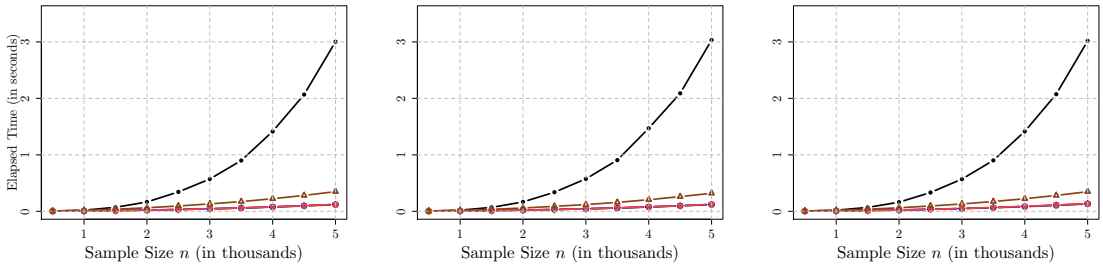


(j) Model (5.1) under $\tau = 0.7$. (k) Model (5.2) under $\tau = 0.7$. (l) Model (5.3) under $\tau = 0.7$.

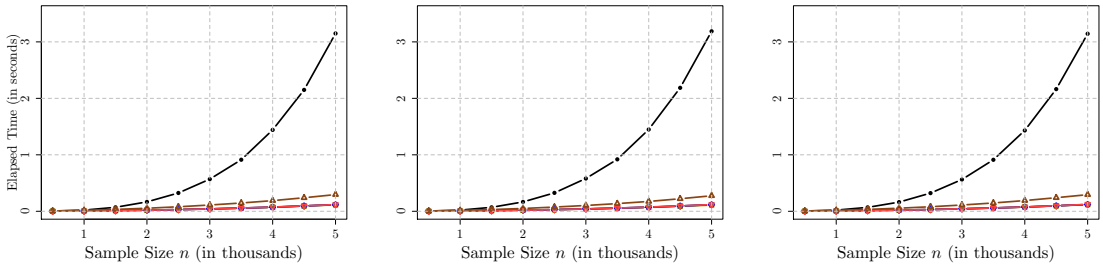
Figure E.3: Results under models (5.1)–(5.3) in Section 5 with $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ and t_2 noise, averaged over 500 simulations. This figure extends the last row of Figure 3 to other quantile levels.



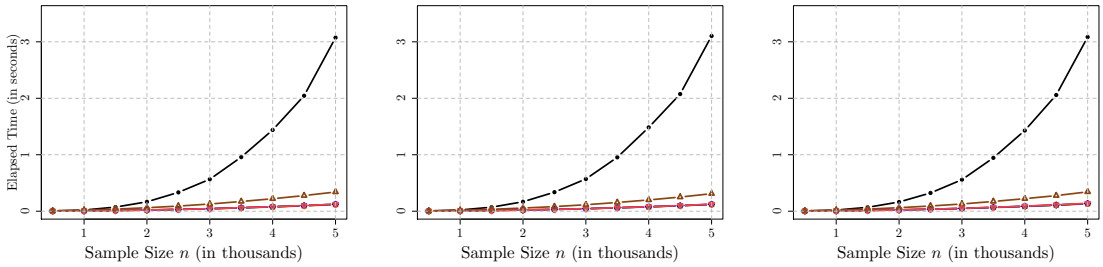
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.2) under $\tau = 0.1$. (c) Model (5.3) under $\tau = 0.1$.



(d) Model (5.1) under $\tau = 0.3$. (e) Model (5.2) under $\tau = 0.3$. (f) Model (5.3) under $\tau = 0.3$.

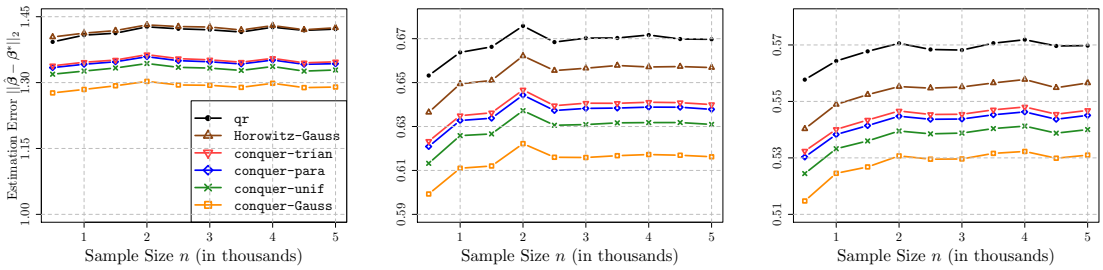


(g) Model (5.1) under $\tau = 0.5$. (h) Model (5.2) under $\tau = 0.5$. (i) Model (5.3) under $\tau = 0.5$.

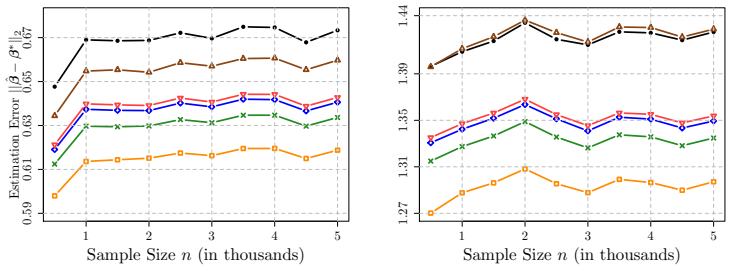


(j) Model (5.1) under $\tau = 0.7$. (k) Model (5.2) under $\tau = 0.7$. (l) Model (5.3) under $\tau = 0.7$.

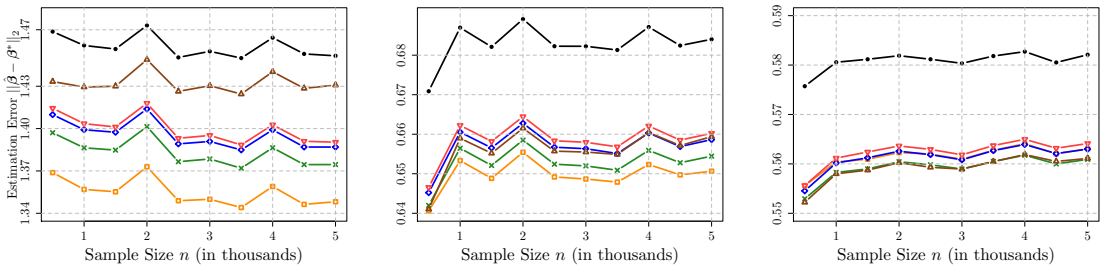
Figure E.4: Elapsed time of standard QR, Horowitz's smoothing and conquer under t_2 noise setting when $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. This figure extends the last row of Figure 4 to other quantile levels.



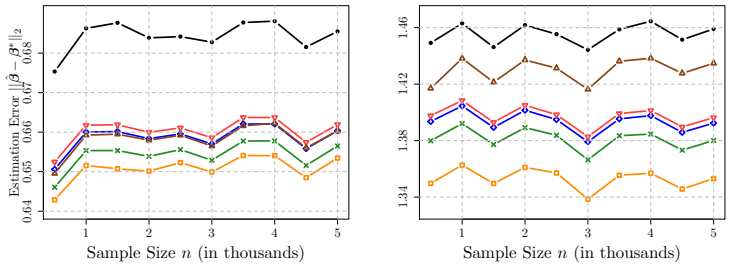
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.1) under $\tau = 0.3$. (c) Model (5.1) under $\tau = 0.5$.



(d) Model (5.1) under $\tau = 0.7$. (e) Model (5.1) under $\tau = 0.9$.

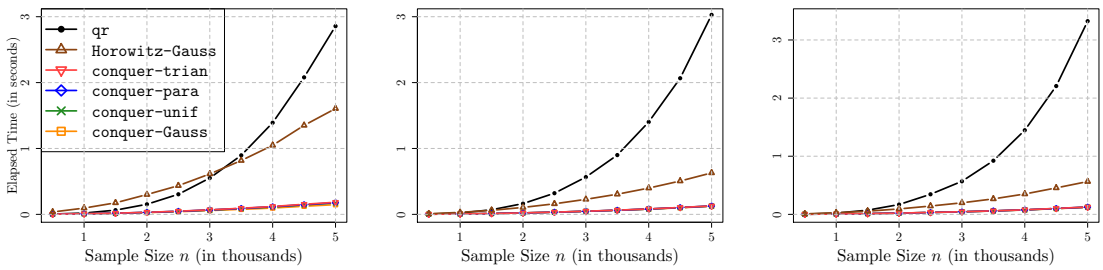


(f) Model (5.3) under $\tau = 0.1$. (g) Model (5.3) under $\tau = 0.3$. (h) Model (5.3) under $\tau = 0.5$.

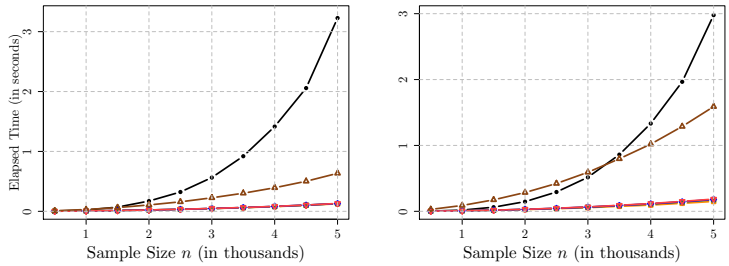


(i) Model (5.3) under $\tau = 0.7$. (j) Model (5.3) under $\tau = 0.9$.

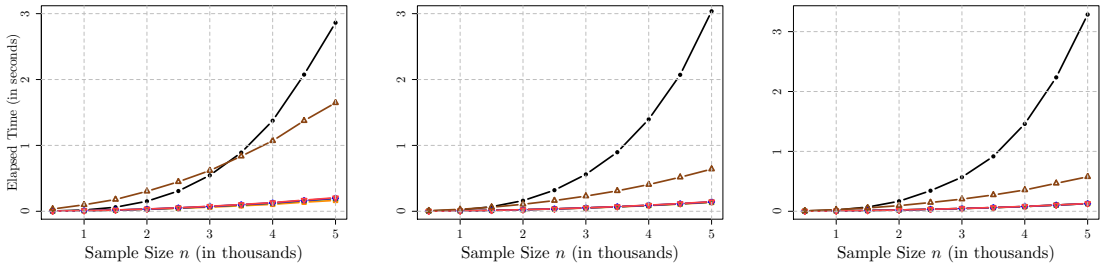
Figure E.5: Estimation errors of standard QR, Horowitz's smoothing and conquer under multivariate normal random design and t_2 noise when $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.



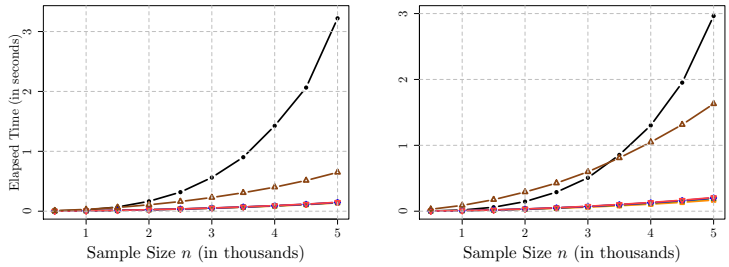
(a) Model (5.1) under $\tau = 0.1$. (b) Model (5.1) under $\tau = 0.3$. (c) Model (5.1) under $\tau = 0.5$.



(d) Model (5.1) under $\tau = 0.7$. (e) Model (5.1) under $\tau = 0.9$.

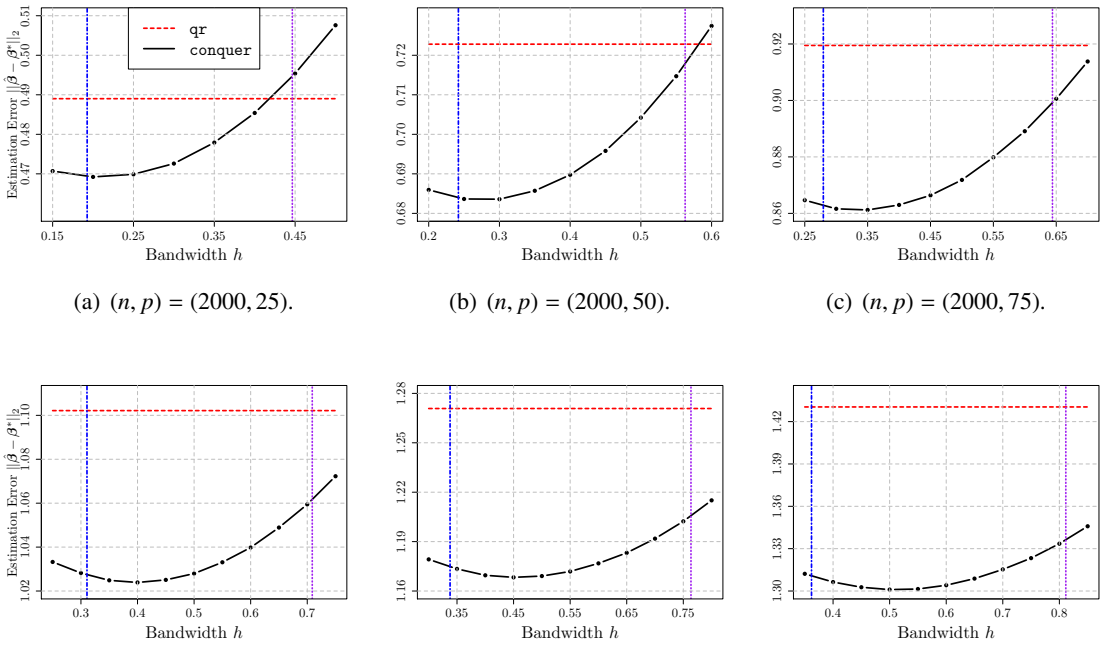


(f) Model (5.3) under $\tau = 0.1$. (g) Model (5.3) under $\tau = 0.3$. (h) Model (5.3) under $\tau = 0.5$.



(i) Model (5.3) under $\tau = 0.7$. (j) Model (5.3) under $\tau = 0.9$.

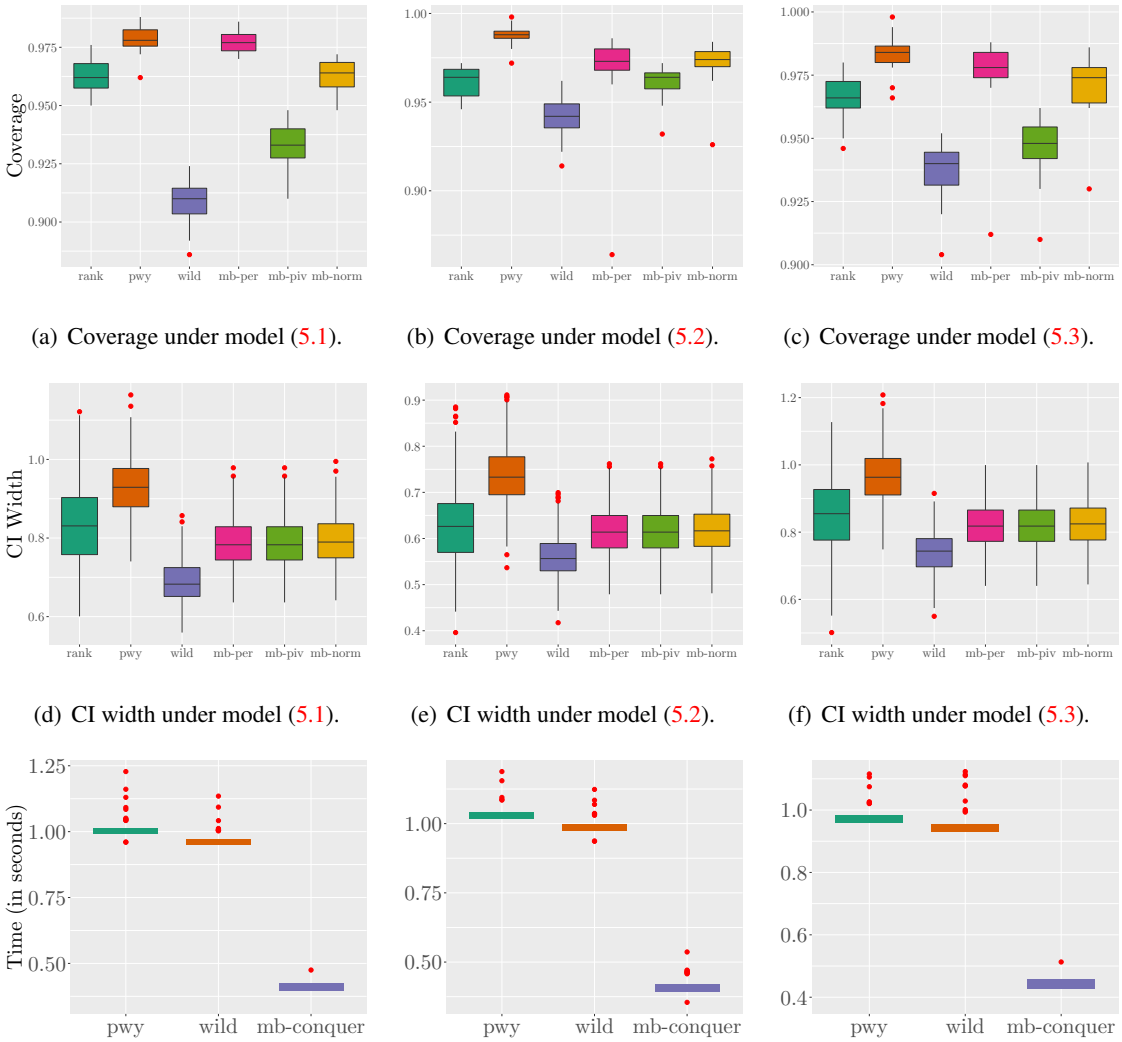
Figure E.6: Elapsed time of standard QR, Horowitz's smoothing and conquer under multivariate normal random design and t_2 noise when $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.



(a) $(n, p) = (2000, 25)$. (b) $(n, p) = (2000, 50)$. (c) $(n, p) = (2000, 75)$.

(d) $(n, p) = (2000, 100)$. (e) $(n, p) = (2000, 125)$. (f) $(n, p) = (2000, 150)$.

Figure E.7: Sensitivity analysis of conquer using a Gaussian kernel under model (5.2) and t_2 error with a range of bandwidth parameter h. Results for $n = 2000$ and $p \in \{25, 50, 75, 100, 125, 150\}$, averaged over 500 datasets. The blue vertical dash line represents our default choice $h_{\text{de}} = \{(p + \log n)/n\}^{2/5}$, the purple vertical dash line refers to the choice h_{AMSE} from (4.12) with some oracle knowledge, and the red horizontal dash line represents the estimation error of standard QR.



(a) Coverage under model (5.1).

(b) Coverage under model (5.2).

(c) Coverage under model (5.3).

(d) CI width under model (5.1).

(e) CI width under model (5.2).

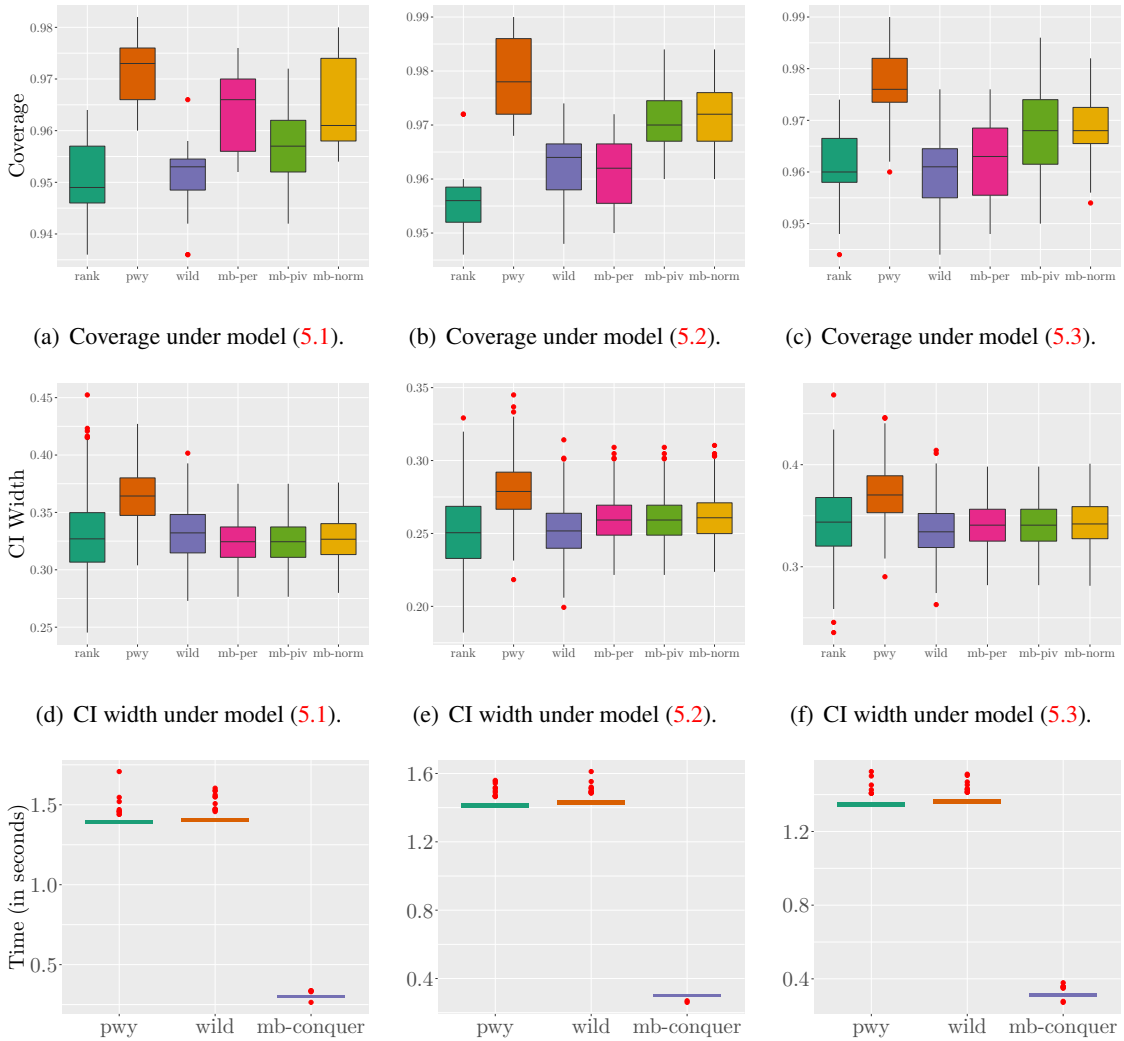
(f) CI width under model (5.3).

(g) Runtime under model (5.1).

(h) Runtime under model (5.2).

(i) Runtime under model (5.3).

Figure E.8: Empirical coverage, confidence interval width, and elapsed time of *six* methods with $\mathcal{N}(0, 4)$ errors under $\tau = 0.9$. Other details are as in Figure 6.



(a) Coverage under model (5.1).

(b) Coverage under model (5.2).

(c) Coverage under model (5.3).

(d) CI width under model (5.1).

(e) CI width under model (5.2).

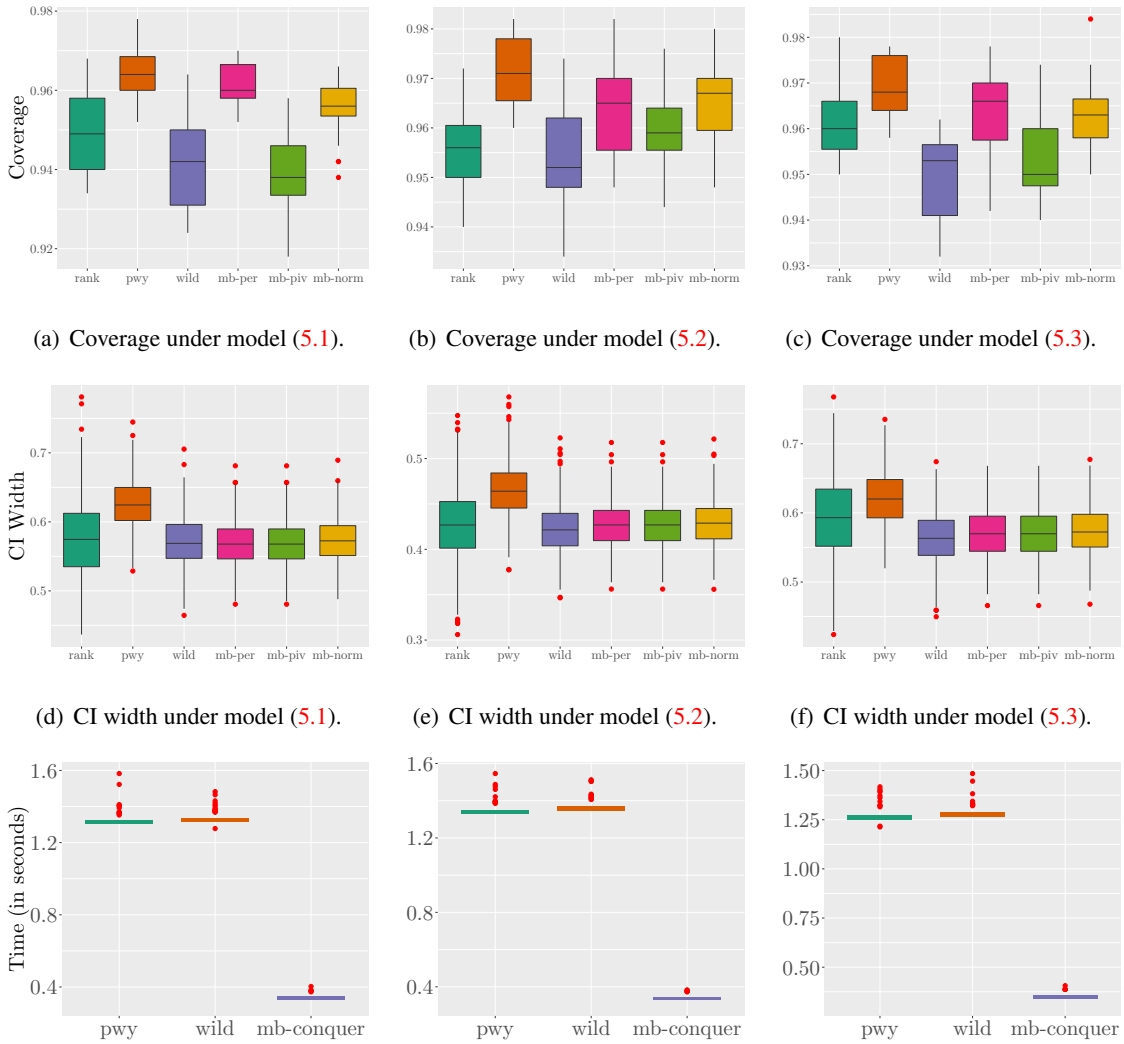
(f) CI width under model (5.3).

(g) Runtime under model (5.1).

(h) Runtime under model (5.2).

(i) Runtime under model (5.3).

Figure E.9: Empirical coverage, confidence interval width, and elapsed time of *six* methods with t_2 errors under $\tau = 0.5$. Other details are as in Figure 6.



(a) Coverage under model (5.1). (b) Coverage under model (5.2). (c) Coverage under model (5.3).

(d) CI width under model (5.1). (e) CI width under model (5.2). (f) CI width under model (5.3).

(g) Runtime under model (5.1). (h) Runtime under model (5.2). (i) Runtime under model (5.3).

Figure E.10: Empirical coverage, confidence interval width, and elapsed time of *six* methods with $\mathcal{N}(0, 4)$ errors under $\tau = 0.5$. Other details are as in Figure 6.

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.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

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

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.

Fernandes, M., Guerre, E. and Horta, E. (2021). Smoothing quantile regressions. *J. Bus. Econ. Statist.* **39** 338–357.

Hestenes, M. R. and Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *J Res. Natl. Bur. Stand.* **49** 409–436.

Hsu, D., Kakade, S. M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17**(52): 1–6.

Kaplan, D. M. and Sun, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econom. Theory* **33** 105–157.

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

Nemirovski, A. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

Oryshchenko, V. (2020). Exact mean integrated squared error and bandwidth selection for kernel distribution function estimators. *Comm. Statist. Theory Methods* **49** 1603–1628.

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

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

Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265.

Tyurin, I. S. (2011). On the convergence rate in Lyapunov's theorem. *Theory Probab. Appl.* **55** 253–270.

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

Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels. *Canad. J. Statist.* **18** 197–204.