Functional Additive Quantile Regression

YINGYING ZHANG[1], HENG LIAN[2], GUODONG LI[3] ZHONGYI ZHU[4]

East China Normal University[1], *City University of Hong Kong*[2],

University of Hong Kong[3], *Fudan University*[4]

Supplementary Material

S1 Proof of Theorem 3.1

We first introduce some notations. Let $J_n = q(K_n + l) + 1$ and

$$\hat{\boldsymbol{W}}_{\mathcal{S}^*} = (\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{1,\mathcal{S}^*}), \ldots, \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{n,\mathcal{S}^*}))^T \in \mathbb{R}^{n \times J_n},$$

$$\hat{\boldsymbol{W}}^2_{B,\mathcal{S}^*} = \hat{\boldsymbol{W}}^T_{\mathcal{S}^*} \boldsymbol{B}_n \hat{\boldsymbol{W}}_{\mathcal{S}^*} \in \mathbb{R}^{J_n \times J_n}, \text{ where } \boldsymbol{B}_n = \text{diag}(f_1(0), \ldots, f_n(0)),$$

$$\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*}) = \hat{\boldsymbol{W}}^{-1}_{B,\mathcal{S}^*} \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*}) \in \mathbb{R}^{J_n},$$

$$\boldsymbol{\delta}_{\mathcal{S}^*} = \hat{\boldsymbol{W}}_{B,\mathcal{S}^*}(\boldsymbol{\theta}_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) \in \mathbb{R}^{J_n},$$

$$R_i = (\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*}) - \boldsymbol{W}(\boldsymbol{\zeta}_{i,\mathcal{S}^*}))^T \boldsymbol{\theta}^0_{\mathcal{S}^*},$$

$$u_i = \boldsymbol{W}(\boldsymbol{\zeta}_{i,\mathcal{S}^*})^T \boldsymbol{\theta}^0_{\mathcal{S}^*} - \alpha(\tau) - \sum_{j=1}^{q} f_{s_j,\tau}(\zeta_{i,s_j}).$$

Define the oracle minimizer of $\boldsymbol{\delta}_{\mathcal{S}^*}$ as

$$\hat{\boldsymbol{\delta}}_{\mathcal{S}^*} = \arg\min_{\boldsymbol{\delta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T \boldsymbol{\delta} - R_i - u_i).$$

First we derive some technical lemmas used in the proof.

Lemma S1.1. *We have the following properties for the spline basis vector:*

(1) $E(\|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2) \leq b_1$*, for some positive constant* b_1 *for all n sufficiently large.*

(2) $b_2 K_n^{-1} \leq E(\lambda_{min}(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T)) \leq E(\lambda_{max}(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T)) \leq b_2^* K_n^{-1}$*, for some positive constants* b_2 *and* b_2^* *for* n *sufficiently large.*

(3) $E(\|\hat{\boldsymbol{W}}_{B,\mathcal{S}^*}^{-1}\|) \geq b_3\sqrt{K_n/n}$*, for some positive* b_3 *for all* n *sufficiently large.*

(4) $\max_i \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 = O_p(\sqrt{\frac{K_n}{n}})$.

Proof.

(1) The result follows if we can show $E(B_m^2(\hat{\zeta}_{i,s_j})) = O_p(\frac{1}{K_n})$ for all $1 \leq m \leq K_n + l$. It holds that $E(B_m^2(\zeta_{i,s_j})) = O_p(\frac{1}{K_n})$ by Lemma 2(1) in Sherwood and Wang (2016). Note that $E(B_m^2(\hat{\zeta}_{i,s_j})) = E(B_m(\hat{\zeta}_{i,s_j}) - B_m(\zeta_{i,s_j}) + B_m(\zeta_{i,s_j}))^2 = E(B_m^{(1)}(\zeta_{i,s_j}^*)(\hat{\zeta}_{i,s_j} - \zeta_{i,s_j}) + B_m(\zeta_{i,s_j}))^2$. By (S.3) in the supplement of Wong et al. (2018), we have $(\hat{\zeta}_{i,s_j} - \zeta_{i,s_j})^2 = O_p(\frac{s_j^2}{n})$, thus

For a matrix A, $\|A\| = \sqrt{\lambda_{max}(A^T A)}$ denotes the spectral norm.

$E(B_m^{(1)}(\zeta^*_{i,s_j})(\hat{\zeta}_{i,s_j} - \zeta_{i,s_j}))^2 = O_p(\frac{K_n s_j^2}{n})$ where $(B_m^{(1)}(\zeta^*_{i,s_j}))^2 = O_p(K_n)$. Note that $\frac{K_n s^2}{n} < \frac{1}{K_n}$. Thus The dominant term is $O_p(1/K_n)$.

(2) By the proof of Lemma 2(2) in Sherwood and Wang (2016), we can see that this result follows if we can prove $E(\boldsymbol{a}_{s_j}^T \boldsymbol{w}(\hat{\zeta}_{i,s_j}))^2 \geq c_{s_j}\|\boldsymbol{a}_{s_j}\|_2^2 K_n^{-1}$ for some constant c_{s_j} and any $(K_n + l)$-dimensional vector $\boldsymbol{a}_{s_j}$ when n is sufficiently large. It holds that $E(\boldsymbol{a}_{s_j}^T \boldsymbol{w}(\zeta_{i,s_j}))^2 \geq c_{s_j}\|\boldsymbol{a}_{s_j}\|_2^2 K_n^{-1}$. Note that $E(\boldsymbol{a}_{s_j}^T \boldsymbol{w}(\hat{\zeta}_{i,s_j}))^2 = E(\boldsymbol{a}_{s_j}^T \boldsymbol{w}(\zeta_{i,s_j}) + \boldsymbol{a}_{s_j}^T(\boldsymbol{w}(\hat{\zeta}_{i,s_j}) - \boldsymbol{w}(\zeta_{i,s_j})))^2$ where the second term is $O_p(\frac{K_n^2 s^2}{n})$ and dominated by $O_p(1/K_n)$.

(3) Similar to Lemma2 (3) in Sherwood and Wang (2016), we can show that $E(\lambda_{\min}(\hat{\boldsymbol{W}}^2_{B,\mathcal{S}^*})) \geq c' n/K_n$ for some positive c' from arguments in (2). The proof finishes by $\|\hat{\boldsymbol{W}}^{-1}_{B,\mathcal{S}^*}\| = \lambda_{\min}^{-1/2}(\hat{\boldsymbol{W}}^2_{B,\mathcal{S}^*})$.

(4) This is the same with Sherwood and Wang (2016) Lemma2 (4) which can be proved as Lemma 5.1 in Shi and Li (1995).

In the proofs C denotes a generic positive constant which may assume different values even on the same line.

Lemma S1.2. *Under conditions (C1)-(C3), we have* $\|\hat{\boldsymbol{\delta}}_{\mathcal{S}^*}\|_2 = O_p(K_n^{1/2} + s + K_n^{-r} n^{1/2})$.

Proof. We will prove that for $\forall \eta > 0$, there exits an $L > 0$ such that

$$P(\inf_{\|\boldsymbol{\delta}\|_2 = L} d_n^{-2} \sum_{i=1}^{n} (Q_i(d_n \boldsymbol{\delta}) - Q_i(0)) > 0) \geq 1 - \eta, \tag{S1.1}$$

where $Q_i(\boldsymbol{\delta}) = \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\delta} - R_i - u_i)$ and $d_n = K_n^{1/2} + s + K_n^{-r}n^{1/2}$. Then the convexity implies $\|\hat{\boldsymbol{\delta}}_{\mathcal{S}^*}\|_2 = O_p(K_n^{1/2} + s + K_n^{-r}n^{1/2})$. Note that

$$
\begin{aligned}
& d_n^{-2}\sum_{i=1}^{n}(Q_i(d_n\boldsymbol{\delta}) - Q_i(0)) \\
&= d_n^{-2}\sum_{i=1}^{n} D_i(d_n\boldsymbol{\delta}) + d_n^{-2}\sum_{i=1}^{n} E[Q_i(d_n\boldsymbol{\delta}) - Q_i(0)|X_i] - d_n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\delta}\psi_\tau(\epsilon_i) \\
&= G_1 + G_2 + G_3,
\end{aligned}
$$

where $D_i(\boldsymbol{\delta}) = Q_i(\boldsymbol{\delta}) - Q_i(0) - E[Q_i(\boldsymbol{\delta}) - Q_i(0)|X_i] + \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\delta}\psi_\tau(\epsilon_i)$ and $\psi_\tau(u) = \tau - I(u < 0)$. Next we will prove (S1.1) by three steps. In the first step, we will prove that $\sup_{\|\boldsymbol{\delta}\|_2 \le L}|G_1| = o_p(1)$. In the second step, we will show that asymptotically G_2 has a positive lower bound CL^2 when L is sufficiently large. In the third step, we obtain $G_3 = O_p(\|\boldsymbol{\delta}\|_2)$. This completes the proof.

Step 1. In this step, we prove that $\forall \varepsilon > 0$,

$$
P(d_n^{-2}\sup_{\|\boldsymbol{\delta}\|_2 \le L}|\sum_{i=1}^{n} D_i(d_n\boldsymbol{\delta})| > \varepsilon) \to 0.
$$

Let F_{n1} denote the event $\max_i\|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 \le \alpha_1\sqrt{\frac{J_n}{n}}$ for some positive α_1. Lemma S1.1(4) implies that $P(F_{n1}) \to 1$ as $n \to \infty$. Let F_{n2} denote the event $\max_i|u_i| \le \alpha_2 K_n^{-r}$ for some positive α_2. Then $P(F_{n2}) \to 1$ follows from Schumaker (1981). Let F_{n3} denote the event $\frac{1}{n}\sum_{i=1}^{n}|R_i| \le \alpha_3 s/\sqrt{n}$ for some positive α_3. In the following we will show that $P(F_{n3}) \to 1$.

Folllowing the calculation

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|R_i| &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{q}|(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,k_t})-\boldsymbol{W}(\boldsymbol{\zeta}_{i,k_t}))^T\boldsymbol{\theta}_{k_t}^0| \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{q}|\boldsymbol{W}^{(1)}(\boldsymbol{\zeta}_{i,k_t})^T\boldsymbol{\theta}_{k_t}^0(\hat{\boldsymbol{\zeta}}_{i,k_t}-\boldsymbol{\zeta}_{i,k_t})| \\ &\leq \Big\{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{q}(\boldsymbol{W}^{(1)}(\boldsymbol{\zeta}_{i,k_t})^T\boldsymbol{\theta}_{k_t}^0)^2\Big\}^{1/2}\Big\{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{q}(\hat{\boldsymbol{\zeta}}_{i,k_t}-\boldsymbol{\zeta}_{i,k_t})^2\Big\}^{1/2}\end{aligned}$$

By Lemma 11 in Stone (1985), we have $|\boldsymbol{W}^{(1)}(\boldsymbol{\zeta}_{i,k_t})^T\boldsymbol{\theta}_{k_t}^0| \leq C\int_0^1(\boldsymbol{W}(t)^T\boldsymbol{\theta}_{k_t}^0)^2dt = C\int_0^1(f_{k_t}(t)+K_n^{-r})^2dt = O(1)$. By Lemma 3.1, we have $E(\hat{\zeta}_{ik}-\zeta_{ik})^2 \leq Ck^2/n$ uniformly for $k \leq s$. So $P(F_{n3}) \to 1$.

Then it's sufficient to show

$$P(d_n^{-2}\sup_{\|\boldsymbol{\delta}\|_2\leq L}|\sum_{i=1}^{n}D_i(d_n\boldsymbol{\delta})| > \varepsilon, F_{n1}\cap F_{n2}\cap F_{n3}) \to 0.$$

Define $\Delta = \{\boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_2 \leq L, \boldsymbol{\delta}\in\mathbb{R}^{J_n}\}$. We can partition Δ as a union of disjoint regions $\Delta_1,\ldots,\Delta_{M_n}$, such that the diameter of each region does not exceed $m_0 = \frac{\varepsilon}{4\alpha_1 J_n^{1/2}n^{1/2}d_n^{-1}}$. This covering can be constructed such that $M_n \leq C(\frac{CJ_n^{1/2}n^{1/2}d_n^{-1}}{\varepsilon})^{J_n}$, where C is a positive constant. Let $\boldsymbol{\delta}_1^\star,\ldots,\boldsymbol{\delta}_{M_n}^\star$ be arbitrary points in $\Delta_1, \ldots, \Delta_{M_n}$ respectively. Then

$$\begin{aligned} & P(\sup_{\|\boldsymbol{\delta}\|_2 \le L} d_n^{-2} |\sum_{i=1}^n D_i(d_n \boldsymbol{\delta})| > \varepsilon, F_{n1} \cap F_{n2} \cap F_{n3}) \\ \le & \sum_{m=1}^{M_n} P(\sup_{\boldsymbol{\delta} \in \Delta_m} d_n^{-2} |\sum_{i=1}^n D_i(d_n \boldsymbol{\delta})| > \varepsilon, F_{n1} \cap F_{n2} \cap F_{n3}) \\ \le & \sum_{m=1}^{M_n} P(|\sum_{i=1}^n D_i(d_n \boldsymbol{\delta}_m^\star)| + \sup_{\boldsymbol{\delta} \in \Delta_m} |\sum_{i=1}^n (D_i(d_n \boldsymbol{\delta}) - D_i(d_n \boldsymbol{\delta}_m^\star))| > d_n^2 \varepsilon, F_{n1} \cap F_{n2} \cap F_{n3}). \end{aligned}$$

We first show that $\sup_{\boldsymbol{\delta} \in \Delta_m} |\sum_{i=1}^n (D_i(d_n \boldsymbol{\delta}) - D_i(d_n \boldsymbol{\delta}_m^\star))| I(F_{n1} \cap F_{n2} \cap F_{n3}) < d_n^2 \varepsilon / 2$. Noting that $\rho_\tau(u) = \frac{1}{2}|u| + (\tau - \frac{1}{2})u$, we have $Q_i(\boldsymbol{\delta}) - Q_i(0) = \frac{1}{2}[|\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T \boldsymbol{\delta} - R_i - u_i| - |\epsilon_i - R_i - u_i|] - (\tau - \frac{1}{2}) \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T \boldsymbol{\delta}$. So

$$\begin{aligned} & \sup_{\boldsymbol{\delta} \in \Delta_m} |\sum_{i=1}^n D_i(d_n \boldsymbol{\delta}) - D_i(d_n \boldsymbol{\delta}_m^\star)| I(F_{n1} \cap F_{n2} \cap F_{n3}) \\ \le & \ 2 n d_n \max_i \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 \sup_{\boldsymbol{\delta} \in \Delta_m} \|\boldsymbol{\delta} - \boldsymbol{\delta}_m^\star\|_2 I(F_{n1} \cap F_{n2} \cap F_{n3}) \\ \le & \ d_n^2 \varepsilon / 2. \end{aligned}$$

The proof is complete if we can verify

$$\sum_{m=1}^{M_n} P(|\sum_{i=1}^n D_i(d_n \boldsymbol{\delta}_m^\star)| > d_n^2 \varepsilon / 2, F_{n1} \cap F_{n2} \cap F_{n3}) \to 0.$$

First applying the definition of D_i and the triangle inequality,

$$\begin{aligned}
& \max_i |D_i(d_n \boldsymbol{\delta}_m^\star)| I(F_{n1} \cap F_{n2} \cap F_{n3}) \\
\leq \;& 2 \max_i \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 d_n \boldsymbol{\delta}_m^\star I(F_{n1} \cap F_{n2} \cap F_{n3}) \\
\leq \;& C d_n J_n^{1/2} n^{-1/2},
\end{aligned}$$

for some positive C. Next,

$$\sum_{i=1}^{n} Var[D_i(d_n \boldsymbol{\delta}_m^\star) I(F_{n1} \cap F_{n2} \cap F_{n3})|X_i] \leq \sum_{i=1}^{n} E[V_i^2(d_n \boldsymbol{\delta}_m^\star) I(F_{n1} \cap F_{n2} \cap F_{n3})|X_i],$$

where $V_i(\boldsymbol{\delta}) = Q_i(\boldsymbol{\delta}) - Q_i(0) + \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T \boldsymbol{\delta} \psi_\tau(\epsilon_i)$ and $D_i(\boldsymbol{\delta}) = V_i(\boldsymbol{\delta}) - E[V_i(\boldsymbol{\delta})|X_i]$ by definition. By Knight's identity,

$$\begin{aligned}
V_i(d_n \boldsymbol{\delta}_m^\star) &= \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n \boldsymbol{\delta}_m^\star [I(\epsilon_i - R_i - u_i < 0) - I(\epsilon_i < 0)] \\
&+ \int_0^{\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n \boldsymbol{\delta}_m^\star} [I(\epsilon_i - R_i - u_i < s) - I(\epsilon_i - R_i - u_i < 0)] \\
&= V_{i1} + V_{i2}.
\end{aligned}$$

We have

$$\begin{aligned}
& \sum_{i=1}^{n} E[V_{i1}^2 I(F_{n1}\cap F_{n2}\cap F_{n3})|X_i] \\
= & \sum_{i=1}^{n} E[(\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n \boldsymbol{\delta}_m^{\star})^2 |I(\epsilon_i - R_i - u_i < 0) - I(\epsilon_i < 0)| I(F_{n1}\cap F_{n2}\cap F_{n3})|X_i] \\
\leq & C\frac{J_n}{n} d_n^2 \sum_{i=1}^{n} E[I(0 \leq |\epsilon_i| \leq |R_i + u_i|) I(F_{n1}\cap F_{n2}\cap F_{n3})|X_i] \\
= & C\frac{J_n}{n} d_n^2 \sum_{i=1}^{n} \int_{-|R_i+u_i|}^{|R_i+u_i|} f_i(s) ds \\
\leq & C\frac{J_n}{n} d_n^2 \sum_{i=1}^{n} |R_i + u_i| \\
\leq & C n^{-1/2} J_n d_n^2 (s + K_n^{-r}\sqrt{n}),
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
& \sum_{i=1}^{n} E[V_{i2}^2 I(F_{n1}\cap F_{n2}\cap F_{n3})|X_i] \\
\leq & C d_n J_n^{1/2} n^{-1/2} \sum_{i=1}^{n} \int_0^{\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n \boldsymbol{\delta}_m^{\star}} [F_i(R_i + u_i + s) - F_i(R_i + u_i)] I(F_{n1}\cap F_{n2}\cap F_{n3}) ds \\
\leq & C d_n^3 J_n^{1/2} n^{-1/2} [\boldsymbol{\delta}_m^{\star T} \sum_{i=1}^{n} f_i(0) \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*}) \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T \boldsymbol{\delta}_m^{\star}](1 + o(1)) \\
\leq & C d_n^3 J_n^{1/2} n^{-1/2}.
\end{aligned}$$

The last inequality follows since $\sum_{i=1}^{n} f_i(0)\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T = \hat{\boldsymbol{W}}_B^{-1}\hat{\boldsymbol{W}} B \hat{\boldsymbol{W}}^T \hat{\boldsymbol{W}}_B^{-1} =$

I. Therefore,

$$\sum_{i=1}^{n} Var[D_i(d_n\boldsymbol{\delta}_m^\star)I(F_{n1}\cap F_{n2}\cap F_{n3})|X_i] \leq Cn^{-1/2}J_n d_n^2(s+K_n^{-r}\sqrt{n}).$$

By Bernstein's inequality,

$$\begin{aligned}
&\sum_{m=1}^{M_n} P(|\sum_{i=1}^{n} D_i(d_n\boldsymbol{\delta}_m^\star)| > d_n^2\varepsilon/2, F_{n1}\cap F_{n2}\cap F_{n3})\\
\leq\ & 2\sum_{m=1}^{M_n}\exp(\frac{-d_n^4\varepsilon^2/4}{Cn^{-1/2}J_n d_n^2(s+K_n^{-r}\sqrt{n})+Cd_n^3J_n^{1/2}n^{-1/2}\varepsilon/2})\\
\leq\ & 2\sum_{m=1}^{M_n}\exp(\frac{-Cd_n^2n^{1/2}}{J_n(s+K_n^{-r}\sqrt{n})})\\
\leq\ & C\exp(CJ_n\log n-\frac{Cd_n^2n^{1/2}}{J_n(s+K_n^{-r}\sqrt{n})}),
\end{aligned}$$

which converges to zero as $\max\{K_n, s^2, K_n^{-2r}n\} \gg K_n^2\{\frac{s}{\sqrt{n}} + K_n^{-r}\}\log n$. Hence the proof of the first step is complete.

Step 2. In this step, we show that asymptotically $G_2 = d_n^{-2}\sum_{i=1}^{n} E[Q_i(d_n\boldsymbol{\delta}) - Q_i(0)|X_i]$ has a positive lower bound CL^2 when L is sufficiently large. By

Knight's identity,

$$
\begin{aligned}
G_2 &= d_n^{-2}\sum_{i=1}^n E[\int_{R_i+u_i}^{\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n\boldsymbol{\delta}+R_i+u_i}(I(\epsilon_i<s)-I(\epsilon_i<0))ds|X_i] \\
&= d_n^{-2}\sum_{i=1}^n \int_{R_i+u_i}^{\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n\boldsymbol{\delta}+R_i+u_i} f_i(0)sds(1+o(1)) \\
&= d_n^{-2}\sum_{i=1}^n f_i(0)\frac{1}{2}\{(\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n\boldsymbol{\delta})^2+2(\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T d_n\boldsymbol{\delta})(R_i+u_i)\} \\
&= C\|\boldsymbol{\delta}\|_2^2 + Cd_n^{-1}\boldsymbol{\delta}^T\hat{\boldsymbol{W}}_B^{-1}\hat{\boldsymbol{W}}\boldsymbol{B}_n(\boldsymbol{R}_n+\boldsymbol{u}_n) \\
&= C\|\boldsymbol{\delta}\|_2^2 + Cd_n^{-1}\boldsymbol{\delta}^T(\boldsymbol{R}_n+\boldsymbol{u}_n),
\end{aligned}
$$

where $\boldsymbol{R}_n = (R_1,\ldots,R_n)^T$ and $\boldsymbol{u}_n = (u_1,\ldots,u_n)^T$. The second last equality follows from $\sum_{i=1}^n f_i(0)\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T = \hat{\boldsymbol{W}}_B^{-1}\hat{\boldsymbol{W}}B\hat{\boldsymbol{W}}^T\hat{\boldsymbol{W}}_B^{-1} = I$. Note that $\|\boldsymbol{u}_n\|_2 = O_p(\sqrt{n}K_n^{-r})$ and $\|\boldsymbol{R}_n\|_2 = \sqrt{\sum_{i=1}^n |R_i|^2} = O_p(s)$by technical arguments similar with the proof of $P(F_{n3}) \to 1$ in Step 1. Thus $|Cd_n^{-1}\boldsymbol{\delta}^T(\boldsymbol{R}_n+\boldsymbol{u}_n)| = O_p(\|\boldsymbol{\delta}\|_2)$, and when L is sufficiently large, the quadratic term will dominant. This completes the proof of Step 2.

Step 3. In this step, we evaluate $G_3 = -d_n^{-1}\sum_{i=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\delta}\psi_\tau(\epsilon_i)$ as Lemma 3.3 in He and Shi (1994). At almost all samples $T = \{X_1, X_2, \cdots,\}$

and for any real number $M > 0$, Chebychev inequality implies

$$\begin{aligned}
& P\big\{d_n^{-1}\|\sum_{i=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})(\tau - I(\epsilon_i < 0))\|_2 > M|T\big\} \\
\leq \quad & E[\|\sum_{i=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})(\tau - I(\epsilon_i < 0))\|_2^2]/(d_n^2 M^2) \\
= \quad & E[\text{trace}(\sum_{i=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})(\tau - I(\epsilon_i < 0)) \sum_{j=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{j,\mathcal{S}^*})^T(\tau - I(\epsilon_j < 0)))]/(d_n^2 M^2) \\
\leq \quad & \frac{\tau(1-\tau)K_n}{M^2 d_n^2}, \qquad \text{(S1.2)}
\end{aligned}$$

where the last equality follows from Lemma S1.1(4) and the fact that $E[(\tau - I(\epsilon_i < 0))(\tau - I(\epsilon_j < 0))] = 0$ for $i \neq j$. So we have $G_3 = O_p(\|\delta\|_2)$.

Proof of Theorem 3.1. From Lemma S1.2, we have

$$\|\hat{\boldsymbol{\delta}}_{\mathcal{S}^*}\|_2 = O_p(K_n^{1/2} + s + K_n^{-r} n^{1/2}).$$

That is, we have $\|\hat{\boldsymbol{W}}_B(\boldsymbol{\theta}^*_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*})\|_2 = O_p(K_n^{1/2} + s + K_n^{-r} n^{1/2})$. In the proof of Lemma S1.1(3), $\lambda_{\min}(\hat{W}_B^2) = O_p(n/K_n)$. So

$$\|\boldsymbol{\theta}^*_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}\|_2 = O_p(\frac{K_n}{\sqrt{n}} + \sqrt{\frac{K_n}{n}} s + K_n^{-r+1/2}).$$

For the second argument, note that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n} f_i(0)(g^*(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*}) - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}))^2 \\
&= n^{-1}\sum_{i=1}^{n} f_i(0)(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T(\boldsymbol{\theta}^*_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) - R_i - u_i)^2 \\
&\le n^{-1}C(\boldsymbol{\theta}^*_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*})^T\hat{\boldsymbol{W}}_B^2(\boldsymbol{\theta}^*_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) + O_p(\frac{s^2}{n}) + O_p(K_n^{-2r}) \\
&= O_p(\frac{K_n}{n} + \frac{s^2}{n} + K_n^{-2r}).
\end{aligned}$$

S2 Proof of Theorem 3.2

Note that the SCAD penalized objective function can be written as $S_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}) - H_n(\boldsymbol{\theta})$, where $G_n(\boldsymbol{\theta})$ and $H_n(\boldsymbol{\theta})$ are convex functions,

$$G_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}) + \sum_{k=1}^{s}\lambda\|\boldsymbol{\theta}_k\|_1,$$

and

$$H_n(\boldsymbol{\theta}) = \sum_{k=1}^{s}\Big\{\frac{\|\boldsymbol{\theta}_k\|_1^2 - 2\lambda\|\boldsymbol{\theta}_k\|_1 + \lambda^2}{2(a-1)}I(\lambda \le \|\boldsymbol{\theta}_k\|_1 \le a\lambda) + [\lambda\|\boldsymbol{\theta}_k\|_1 - (a+1)\lambda^2/2]I(\|\boldsymbol{\theta}_k\|_1 > a\lambda)\Big\}.$$

Here neither $G_n(\boldsymbol{\theta})$ nor $H_n(\boldsymbol{\theta})$ are differentiable, while H_n in Sherwood and Wang (2016) is differentiable everywhere. We formally define the subdifferentials of $G_n(\boldsymbol{\theta})$ and $H_n(\boldsymbol{\theta})$.

$$
\begin{aligned}
\frac{\partial G_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \{\boldsymbol{\pi} &= (\pi_0, \boldsymbol{\pi}_1^T, \ldots, \boldsymbol{\pi}_s^T)^T \in \mathbb{R}^{s(K_n+l)+1} : \\
\pi_0 &= -\tau n^{-1} \sum_{i=1}^n K_n^{-1/2} I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta} > 0) \\
&\quad + (1-\tau) n^{-1} \sum_{i=1}^n K_n^{-1/2} I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta} < 0) \\
&\quad - n^{-1} \sum_{i=1}^n K_n^{-1/2} a_i \equiv \nu_0(\boldsymbol{\theta}); \\
\boldsymbol{\pi}_k &= -\tau n^{-1} \sum_{i=1}^n \boldsymbol{w}(\hat{\boldsymbol{\zeta}}_{ik}) I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta} > 0) \\
&\quad + (1-\tau) n^{-1} \sum_{i=1}^n \boldsymbol{w}(\hat{\boldsymbol{\zeta}}_{ik}) I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta} < 0) \\
&\quad - n^{-1} \sum_{i=1}^n \boldsymbol{w}(\hat{\boldsymbol{\zeta}}_{ik}) a_i + \lambda \boldsymbol{l}_k \equiv \boldsymbol{\nu}_k(\boldsymbol{\theta}) + \lambda \boldsymbol{l}_k,\ for\ 1 \le k \le s\},
\end{aligned}
$$

where $a_i = 0$ if $y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta} \neq 0$ and $a_i \in [\tau - 1, \tau]$ otherwise; $\boldsymbol{l}_k = (l_{k1}, \ldots, l_{k,K_n+l})^T \in \mathbb{R}^{K_n+l}$ and $l_{km} = sgn(\theta_{km})$ if $\theta_{km} \neq 0$ and $l_{km} \in [-1, 1]$ otherwise for $1 \le m \le K_n + l$.

$$
\begin{aligned}
\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \{\boldsymbol{\varpi} &= (0, \boldsymbol{\varpi}_1^T, \ldots, \boldsymbol{\varpi}_s^T)^T \in \mathbb{R}^{s(K_n+l)+1} : \\
\boldsymbol{\varpi}_k &= \mathbf{0},\ \text{if } 0 \le \|\boldsymbol{\theta}_k\|_1 < \lambda, \\
\boldsymbol{\varpi}_k &= [(\|\boldsymbol{\theta}_k\|_1 - \lambda)/(a-1)] \boldsymbol{h}_k,\ \text{if } \lambda \le \|\boldsymbol{\theta}_k\|_1 \le a\lambda, \\
\boldsymbol{\varpi}_k &= \lambda \boldsymbol{h}_k,\ \text{if } \|\boldsymbol{\theta}_k\|_1 > a\lambda,\ \text{for all } 1 \le k \le s\},
\end{aligned}
$$

where $\boldsymbol{h}_k = (h_{k1}, \ldots, h_{k,K_n+l})^T \in \mathbb{R}^{K_n+l}$ and $h_{km} = sgn(\theta_{km})$ if $\theta_{km} \neq 0$ and $h_{km} \in [-1, 1]$ otherwise for $1 \leq m \leq K_n + l$. In the following, we analyze the subgradient of the unpenalized objective function, which is given by $\boldsymbol{\nu}(\boldsymbol{\theta}) = (\nu_0(\boldsymbol{\theta}), \boldsymbol{\nu}_1(\boldsymbol{\theta})^T, \ldots, \boldsymbol{\nu}_s(\boldsymbol{\theta})^T)^T$ where $\boldsymbol{\nu}_k(\boldsymbol{\theta}) = (\nu_{k1}(\boldsymbol{\theta}), \ldots, \nu_{k,K_n+l}(\boldsymbol{\theta}))^T$. The following lemma states the behavior of $\boldsymbol{\nu}(\boldsymbol{\theta}^*)$ when being evaluated at the oracle estimator.

Lemma S2.1. *Assume conditions in Theorem 3.2 are satisfied. For the oracle estimator $\boldsymbol{\theta}^*$, there exists a_i^* with $a_i^* = 0$ if $y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}^* \neq 0$ and $a_i^* \in [\tau - 1, \tau]$ otherwise, such that for $\boldsymbol{\nu}(\boldsymbol{\theta}^*)$ with $a_i = a_i^*$, with probability approaching one,*

(1) $\nu_0(\boldsymbol{\theta}^*) = 0$, $\boldsymbol{\nu}_k(\boldsymbol{\theta}^*) = \boldsymbol{0}$ *for* $k \in \mathcal{S}^*$,

(2) $|\nu_{km}(\boldsymbol{\theta}^*)| \leq c\lambda$, $\forall c > 0$, $k \notin \mathcal{S}^*$, $1 \leq m \leq K_n + l$,

(3) $\|\boldsymbol{\theta}_k^*\|_2 \geq (a + 1/2)\lambda$ *for* $k \in \mathcal{S}^*$.

To obtain the property of the SCAD penalized estimator, we require the following lemma which is a sufficient condition of a local minimizer for a convex-difference objective function.

Lemma S2.2. *(Lemma 2.1 in Wang et al. (2012)). If there exists a neighborhood U around the point $\boldsymbol{\theta}^*$ such that $\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigcap \frac{\partial G_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^*} \neq \emptyset$, $\forall \boldsymbol{\theta} \in U \bigcap dom(G_n)$, then $\boldsymbol{\theta}^*$ is a local minimizer of $G_n(\boldsymbol{\theta}) - H_n(\boldsymbol{\theta})$.*

Now we use Lemma S2.1 to prove that the oracle estimator satisfies Lemma

S2.2. Recall that

$$\frac{\partial G_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^*} = \{\boldsymbol{\pi}^* = (\pi_0^*, \boldsymbol{\pi}_1^{*T}, \ldots, \boldsymbol{\pi}_s^{*T})^T \in \mathbb{R}^{s(K_n+l)+1} : \pi_0^* \equiv \nu_0(\boldsymbol{\theta}^*);\ \boldsymbol{\pi}_k^* \equiv \boldsymbol{\nu}_k(\boldsymbol{\theta}^*) + \lambda \boldsymbol{l}_k,\ for\ 1 \le k \le s\},$$

where $\boldsymbol{l}_k = (l_{k1}, \ldots, l_{k,K_n+l})^T \in \mathbb{R}^{K_n+l}$ and $l_{km} = sgn(\theta_{km})$ if $\theta_{km} \neq 0$ and $l_{km} \in [-1, 1]$ otherwise for $1 \le m \le K_n + l$.

$$\begin{aligned}\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \{\boldsymbol{\varpi} &= (0, \boldsymbol{\varpi}_1^T, \ldots, \boldsymbol{\varpi}_s^T)^T \in \mathbb{R}^{s(K_n+l)+1} :\\ \boldsymbol{\varpi}_k &= \mathbf{0},\ \text{if } 0 \le \|\boldsymbol{\theta}_k\|_1 < \lambda,\\ \boldsymbol{\varpi}_k &= [(\|\boldsymbol{\theta}_k\|_1 - \lambda)/(a-1)]\boldsymbol{h}_k,\ \text{if } \lambda \le \|\boldsymbol{\theta}_k\|_1 \le a\lambda,\\ \boldsymbol{\varpi}_k &= \lambda \boldsymbol{h}_k,\ \text{if } \|\boldsymbol{\theta}_k\|_1 > a\lambda,\ \text{for all } 1 \le k \le s\},\end{aligned}$$

where $\boldsymbol{h}_k = (h_{k1}, \ldots, h_{k,K_n+l})^T \in \mathbb{R}^{K_n+l}$ and $h_{km} = sgn(\theta_{km})$ if $\theta_{km} \neq 0$ and $h_{km} \in [-1, 1]$ otherwise for $1 \le m \le K_n + l$.

Consider any $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*, \lambda/(2(\sqrt{K_n + l})))$ where $\mathcal{B}(\boldsymbol{\theta}^*, \lambda/(2(\sqrt{K_n + l})))$ denotes the ball with the center $\boldsymbol{\theta}^*$ and radius $\lambda/(2(\sqrt{K_n + l}))$. First consider $k \in \mathcal{S}^*$. From Lemma S2.1(1), there exists a_i^* such that $\pi_0^* = 0$ and $\boldsymbol{\pi}_k^* = \lambda \boldsymbol{l}_k$. On the other hand, from Lemma S2.1(3) we have $\|\boldsymbol{\theta}_k\|_1 \ge \|\boldsymbol{\theta}_k\|_2 \ge \|\boldsymbol{\theta}_k^*\|_2 - \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\|_2 \ge (a + 1/2)\lambda - \lambda/(2\sqrt{K_n + l}) \ge a\lambda$. Thus $\boldsymbol{\varpi}_k = \lambda \boldsymbol{h}_k$. Obviously, $\boldsymbol{\varpi}_k = \boldsymbol{\pi}_k^*$ if $\boldsymbol{l}_k = \boldsymbol{h}_k$.

Then consider $k \notin \mathcal{S}^*$. From Lemma S2.1(2), we have $|\nu_{km}(\boldsymbol{\theta}^*)| < \lambda$ for any $k \notin \mathcal{S}^*$ and $1 \leq m \leq K_n+l$. By definition, $\boldsymbol{\pi}_k^* = (\nu_{k1}(\boldsymbol{\theta}^*), \ldots, \nu_{k,K_n+l}(\boldsymbol{\theta}^*))^T + \lambda \boldsymbol{l}_k$ where $\boldsymbol{l}_k \in [-1,1]^{K_n+l}$. Thus there exists $\boldsymbol{l}_k^*$ such that $\boldsymbol{\pi}_k^* = \mathbf{0}$. On the other hand, $\boldsymbol{\theta}_k^* = \mathbf{0}$ for $k \notin \mathcal{S}^*$. And $\|\boldsymbol{\theta}_k\|_1 \leq \sqrt{K_n+l}\|\boldsymbol{\theta}_k\|_2 \leq \sqrt{K_n+l}(\|\boldsymbol{\theta}_k^*\|_2 + \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\|_2) = \lambda/2 \leq \lambda$. Thus $\boldsymbol{\varpi}_k = \mathbf{0}$ from the definition.

We have shown that there exists a neighborhood U around the point $\boldsymbol{\theta}^*$ such that $\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigcap \frac{\partial G_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^*} \neq \emptyset$, $\forall \boldsymbol{\theta} \in U \bigcap dom(G_n)$. Applying Lemma S2.2, we can get Theorem 3.2.

Proof of Lemma S2.1. (1) By convex optimization theory, $\mathbf{0}$ is in the subdifferential of the oracle objective function. Thus, there exists a_i^* as described in the lemma such that (1) is satisfied.

(2) From the definition, we have

$$\begin{aligned}\nu_{km}(\boldsymbol{\theta}^*) \;=\; & -\tau n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik}) I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}^* > 0) + (1-\tau)n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik}) I(y_i \\ & -\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}^* < 0) - n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik}) a_i^*,\end{aligned}$$

where $k \notin \mathcal{S}^*$, $1 \leq m \leq K_n + l$ and a_i^* satisfies the condition in (1). Let

$\mathcal{D} = \{i : y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta}^* = 0\}$. Then

$$\nu_{km}(\boldsymbol{\theta}^*) = n^{-1} \sum_{i=1}^{n} B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta}^* \leq 0) - \tau] - n^{-1} \sum_{i \in \mathcal{D}} B_m(\hat{\boldsymbol{\zeta}}_{ik})(a_i^* + (1-\tau)).$$

With probability one (Section 2.2 Koenker, 2005), $|\mathcal{D}| = K_n$. Therefore,

$$n^{-1} \sum_{i \in \mathcal{D}} B_m(\hat{\boldsymbol{\zeta}}_{ik})(a_i^* + (1-\tau)) = O_p(K_n^{1/2}/n) = o_p(\lambda),$$

since $K_n^{1/2}/n \ll n^{-1/2} = o(\lambda)$. We will show that

$$P(\max_{\substack{k \in \mathcal{S}^{*c} \\ 1 \leq m \leq K_n + l}} |n^{-1} \sum_{i=1}^{n} B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T \boldsymbol{\theta}^* \leq 0) - \tau]| > c\lambda) \to 0.$$

Define $\Theta_{\mathcal{S}^*,n} = \mathcal{B}(\boldsymbol{\theta}^0_{\mathcal{S}^*}, \sqrt{\frac{K_n}{n}} d_n)$. Note that

$$\begin{aligned}
& P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}^* \leq 0) - \tau]| > c\lambda) \\
\leq\ & P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T\boldsymbol{\theta}^* \leq 0) - I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| > c\lambda/2) \\
& +P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0) - \tau]| > c\lambda/2) \\
\leq\ & P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) \\
& -I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| > c\lambda/2) + A_1 \\
\leq\ & P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) - I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0) \\
& -P(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) + P(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| > c\lambda/4) \\
& +P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[P(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) \\
& -P(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| > c\lambda/4) + A_1 \\
=\ & A_3 + A_2 + A_1.
\end{aligned}$$

Next we will show that A_1, A_2 and A_3 converge to zero one by one.

Step 1. By definition, we have

$$A_1 = P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0) - \tau]| > c\lambda/2).$$

Since $|B_m(\hat{\boldsymbol{\zeta}}_{ik})| = O_P(1/\sqrt{K_n})$, it holds by Hoeffding's inequality

$$A_1 \leq 2sK_n \exp\{-CnK_n\lambda^2\} = 2\exp(C\log(n) - CnK_n\lambda^2) \to 0.$$

Step 2. By definition, we have

$$\begin{aligned} A_2 &= P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[P(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) \\ &\quad -P(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| > c\lambda/4). \end{aligned}$$

Note that

$$\begin{aligned} &\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[P(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \leq 0) - P(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \leq 0)]| \\ = &\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\leq m\leq K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[F_i(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T(\boldsymbol{\theta}_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) - R_i - u_i) - F_i(0)]| \\ \leq &CK_n^{-1/2} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} n^{-1}\sum_{i=1}^n(|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T(\boldsymbol{\theta}_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) + R_i + u_i|) \qquad \text{(S2.1)} \\ \leq &CK_n^{-1/2} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} [\sqrt{n^{-1}(\boldsymbol{\theta}_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*})^T\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}^T(\boldsymbol{\theta}_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*})} + \sum_{i=1}^n |R_i|/n + \sup_i |u_i|] \\ \leq &CK_n^{-1/2}O_p(\frac{d_n}{n^{1/2}} + \frac{s}{n^{1/2}} + K_n^{-r}) = O_p(\frac{d_n}{K_n^{1/2}n^{1/2}}) = o(\lambda), \end{aligned}$$

where the second inequality applies Jensen's inequality (similar to Lemma B.5 in Sherwood and Wang (2016)) and the last inequality follows from $\lambda_{\max}(\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}^T) = O_p(\frac{n}{K_n})$ (Lemma S1.1(3)), $\sum_{i=1}^n |R_i|/n = O_p(\frac{s}{n^{1/2}})$ and $\sup_i |u_i| = O_p(K_n^{-r})$.

Since $\max\{n^{-1/2}, sK_n^{-1/2}n^{-1/2}\} = o(\lambda)$, we have the last equality. Thus we can conclude that $A_2 \to 0$.

Step 3. By definition, we have

$$A_3 = P(\max_{\substack{k\in\mathcal{S}^{*c} \\ 1\le m\le K_n+l}} \sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |n^{-1}\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \le 0) - I(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \le 0)$$
$$- P(y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*} \le 0) + P(y_i - g(\boldsymbol{\zeta}_{i,\mathcal{S}^*}) \le 0)]| > c\lambda/4).$$

The set $\Theta_{\mathcal{S}^*,n}$ can be covered by a set of balls denoted as $\{\Theta^1_{\mathcal{S}^*,n}, \ldots, \Theta^N_{\mathcal{S}^*,n}\}$ with radius $C\sqrt{\frac{K_n}{n}\frac{d_n}{n^2}}$ with cardinality $N \le n^{2(q(K_n+l)+1)}$. Denote by $\boldsymbol{\theta}^l_{\mathcal{S}^*}$, $l = 1, \ldots, N$, the centers in the balls. Let $\epsilon_i(\boldsymbol{\theta}_{\mathcal{S}^*}) = y_i - \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T\boldsymbol{\theta}_{\mathcal{S}^*}$, we have for each k and m,

$$\begin{aligned}
& P(\sup_{\boldsymbol{\theta}_{\mathcal{S}^*}\in\Theta_{\mathcal{S}^*,n}} |\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(\epsilon_i(\boldsymbol{\theta}_{\mathcal{S}^*}) \le 0) - I(\epsilon_i \le 0) - P(\epsilon_i(\boldsymbol{\theta}_{\mathcal{S}^*}) \le 0) + P(\epsilon_i \le 0)]| > n\lambda) \\
\le\ & \sum_{l=1}^N P(|\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) - I(\epsilon_i \le 0) - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) + P(\epsilon_i \le 0)]| > n\lambda/2) \\
& + \sum_{l=1}^N P(\sup_{\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*}\in\Theta^l_{\mathcal{S}^*,n}} |\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(\epsilon_i(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*}) \le 0) - I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) - P(\epsilon_i(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*}) \le 0) \\
& + P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0)]| > n\lambda/2) \\
=\ & T_{1km} + T_{2km}.
\end{aligned}$$

In the following, we will show that $T_{1km} \le C\exp(K_n\log(n) - CnK_n^{1/2}\lambda)$ and $T_{2km} \le C\exp(K_n\log(n) - CnK_n^{1/2}\lambda)$. If so, then the following completes the proof:

$$\begin{aligned} A_3 &\le \sum_{\substack{k\in\mathcal{S}^{*c}\\ 1\le m\le K_n+l}} (T_{1km}+T_{2km}) \\ &\le CsK_n\exp(K_n\log(n) - CnK_n^{1/2}\lambda) \\ &= C\exp(C\log(n) + K_n\log(n) - CnK_n^{1/2}\lambda) = o(1). \end{aligned}$$

To evaluate T_{1km}, let $\vartheta_{ikm} = B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) - I(\epsilon_i \le 0) - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) + P(\epsilon_i \le 0)]$. Note that $\max_i|\vartheta_{ikm}| \le \frac{1}{\sqrt{K_n}}$ and

$$\begin{aligned} \sum_{i=1}^n Var(\vartheta_{ikm}) &\le \sum_{i=1}^n EB_m(\hat{\boldsymbol{\zeta}}_{ik})^2[I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le 0) - I(\epsilon_i \le 0)]^2 \\ &\le \frac{1}{K_n}\sum_{i=1}^n P(|\epsilon_i| \le |\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T(\boldsymbol{\theta}^l_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) + R_i + u_i|) \\ &\le \frac{C}{K_n}\sum_{i=1}^n |\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_i)^T(\boldsymbol{\theta}^l_{\mathcal{S}^*} - \boldsymbol{\theta}^0_{\mathcal{S}^*}) + R_i + u_i| = O_p(\frac{n^{1/2}d_n}{K_n}), \end{aligned}$$

where the last equality follows from (S2.1). Applying Bernstein's inequality,

$$\begin{aligned} T_{1km} &\le N\exp(-\frac{Cn^2\lambda^2}{Cn^{1/2}d_nK_n^{-1} + Cn\lambda K_n^{-1/2}}) \\ &\le N\exp(-CnK_n^{1/2}\lambda) = C\exp(K_n\log(n) - CnK_n^{1/2}\lambda). \end{aligned}$$

To evaluate T_{2km}, note that $I(\epsilon_i(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*} \le 0) = I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \le \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})^T(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*} -$

$\boldsymbol{\theta}^l_{\mathcal{S}^*}))$ and $I(x \leq s)$ is an increasing function of s. Thus we have

$$\begin{aligned}
& \sup_{\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*} \in \Theta^l_{\mathcal{S}^*,n}} |\sum_{i=1}^n B_m(\hat{\boldsymbol{\zeta}}_{ik})[I(\epsilon_i(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*}) \leq 0) - I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0) - P(\epsilon_i(\tilde{\boldsymbol{\theta}}_{\mathcal{S}^*}) \leq 0) + P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0)]| \\
\leq & \sum_{i=1}^n |B_m(\hat{\boldsymbol{\zeta}}_{ik})| \times |I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) - I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0) \\
& - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq -\|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) + P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0)| \\
\leq & \sum_{i=1}^n |B_m(\hat{\boldsymbol{\zeta}}_{ik})| \times |I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) - I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0) \\
& - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) + P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0)| \\
& + \sum_{i=1}^n |B_m(\hat{\boldsymbol{\zeta}}_{ik})| \times |P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq -\|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}})|.
\end{aligned}$$

Note that

$$\begin{aligned}
& \sum_{i=1}^n |B_m(\hat{\boldsymbol{\zeta}}_{ik})| \times |P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq -\|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}})| \\
\leq & \frac{C}{\sqrt{K_n}} \sum_{i=1}^n \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\| \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}} = O_p(d_n n^{-3/2}) = o_p(n\lambda).
\end{aligned}$$

Hence for n sufficiently large, $T_{2km} \leq \sum_{l=1}^N P(\sum_{i=1}^n \varsigma_{ilkm} \geq n\lambda/4)$, where

$$\begin{aligned}
\varsigma_{ilkm} = & |B_m(\hat{\boldsymbol{\zeta}}_{ik})| \times |I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) - I(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0) \\
& - P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 \sqrt{\frac{K_n}{n} \frac{d_n}{n^2}}) + P(\epsilon_i(\boldsymbol{\theta}^l_{\mathcal{S}^*}) \leq 0)|.
\end{aligned}$$

Similarly to the evaluation of T_{1km}, we cam show that

$$\sum_{i=1}^{n} Var(\varsigma_{ilkm}) \leq \frac{n}{K_n} \|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\|_2 \sqrt{\frac{K_n}{n}} \frac{d_n}{n^2} = O_p(\frac{d_n}{n^{3/2} K_n^{1/2}}).$$

Applying Bernstein's inequality, we have

$$\begin{aligned} T_{2km} &\leq N \exp(-\frac{Cn^2\lambda^2}{Cn^{-3/2} d_n K_n^{-1/2} + Cn\lambda K_n^{-1/2}}) \\ &\leq N \exp(-CnK_n^{1/2}\lambda) = C\exp(K_n \log(n) - CnK_n^{1/2}\lambda). \end{aligned}$$

(3) Note that $\min_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^*\|_2 \geq \min_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^0\|_2 - \max_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}_k^0\|_2$. From the proof of Theorem 3.1, we have $\max_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}_k^0\|_2 \leq \|\boldsymbol{\theta}_{\mathcal{S}^*}^* - \boldsymbol{\theta}_{\mathcal{S}^*}^0\|_2 = O_p(\frac{K_n}{\sqrt{n}} + \sqrt{\frac{K_n}{n}} s)$. By Condition 5, we have $\min_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^0\|_2 \geq C(\frac{K_n}{\sqrt{n}} + \sqrt{\frac{K_n}{n}} s) n^\alpha$. Thus for $k \in \mathcal{S}^*$, $\|\boldsymbol{\theta}_k^*\|_2 \geq C(\frac{K_n}{\sqrt{n}} + \sqrt{\frac{K_n}{n}} s) n^\alpha \geq (a + 1/2)\lambda$.

S3 Proof of Theorem 3.3

For each candidate model $\mathcal{S}$, similarly we can define $J_{\mathcal{S}} = (K_n + l)|\mathcal{S}| + 1$ and

$$\hat{\boldsymbol{W}}_{\mathcal{S}} = (\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{1,\mathcal{S}}), \ldots, \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{n,\mathcal{S}}))^T \in \mathbb{R}^{n \times J_{\mathcal{S}}},$$

$$\hat{\boldsymbol{W}}^2_{B,\mathcal{S}} = \hat{\boldsymbol{W}}^T_{\mathcal{S}} \boldsymbol{B}_n \hat{\boldsymbol{W}}_{\mathcal{S}} \in \mathbb{R}^{J_{\mathcal{S}} \times J_{\mathcal{S}}}, \text{ where } \boldsymbol{B}_n = \text{diag}(f_1(0), \ldots, f_n(0)),$$

$$\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}}) = \hat{\boldsymbol{W}}^{-1}_{B,\mathcal{S}} \boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}}) \in \mathbb{R}^{J_{\mathcal{S}}},$$

$$\boldsymbol{\delta}_{\mathcal{S}} = \hat{\boldsymbol{W}}_{B,\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}^0_{\mathcal{S}}) \in \mathbb{R}^{J_{\mathcal{S}}}.$$

$$R_{i,\mathcal{S}} = (\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}}) - \boldsymbol{W}(\boldsymbol{\zeta}_{i,\mathcal{S}}))^T \boldsymbol{\theta}^0_{\mathcal{S}},$$

We first show lemmas used in proof. With condition (C5), the following lemma holds parallelly with Lemma S1.1. All constants in the following lemma do not depend on $\mathcal{S}$.

Lemma S3.1. *We have the following properties for the spline basis vector:*

(1) $E(\|\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\|_2) \leq b_1|\mathcal{S}|$*, for some positive constant* b_1 *for all* n *sufficiently large.*

(2) $b_2 K_n^{-1} \leq E(\lambda_{min}(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T)) \leq E(\lambda_{max}(\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\boldsymbol{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T)) \leq b_2^* K_n^{-1}$*, for some positive constants* b_2 *and* b_2^* *for* n *sufficiently large.*

(3) $E(\|\hat{\boldsymbol{W}}^{-1}_{B,\mathcal{S}}\|) \geq b_3\sqrt{K_n/n}$*, for some positive* b_3 *for all* n *sufficiently large.*

For a matrix A, $\|A\| = \sqrt{\lambda_{max}(A^T A)}$ denotes the spectral norm.

(4) $\max_i \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\|_2 = O_p(\sqrt{\frac{J_{\mathcal{S}}}{n}})$.

Let $\mathcal{M}^{OF} = \{\mathcal{S} : \mathcal{S}^* \subseteq \mathcal{S}\}$ be the set of overfitted model and $B_\eta(\mathcal{S}) = \{\boldsymbol{\delta} \in \mathbb{R}^{J_{\mathcal{S}}} : \|\boldsymbol{\delta}\| \leq \eta\}$. We denote the maximum of $J_{\mathcal{S}}$ over $\mathcal{S} \in \mathcal{M}^{OF}$ by J. For $\mathcal{S} \in \mathcal{M}^{OF}$, $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$ is defined as

$$\hat{\boldsymbol{\delta}}_{\mathcal{S}} = \arg\min_{\boldsymbol{\delta}_{\mathcal{S}}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}} - R_{i,\mathcal{S}} - u_i).$$

Denote $Q_i(\boldsymbol{\delta}_{\mathcal{S}}) = \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}} - R_{i,\mathcal{S}} - u_i)$ and $D_i(\boldsymbol{\delta}_{\mathcal{S}}) = Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0) - E[Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0)|X_i] + \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}}\psi_\tau(\epsilon_i)$ and $\psi_\tau(u) = \tau - I(u < 0)$.

Lemma S3.2. *Assume conditions in Theorem 3.3 hold. Then for any sequence $L_n = O(n^\gamma)$ with small $\gamma > 0$ satisfying $L_n^3/\sqrt{n} \to 0$ and $L_n^2(s+\sqrt{K_n})/\sqrt{n} \to 0$, we have*

$$\sup_{\mathcal{S}\in\mathcal{M}^{OF}} \sup_{\|\delta_{\mathcal{S}}\|\leq L_n d_{\mathcal{S}}} |d_{\mathcal{S}}^{-2}\sum_{i=1}^{n} D_i(\boldsymbol{\delta}_{\mathcal{S}})| = o_p(1), \tag{S3.1}$$

where $d_{\mathcal{S}} = \sqrt{J_{\mathcal{S}}} + s$.

This lemma provides a uniform approximation of $\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0)$ and can be proved by the same technical arguments in the proof of step 1 for Lemma S1.2.

Proof. It's equivalent to show

$$\sup_{\mathcal{S}\in\mathcal{M}^{OF}} \sup_{\boldsymbol{\delta}_{\mathcal{S}}\in B_1(\mathcal{S})} |d_{\mathcal{S}}^{-2}\sum_{i=1}^{n} D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}})| = o_p(1). \tag{S3.2}$$

Let F_{n4} denote the event $\max_i \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\|_2 \leq \alpha_1\sqrt{\frac{J_{\mathcal{S}}}{n}}$ for some positive α_1. Lemma S3.1(4) implies that $P(F_{n4}) \to 1$ as $n \to \infty$. F_{n2} and F_{n3} is defined in the proof of Lemma S1.2. Then it's sufficient to show for any $\varepsilon > 0$

$$P(\sup_{\mathcal{S}\in\mathcal{M}^{OF}} \sup_{\boldsymbol{\delta}_{\mathcal{S}}\in B_1(\mathcal{S})} d_{\mathcal{S}}^{-2}|\sum_{i=1}^{n} D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}})| > \varepsilon, F_{n2}\cap F_{n3}\cap F_{n4}) \to 0.$$

Partition $B_1(\mathcal{S})$ as a union of balls with radius $m_0 = \frac{\varepsilon}{4\alpha_1 J_{\mathcal{S}}^{1/2} n^{1/2} L_n d_{\mathcal{S}}^{-1}}$, say $\Delta_1, \ldots, \Delta_{M_n}$. We have $M_n \leq C(\frac{CJ_{\mathcal{S}}^{1/2} n^{1/2} L_n d_n^{-1}}{\varepsilon})^{J_n}$, where C is a positive constant. Let $\boldsymbol{\delta}_{\mathcal{S}}^1, \ldots, \boldsymbol{\delta}_{\mathcal{S}}^{M_n}$ be arbitrary points in $\Delta_1, \ldots, \Delta_{M_n}$ respectively. Similarly we can show for all $\mathcal{S}$:

(i) $\sup_{\boldsymbol{\delta}_{\mathcal{S}}\in\Delta_m} |\sum_{i=1}^n (D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}) - D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}^m)|I(F_{n2}\cap F_{n3}\cap F_{n4}) < d_{\mathcal{S}}^2\varepsilon/2$.

(ii) $\max_i |D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}^m)|I(F_{n2}\cap F_{n3}\cap F_{n4}) \leq CL_n d_{\mathcal{S}} J_{\mathcal{S}}^{1/2} n^{-1/2}$.

(iii) $\sum_{i=1}^n Var[D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}^m)I(F_{n2}\cap F_{n3}\cap F_{n4})|X_i] \leq CJ_{\mathcal{S}} L_n^2 d_{\mathcal{S}}^2(\frac{s}{\sqrt{n}} + K_n^{-r}) + CL_n^3 d_{\mathcal{S}}^3 J_{\mathcal{S}}^{1/2} n^{-1/2}$.

By Bernstein inequality, we have

$$
\begin{aligned}
& P(\sup_{\mathcal{S}\in\mathcal{M}^{OF}} \sup_{\boldsymbol{\delta}_{\mathcal{S}}\in B_1(\mathcal{S})} d_{\mathcal{S}}^{-2}|\sum_{i=1}^{n} D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}})| > \varepsilon, F_{n2}\cap F_{n3}\cap F_{n4}) \\
\leq\ & \sum_{\mathcal{S}\in\mathcal{M}^{OF}} \sum_{m=1}^{M_n} P(|\sum_{i=1}^{n} D_i(L_n d_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}}^{m})| > d_{\mathcal{S}}^{2}\varepsilon/2, F_{n2}\cap F_{n3}\cap F_{n4}) \\
\leq\ & 2\sum_{\mathcal{S}\in\mathcal{M}^{OF}} \sum_{m=1}^{M_n} \exp\Big(\frac{-d_{\mathcal{S}}^{4}\varepsilon^2/4}{Cn^{-1/2}J_{\mathcal{S}}L_n^2 d_{\mathcal{S}}^2(s+K_n^{-r}\sqrt{n}) + CL_n^3 d_{\mathcal{S}}^3 J_{\mathcal{S}}^{1/2} n^{-1/2} + Cd_{\mathcal{S}}^3 L_n J_{\mathcal{S}}^{1/2} n^{-1/2}\varepsilon/2}\Big) \\
\leq\ & 2\sum_{\mathcal{S}\in\mathcal{M}^{OF}} \sum_{m=1}^{M_n} \exp\Big(\frac{-Cd_{\mathcal{S}}^{2} n^{1/2}}{J_{\mathcal{S}}L_n^2(s+K_n^{-r}\sqrt{n}) + CL_n^3 d_{\mathcal{S}} J_{\mathcal{S}}^{1/2}}\Big) \\
\leq\ & 2^s \exp\Big(CJ\log n - \frac{Cn^{1/2}}{L_n^2(s+K_n^{-r}\sqrt{n}) + L_n^3}\Big),
\end{aligned}
$$

which converges to zero. Hence the proof of the first step is complete.

Lemma S3.3. *Assume conditions in Theorem 3.3 hold. We have*

$$
\lim_{L\to\infty}\lim_{n\to\infty} P(\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\| \leq Ld_{\mathcal{S}}(\log n)^{1/2} \textit{ for all } \mathcal{S}\in\mathcal{M}^{OF}) = 1. \tag{S3.3}
$$

This lemma is different with Lemma S1.2 in that we provide a uniform bound for $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$ for all $\mathcal{S}\in\mathcal{M}^{OF}$.

Proof. By the convexity of ρ_τ, it suffices to show that, for any $\varepsilon > 0$, there exists a large constant $L > 0$ such that

$$
\liminf_{n} P(\inf_{\mathcal{S}\in\mathcal{M}^{OF}} \inf_{\|\boldsymbol{\delta}_{\mathcal{S}}\|=Ld_{\mathcal{S}}(\log n)^{1/2}} \sum_{i=1}^{n} Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0) > 0) > 1-\varepsilon. \tag{S3.4}
$$

From Lemma S3.2, if follows that for any $\boldsymbol{\delta}_{\mathcal{S}} : \|\boldsymbol{\delta}_{\mathcal{S}}\| = Ld_{\mathcal{S}}(\log n)^{1/2}$ with $\mathcal{S} \in \mathcal{M}^{OF}$,

$$\begin{aligned}\sum_{i=1}^{n} Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0) &= -\sum_{i=1}^{n} \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T \boldsymbol{\delta}_{\mathcal{S}} \psi_\tau(\epsilon_i) + \sum_{i=1}^{n} E[Q_i(\boldsymbol{\delta}_{\mathcal{S}}) - Q_i(0)|X_i] + d_{\mathcal{S}}^2 o_p(1) \\ &= A_n(\boldsymbol{\delta}_{\mathcal{S}}) + B_n(\boldsymbol{\delta}_{\mathcal{S}}) + d_{\mathcal{S}}^2 o_p(1).\end{aligned}$$

For $A_n(\boldsymbol{\delta}_{\mathcal{S}})$, we get $|A_n(\boldsymbol{\delta}_{\mathcal{S}})| \leq \max_{1\leq k\leq s} \| \sum_{i=1}^n \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,k})^T \psi_\tau(\epsilon_i)\| |S|^{1/2} \|\boldsymbol{\delta}_{\mathcal{S}}\|$.

Since $\max_{1\leq k\leq s} \sum_{i=1}^n \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,k})\|^2 \leq MK_n$ for sufficiently large M, we have

$$\begin{aligned}& P(\max_{1\leq k\leq s} \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,k})\psi_\tau(\epsilon_i)\|^2 \geq M^2 K_n \log n | T) \\ \leq\ & sK_n \max_{k,m} P(|\sum_{i=1}^{n} \tilde{\boldsymbol{W}}_m(\hat{\boldsymbol{\zeta}}_{i,k})\psi_\tau(\epsilon_i)| > \{M \sum_{i=1}^{n} (\tilde{\boldsymbol{W}}_m(\hat{\boldsymbol{\zeta}}_{i,k}))^2 \log n\}^{1/2} | T) \\ \leq\ & 2sK_n \exp(-M \log n/8),\end{aligned}$$

where the last inequality is from Hoeffding's inequality. This implies

$$\max_{1\leq k\leq s} \|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,k})\psi_\tau(\epsilon_i)\| = O_p((K_n \log n)^{1/2}).$$

Consequently, we have

$$P(|A_n(\boldsymbol{\delta}_{\mathcal{S}})| < (J_{\mathcal{S}} \log n)^{1/2} \|\boldsymbol{\delta}_{\mathcal{S}}\| \text{ for all } \mathcal{S} \in \mathcal{M}^{OF}) \to 1.$$

We deal with $B_n(\boldsymbol{\delta}_{\mathcal{S}})$ similar with step 2 of Lemma S1.2. Applying Knight's identity twice,

$$\begin{aligned} B_n(\boldsymbol{\delta}_{\mathcal{S}}) &= \sum_{i=1}^{n} E[\int_{R_{i,\mathcal{S}}+u_i}^{\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}}+R_{i,\mathcal{S}}+u_i} (I(\epsilon_i < s) - I(\epsilon_i < 0))ds|X_i] \\ &= C\|\boldsymbol{\delta}_{\mathcal{S}}\|^2 + C\|\boldsymbol{\delta}_{\mathcal{S}}\|(s + K_n^{-r}\sqrt{n}). \end{aligned}$$

The last equality holds because $R_{i,\mathcal{S}} = R_{i,\mathcal{S}^*}$ for any overfitted model $\mathcal{S}$. Consequently, for sufficient large L, $C\|\boldsymbol{\delta}_{\mathcal{S}}\|^2$ dominates all other terms and impies (S3.4).

Lemma S3.4. *Assume conditions in Theorem 3.3 hold. Then given a constant $\eta > 0$ we have*

$$\sup_{|\mathcal{S}|\le s} \sup_{\boldsymbol{\delta}_{\mathcal{S}}\in B_\eta(\mathcal{S})} |\sum_{i=1}^{n} g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}})| = O_p((nJ\log n)^{1/2})$$

where $g_i(\boldsymbol{\delta}_{\mathcal{S}}) = \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}} - R_{i,\mathcal{S}} - u_i) - \rho_\tau(\epsilon_i - R_{i,\mathcal{S}} - u_i) - E(\rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\boldsymbol{\delta}_{\mathcal{S}} - R_{i,\mathcal{S}} - u_i) - \rho_\tau(\epsilon_i - R_{i,\mathcal{S}} - u_i)|X_i)$.

Proof. This lemma can be proved by the arguments of Lemma A.3 in Lee et al. (2014), where chain technique is used. For $m \ge 0$, let $\Theta_n(2^{-m}\eta, \mathcal{S})$ denote a grid of points in $B_\eta(\mathcal{S})$ such that for every $\boldsymbol{\delta}_{\mathcal{S}} \in B_\eta(\mathcal{S})$ there exists $\boldsymbol{\delta}_{\mathcal{S}}^m \in \Theta_n(2^{-m}\eta, \mathcal{S})$ such that $\|\boldsymbol{\delta}_{\mathcal{S}} - \boldsymbol{\delta}_{\mathcal{S}}^m\| \le 2^{-m}\eta$. For a given constant $C > 0$, define

$M_n = \min\{m : 2^{-m}\eta \le (C/8M)n^{-1/2}(\log n)^{1/2}\}$. Then

$$\sup_{\boldsymbol{\delta}_{\mathcal{S}} \in B_\eta(\mathcal{S})} |\sum_{i=1}^n g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}) - g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^{M_n})| \le 4\sqrt{n}\sum_{i=1}^n |\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T(\boldsymbol{\delta}_{\mathcal{S}} - \boldsymbol{\delta}_{\mathcal{S}}^{M_n})| \le \frac{C}{2}(nJ_{\mathcal{S}}\log n)^{1/2}.$$

Consequently, we have

$$\begin{aligned}
\boldsymbol{I}_n(\mathcal{X}) &= P(\sup_{|\mathcal{S}|\le s}\sup_{\boldsymbol{\delta}_{\mathcal{S}} \in B_\eta(\mathcal{S})} |\sum_{i=1}^n g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}})| \ge C(nJ\log n)^{1/2}|T) \\
&\le P(\sup_{|\mathcal{S}|\le s}\sup_{\boldsymbol{\delta}_{\mathcal{S}} \in B_\eta(\mathcal{S})} |\sum_{i=1}^n g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^{M_n})| \ge \frac{C}{2}(nJ\log n)^{1/2}|T) \\
&\le \sum_{|\mathcal{S}|\le s} P(\sup_{\boldsymbol{\delta}_{\mathcal{S}} \in B_\eta(\mathcal{S})} \sum_{m=1}^{M_n} |\sum_{i=1}^n g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^m) - g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^{m-1})| \ge \frac{C}{2}(nJ\log n)^{1/2}|T) \\
&\le \sum_{|\mathcal{S}|\le s}\sum_{m=1}^{M_n} N_m(\mathcal{S})N_{m-1}(\mathcal{S}) \times \max_{*} P(|\sum_{i=1}^n g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^m) - g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^{m-1})| \ge \frac{C}{2}a_m(nJ\log n)^{1/2}|T).
\end{aligned}$$

For the first inequality, note that $\boldsymbol{\delta}_{\mathcal{S}}^{M_n}$ depends on $\boldsymbol{\delta}_{\mathcal{S}}$. For the second inequality, we take $\boldsymbol{\delta}_{\mathcal{S}}^m = 0$ when $m = 0$. For the last inequality, $N_m(\mathcal{S})$ is the cardinality of the set $\Theta_n(2^{-m}\eta, \mathcal{S})$ which is bounded by $(1 + 4 \cdot 2^m)^{J_{\mathcal{S}}}$; a_m is positive numbers such that $\sum_{m=1}^{M_n} a_m \le 1$; and $\max_*$ is taken over all $\boldsymbol{\delta}_{\mathcal{S}}^m$ and $\boldsymbol{\delta}_{\mathcal{S}}^{m-1}$ such that $\|\boldsymbol{\delta}_{\mathcal{S}}^m - \boldsymbol{\delta}_{\mathcal{S}}^{m-1}\| \le 3(2^{-m}\eta)$. Note that $|g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^m) - g_i(\sqrt{n}\boldsymbol{\delta}_{\mathcal{S}}^{m-1})| \le 4\sqrt{n}|\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T(\boldsymbol{\delta}_{\mathcal{S}}^m - \boldsymbol{\delta}_{\mathcal{S}}^{m-1})|$ and $\sum_{i=1}^n |\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T(\boldsymbol{\delta}_{\mathcal{S}}^m - \boldsymbol{\delta}_{\mathcal{S}}^{m-1})|^2 \le 9\bar{f}2^{-2m}\eta^2$ for some constant $\bar{f} > 0$ since $\sum_{i=1}^n f_i(0)\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})\tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T = \hat{\boldsymbol{W}}_{B,\mathcal{S}}^{-1}\hat{\boldsymbol{W}}_{\mathcal{S}}B\hat{\boldsymbol{W}}_{\mathcal{S}}^T\hat{\boldsymbol{W}}_{B,\mathcal{S}}^{-1} =$

I. Similar to (A.14) in Lee et al. (2014), we can take

$$a_m = \max\{2^{-m}m^{1/2}/8, \frac{96\bar{f}^{1/2}2^{-m}\eta(\log(1+4\cdot 2^m))^{1/2}}{C(\log n)^{1/2}}\}.$$

Applying Hoeffding's inequality, we get that

$$\boldsymbol{I}_n(\mathcal{X}) \leq 2\sum_{|\mathcal{S}|\leq s}\sum_{m=1}^{M_n}\exp(2J\log(1+4\cdot 2^m) - \frac{C^2 a_m^2 J\log n}{48^2\bar{f}2^{-2m}\eta^2}),$$

which converges to zero for sufficiently large $C > 0$.

Proof of Theorem 3.3. Let $\mathcal{M}^{UF} = \{\mathcal{S} : \mathcal{S}^* \nsubseteq \mathcal{S}\}$ denote the underfitted model. It suffices to show that

$$P(\min_{\mathcal{S}\in\mathcal{M}^{OF},\mathcal{S}\neq\mathcal{S}^*} \text{BIC}(\mathcal{S}) > \text{BIC}(\mathcal{S}^*)) \to 1, \quad \text{(S3.5)}$$

$$P(\min_{\mathcal{S}\in\mathcal{M}^{UF}} \text{BIC}(\mathcal{S}) > \text{BIC}(\mathcal{S}^*)) \to 1. \quad \text{(S3.6)}$$

First we prove (S3.5). Using similar arguments as in the proof of Lemma S3.3, and the fact that $|B_n(\boldsymbol{\delta}_{\mathcal{S}})| \leq C\|\boldsymbol{\delta}_{\mathcal{S}}\|^2$, we can choose a sequence $\{L_n\}$, not depending on $\mathcal{S}$, such that $\frac{L_n}{C_n} \to 0$ and $\frac{L_n s^2}{JC_n} \to 0$, and

$$|\sum_{i=1}^{n} Q_i(\hat{\boldsymbol{\delta}}_{\mathcal{S}}) - Q_i(0)| \leq L_n d_{\mathcal{S}}^2 \log n, \quad \text{(S3.7)}$$

for any $\mathcal{S} \in \mathcal{M}^{OF}$ with probability tending to one. Then we have

$$
\begin{aligned}
&\min_{\mathcal{S}\in\mathcal{M}^{OF},\mathcal{S}\neq\mathcal{S}^*} \mathrm{BIC}(\mathcal{S}) - \mathrm{BIC}(\mathcal{S}^*) \\
&= \min_{\mathcal{S}\in\mathcal{M}^{OF},\mathcal{S}\neq\mathcal{S}^*} \log(1 + \frac{n^{-1}\sum_{i=1}^n Q_i(\hat{\boldsymbol{\delta}}_{\mathcal{S}}) - Q_i(\hat{\boldsymbol{\delta}}_{\mathcal{S}^*})}{n^{-1}\sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\hat{\boldsymbol{\delta}}_{\mathcal{S}^*} - R_i - u_i)}) \\
&\qquad + (J_{\mathcal{S}} - J_{\mathcal{S}^*})\frac{\log n}{2n}C_n \\
&\geq \min_{\mathcal{S}\in\mathcal{M}^{OF},\mathcal{S}\neq\mathcal{S}^*} -2|\frac{n^{-1}\sum_{i=1}^n Q_i(\hat{\boldsymbol{\delta}}_{\mathcal{S}}) - Q_i(\hat{\boldsymbol{\delta}}_{\mathcal{S}^*})}{n^{-1}\sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}^*})\hat{\boldsymbol{\delta}}_{\mathcal{S}^*} - R_i - u_i)}| + (J_{\mathcal{S}} - J_{\mathcal{S}^*})\frac{\log n}{2n}C_n \\
&\geq \min_{\mathcal{S}\in\mathcal{M}^{OF},\mathcal{S}\neq\mathcal{S}^*} \big\{ - CL_n(J_{\mathcal{S}} + s^2)\frac{\log n}{2n} + (J_{\mathcal{S}} - J_{\mathcal{S}^*})\frac{\log n}{2n}C_n\big\},
\end{aligned}
$$

where the first inequality follows from $\log(1+x) \geq -2|x|$ for any $x : |x| < 1/2$. This completes the proof of (S3.5).

Now we prove (S3.6). By assumption, we can take $\eta > 0$(not depending on n) such that $\min_{k\in\mathcal{S}^*} \|\boldsymbol{\theta}_k^0\| > \sqrt{K_n}\eta$ (every B-spline covariate is $O_p(1/\sqrt{K_n})$). Let $\tilde{\mathcal{S}} = \mathcal{S} \cup \mathcal{S}^*$. Then $\tilde{\mathcal{S}} \in \mathcal{M}^{OF}$. Let's extend $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ from $\mathbb{R}^{J_{\mathcal{S}}}$ to $\mathbb{R}^{J_{\tilde{\mathcal{S}}}}$ by setting zero on elements in $\tilde{\mathcal{S}}/\mathcal{S}$. Denote the extended vector by $\hat{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}(\mathcal{S})$. Note that it's different with $\hat{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}$ which is the estimator under model $\tilde{\mathcal{S}}$. Clearly, $\|\hat{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}(\mathcal{S}) - \boldsymbol{\theta}_{\tilde{\mathcal{S}}}^0\| > \sqrt{K_n}\eta$. Accordingly, define $\hat{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}(\mathcal{S}) = \hat{\mathbf{W}}_{B,\tilde{\mathcal{S}}}(\hat{\boldsymbol{\theta}}_{\tilde{\mathcal{S}}}(\mathcal{S}) - \boldsymbol{\theta}_{\tilde{\mathcal{S}}}^0)$ and $\|\hat{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}(\mathcal{S})\| > \sqrt{n}\eta$ (from Lemma S3.1(3)). On the other hand, we have $\|\hat{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}\| \leq \sqrt{n}\eta$ from Lemma

S3.3. By the convexity of $\rho_\tau(\cdot)$, there exists $\bar{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}$ with $\|\bar{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}\| = \sqrt{n}\eta$ such that

$$\begin{aligned} & \sum_{i=1}^n \rho_\tau(y_i - \mathbf{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T \hat{\boldsymbol{\theta}}_{\mathcal{S}}) \\ = & \sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\hat{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}(\mathcal{S}) - R_i - u_i) \\ \geq & \sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\bar{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i). \end{aligned}$$

Consequently,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \mathbf{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T \hat{\boldsymbol{\theta}}_{\mathcal{S}}) - \frac{1}{n}\sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\tilde{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i) \\ \geq & \frac{1}{n}\Big[\inf_{\boldsymbol{\delta}_{\tilde{\mathcal{S}}} \in B_{\sqrt{n}\eta}(\tilde{\mathcal{S}})} \sum_{i=1}^n E[Q_i(\boldsymbol{\delta}_{\tilde{\mathcal{S}}}) - Q_i(0)|X_i] \\ & \quad - \sup_{\boldsymbol{\delta}_{\tilde{\mathcal{S}}} \in B_{\sqrt{n}\eta}(\tilde{\mathcal{S}})} |\sum_{i=1}^n [Q_i(\boldsymbol{\delta}_{\tilde{\mathcal{S}}}) - Q_i(0)] - (\sum_{i=1}^n E[Q_i(\boldsymbol{\delta}_{\tilde{\mathcal{S}}}) - Q_i(0)|X_i])| \\ & \quad - (\sum_{i=1}^n [Q_i(\hat{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}}) - Q_i(0)])\Big]. \end{aligned} \tag{S3.8}$$

Similar to arguments in Lemma S3.3, $n^{-1}\inf_{\boldsymbol{\delta}_{\tilde{\mathcal{S}}} \in B_{\sqrt{n}\eta}(\tilde{\mathcal{S}})} \sum_{i=1}^n E[Q_i(\boldsymbol{\delta}_{\tilde{\mathcal{S}}}) - Q_i(0)|X_i]$ is positive and bounded away uniformly over $\tilde{\mathcal{S}} \in \mathcal{OF}$. From Lemma S3.4, the second term converges to 0. From (S3.7), the third term converges to 0. So we can take a constant $c > 0$ not depending on $\mathcal{S}$ such that

$$\frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \mathbf{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T \hat{\boldsymbol{\theta}}_{\mathcal{S}}) - \frac{1}{n}\sum_{i=1}^n \rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\tilde{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i) \geq 2c > 0,$$

for all $\mathcal{S} \in \mathcal{S}^{UF}$ with probability tending to one. Then we have

$$\begin{aligned}
& \min_{\mathcal{S}\in\mathcal{M}^{UF}} \text{BIC}(\mathcal{S}) - \text{BIC}(\tilde{\mathcal{S}}) \\
= & \min_{\mathcal{S}\in\mathcal{M}^{UF}} \log(1 + \frac{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{W}(\hat{\boldsymbol{\zeta}}_{i,\mathcal{S}})^T\hat{\boldsymbol{\theta}}_{\mathcal{S}}) - \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\tilde{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i)}{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\tilde{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i)}) \\
& + (J_{\mathcal{S}} - J_{\tilde{\mathcal{S}}})\frac{\log n}{2n}C_n \\
\geq & \min_{\mathcal{S}\in\mathcal{M}^{UF}} \min\{\log 2, \frac{c}{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - \tilde{\boldsymbol{W}}(\hat{\boldsymbol{\zeta}}_{i,\tilde{\mathcal{S}}})\tilde{\boldsymbol{\delta}}_{\tilde{\mathcal{S}}} - R_i - u_i)}\} - |\mathcal{S}^*|K_n\frac{\log n}{2n}C_n > 0,
\end{aligned}$$

with probability tending to 1. The first inequality follows from $\log(1+x) \geq \min\{x/2, \log 2\}$ for any $x > 0$. Then we have

$$\begin{aligned}
& \min_{\mathcal{S}\in\mathcal{M}^{UF}}[\text{BIC}(\mathcal{S}) - \text{BIC}(\mathcal{S}^*)] \\
& = \min_{\mathcal{S}\in\mathcal{M}^{UF}}[\text{BIC}(\mathcal{S}) - \text{BIC}(\tilde{\mathcal{S}}) + \text{BIC}(\tilde{\mathcal{S}}) - \text{BIC}(\mathcal{S}^*)] \\
& \geq \min_{\mathcal{S}\in\mathcal{M}^{UF}}[\text{BIC}(\mathcal{S}) - \text{BIC}(\tilde{\mathcal{S}})] > 0,
\end{aligned}$$

where the first inequality comes from (S3.5). This completes the proof.

Bibliography

He, X. and Shi, P. (1994) Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, **3**, 299–308.

Lee, E. R., Noh, H. and Park, B. U. (2014) Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, **109**, 216–229.

Sherwood, B. and Wang, L. (2016) Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, **44**, 288–317.

Shi, P. and Li, G. (1995) Global convergence rates of b-spline m-estimators in nonparametric regression. *Statistica Sinica*, 303–318.

Stone, C. J. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

Wang, L., Wu, Y. and Li, R. (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, **107**, 214–222.

Wong, R. K. W., Li, Y. and Zhu, Z. (2018) Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, just–accepted.