Supplemenary Material: Proofs

Denote by $F(.|\mathbf{W},\mathbf{X})$ the conditional cumulative distribution function of the error term e given the covariates. The true conditional quantile $\mathbf{g}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}_i$ is also written as m_i for simplicity of notation. Below C denotes a generic positive constant.

Lemma 1. *Let* $r_n = \sqrt{K/n} + K^{-d}$.

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Cr_n} \sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) - \sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}) \\
&+\sum_{i=1}^n\sum_{j=1}^p (X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})(\tau - I\{e_i \le 0\}) \\
&-E\sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) + E\sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}) = o_p(nr_n^2).
\end{aligned}
$$

Proof. As argued in He and Shi (1994), without loss of generality we only need to consider median regression with $\tau = 1/2$, $\rho_\tau(u) = |u|/2$. Below we use covering arguments to achieve uniformility of bounds. Let $\mathcal{N} = \{(\boldsymbol{\beta}^{(1)},\boldsymbol{\theta}^{(1)}),\ldots,(\boldsymbol{\beta}^{(N)},\boldsymbol{\theta}^{(N)})\}$ be a δ_n covering of $\{(\boldsymbol{\beta},\boldsymbol{\theta}) : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| + \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le Cr_n\}$, with size bounded by $N \le (Cr_n/\delta_n)^{CK}$ and thus $\log N \le CK\log n$ if we choose $\delta_n \sim n^{-a}$ for some $a > 0$ (we will choose a to be large enough).

Let $M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{1}{2}|Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j| - \frac{1}{2}|Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}| + \sum_{j=1}^p(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})(1/2 - I\{e_i \le 0\})$, and $M_n(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^n M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})$. Since the absolute value function $|u|$ is Lipschitz, for any $(\boldsymbol{\beta},\boldsymbol{\theta})$ there is a $(\boldsymbol{\beta}^{(l)},\boldsymbol{\theta}^{(l)})$ that satisfies $\|\boldsymbol{\beta}-\boldsymbol{\beta}^{(l)}\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\theta}^{(l)}\|^2 \le \delta_n^2$, we have

$$
\begin{aligned}
& M_n(\boldsymbol{\beta},\boldsymbol{\theta}) - EM_n(\boldsymbol{\beta},\boldsymbol{\theta}) - M_n(\boldsymbol{\beta}^{(l)},\boldsymbol{\theta}^{(l)}) + EM_n(\boldsymbol{\beta}^{(l)},\boldsymbol{\theta}^{(l)}) \\
\le\ & C\sum_{i=1}^n\sum_{j=1}^p |\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^{(l)})\boldsymbol{\theta}_j^{(l)}| + C\sum_{i=1}^n\sum_{j=1}^p E|\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^{(l)})\boldsymbol{\theta}_j^{(l)}|,
\end{aligned}
$$

which can obviously be made $o_p(nr_n^2)$ by the uniform continuity of the spline functions, if one sets $\delta_n \sim n^{-a}$ for a sufficiently big.

Then, we easily have

$$
\begin{aligned}
|M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})| &= \Bigg|\frac{1}{2}|Y_i-\sum_{j=1}^{p}X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j|-\frac{1}{2}|Y_i-\sum_{j=1}^{p}X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}| \\
&\quad +\sum_{j=1}^{p}(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})(1/2-I\{e_i\le 0\})\Bigg| \\
&= \Bigg|\frac{1}{2}|e_i+m_i-\sum_{j=1}^{p}X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j|-\frac{1}{2}|e_i+m_i-\sum_{j=1}^{p}X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}| \\
&\quad +\sum_{j=1}^{p}(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})(1/2-I\{e_i\le 0\})\Bigg| \\
&\le |\sum_{j=1}^{p}(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})|\cdot \\
&\quad I\{|e_i|\le|\sum_{j=1}^{p}(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})| \\
&\quad +|m_i-\sum_{j=1}^{p}X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}|\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
|M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})| &\le |\sum_{j=1}^{p}(X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})| \\
&\le C\sum_j(|\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{\theta}_j\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|+|\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{0j})|) \\
&\le C\sum_j(|\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|+|\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{0j})| \\
&\quad +|\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*)(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{0j})\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|) \\
&\le C(r_n+\sqrt{K}r_n+K^{3/2}r_n^2) \\
&\le C\sqrt{K}r_n \\
&=: A,
\end{aligned}
$$

where we used that $\|\mathbf{B}(x)\|\le C\sqrt{K}$ and $\|\mathbf{B}^{(1)}(x)\|\le CK^{3/2}$ for any $x\in[a,b]$.

Furthermore, using

$$
E|M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})|^2 \le C(\sqrt{K}r_n)\sum_j E|\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}|^2
$$

$$\leq \quad C(\sqrt{K}r_n)(r_n^2) =: D^2, \tag{10}$$

application of Bernstein's inequality together with the union bound yields

$$P(\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{N}} |M_n(\boldsymbol{\beta},\boldsymbol{\theta}) - EM_n(\boldsymbol{\beta},\boldsymbol{\theta})| > a) \leq C\exp\{-\frac{a^2}{aA+nD^2} + CK\text{log}n\}.$$

It is clear that the right hand side above will converge to zero when $a = O\left(\max\{K^{3/2}r_n\text{log}n, \sqrt{nK^{3/2}r_n^3\text{log}n}\}\right.$
$o(nr_n^2)$. □

Lemma 2. *For sufficiently large $L > 0$,*

$$\begin{aligned}
&\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=Lr_n} \sum_i E\rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) \\
&\qquad - \sum_i E\rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}) \\
&\geq L^2Cnr_n^2.
\end{aligned}$$

Proof. We make use of the Knight's identity that $\rho_\tau(x-y) - \rho_\tau(x) = -y(\tau - I\{x \leq 0\}) + \int_0^y (I\{x \leq t\} - I\{x \leq 0\})dt$, which implies that

$$\begin{aligned}
&E\sum_{i=1}^n \rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) - E\sum_{i=1}^n \rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}) \\
&= E\left\{\sum_i \int_{\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}-m_i}^{\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j-m_i} F(t|\mathbf{W}_i,\mathbf{X}_i) - F(0|\mathbf{W}_i,\mathbf{X}_i)dt\right\} \\
&\geq CE\left\{\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)\left[(\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})^2\right.\right. \\
&\qquad \left.\left. +2(\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})(\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j} - m_i)\right]\right\}
\end{aligned}$$

We have, by Taylor's expansion,

$$\begin{aligned}
&\sum_i(\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})^2 \\
\geq \quad & C\sum_i\left(\sum_{j=1}^p X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_{0j})\right)^2 + o_p(nr_n^2)
\end{aligned}$$

$$\geq \quad CL^2 n r_n^2.$$

Note that we have used the eigenvalue property as in Lemma 3 stated below. Furthermore, as in (10) we can derive a corresponding upper bound

$$\sum_i (\sum_{j=1}^p X_{ij} \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}) \boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij} \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \boldsymbol{\theta}_{0j})^2 \leq CL^2 n r_n^2,$$

and using the property of polynomial splines,

$$\sum_i (\sum_{j=1}^p X_{ij} \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \boldsymbol{\theta}_{0j} - m_i)^2 \leq CnK^{-2d}. \tag{11}$$

Combining several bounds stated above, we have

$$E \sum_{i=1}^n \rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij} \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}) \boldsymbol{\theta}_j) - E\rho_\tau(e_i + m_i - \sum_{j=1}^p X_{ij} \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \boldsymbol{\theta}_{0j})$$
$$\geq CL^2 n r_n^2,$$

when L is sufficiently large. □

Lemma 3. *With probability approaching one, the eigenvalues of*

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} (\mathbf{X}_i \otimes \mathbf{B}^{(1)}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{\theta}_0 \mathbf{W}_i \\ \mathbf{X}_i \otimes \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \end{pmatrix} \left((\mathbf{X}_i \otimes \mathbf{B}^{(1)}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{\theta}_0 \mathbf{W}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \right)$$

are bounded away from zero and infinity.

Proof. Using Markov's inequality, we only need to show the population version that the eigenvalues of

$$E \left[\begin{pmatrix} (\mathbf{X} \otimes \mathbf{B}^{(1)}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{\theta}_0 \mathbf{W} \\ \mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \end{pmatrix} \left((\mathbf{X} \otimes \mathbf{B}^{(1)}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{\theta}_0 \mathbf{W}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \right) \right] \tag{12}$$

are bounded away from zero and infinity.

Since $|(\mathbf{X} \otimes \mathbf{B}^{(1)}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{\theta}_0 - \mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{X}| \leq CK^{-d+1}$, we only need to show that the eigenvalues of

$$E \left[\begin{pmatrix} \mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{X} \mathbf{W} \\ \mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \end{pmatrix} \left(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{X} \mathbf{W}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}} \boldsymbol{\beta}_0) \right) \right] \tag{13}$$

are bounded away from zero and infinity.

By (A4), we can find a $pK \times q$ matrix $\boldsymbol{\gamma}_0$ satisfying $\|E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}] - \boldsymbol{\gamma}_0^{\mathrm{T}}(\mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))\| \leq CK^{-d'}$.

Pre-multiplying (13) by

$$\begin{pmatrix} \mathbf{I} & -\boldsymbol{\gamma}_0^{\mathrm{T}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{14}$$

and post-multiplying (13) by its transposition, we obtain the new matrix

$$E\left[\begin{pmatrix} \mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W} - \boldsymbol{\gamma}_0^{\mathrm{T}}(\mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)) \\ \mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0) \end{pmatrix}^{\otimes 2}\right]. \tag{15}$$

Since it can be directly verified that singular values of (14) are bounded away from zero and infinity, we only need to show that the eigenvalues of (15) are bounded away from zero and infinity. Since $\|E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}] - \boldsymbol{\gamma}_0^{\mathrm{T}}(\mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))\| \leq CK^{-d'}$, we can replace $\boldsymbol{\gamma}_0^{\mathrm{T}}(\mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))$ with $E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}]$ in the displayed expression above. By the assumed boundedness of $f(0|\mathbf{W}, \mathbf{X})$ in (A2), we only need to consider the matrix

$$E\left[f(0|\mathbf{W}, \mathbf{X})\begin{pmatrix} \mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W} - E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}] \\ \mathbf{X} \otimes \mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0) \end{pmatrix}^{\otimes 2}\right]. \tag{16}$$

By our specific definition of the projection in the main text, $E[f(0|\mathbf{W}, \mathbf{X})(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W} - E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}])(\mathbf{X}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))] = \mathbf{0}$ and (16) is block-diagonal and it is easy to see by (A5) that the matrix has eigenvalues bounded away from zero and infinity. $\square$

Lemma 4.

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=Lr_n} \sum_i \left(\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\right)$$
$$(\tau - I\{e_i \leq 0\}) = L \cdot O_p(nr_n^2).$$

Proof. For simplicity of presentation below, we denote $\epsilon_i = \tau - I\{e_i \le 0\}$. We have

$$\begin{aligned}
& \sum_i (\sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})\epsilon_i \\
= & \sum_i \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i \qquad (17) \\
& + \sum_i \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j})\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i \qquad (18) \\
& + \sum_i \sum_j X_{ij}(\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*) - \mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i \qquad (19) \\
& + \sum_i \sum_j X_{ij}(\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*) - \mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j})\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i \qquad (20) \\
& + \sum_i \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j})\epsilon_i. \qquad (21)
\end{aligned}$$

The term (17) obviously has order $L{\cdot}O_p(\sqrt{n}r_n)$. For (21), we have that $\|\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\epsilon_i\|^2 = O_p(\sum_i \|\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\|^2) = O_p(nK)$ and thus (21) is $O_p(\sqrt{nK}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|) = L\cdot O_p(\sqrt{nK}r_n)$.

For the term (18), since $\|\sum_i \mathbf{B}^{(1)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\epsilon_i\|^2 = O_p(\sum_i \|\mathbf{B}^{(1)}(\mathbf{W}_i^{T}\boldsymbol{\beta}_0)\|^2) = O_p(nK^3)$ we have

$$\sum_i \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j})\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i = O_p(\sqrt{n}K^{3/2}r_n^2) = o_p(nr_n^2).$$

With further Taylor expansion $\mathbf{B}^{(1)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*) - \mathbf{B}^{(1)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0) = \mathbf{B}^{(2)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^{**})\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)$, (19) and (20) are also of order $o_p(nr_n^2)$ and the proof is complete. □

Proof of Theorem 1. Combining Lemmas stated above, we get that

$$P\left(\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=Lr_n} \sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) > \sum_{i=1}^n \rho_\tau(Y_i - \sum_{j=1}^p X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j})\right) \to 1,$$

and thus there is a local minimizer of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ with $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(r_n)$. □

Next we try to establish asymptotic normality of $\widehat{\boldsymbol{\beta}}$. Orthogonality step plays an important role in this part. Due to the more complicated model structure here, this procedure is more complicated than partially linear models considered in some previous works (Wang et al., 2009, 2011).

Let $\mathbf{\Pi}_i = \mathbf{X}_i \otimes \mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)$ and let $\mathbf{\Pi}$ be the $n \times (pK)$ matrix with rows $\mathbf{\Pi}_i^{\mathrm{T}}$. The empirical counterpart of the previously defined projection is

$$\min_{\boldsymbol{\theta}} \sum_i f(0|\mathbf{W}_i, \mathbf{X}_i)(V_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\theta})^2,$$

with the minimizer $(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{V}$ where $\mathbf{\Gamma}$ is the diagonal matrix containing $f(0|\mathbf{W}_i, \mathbf{X}_i)$ as its diagonal entries, and $\mathbf{V} = (V_1, \ldots, V_n)^{\mathrm{T}}$. Define $\mathbf{P} = \mathbf{\Pi}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}$.

We write

$$\begin{aligned}
& \rho_\tau(e_i + m_i - \sum_j X_{ij}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) \\
= \; & \rho_\tau(e_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j}) - \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta})) \\
= \; & \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta})),
\end{aligned}$$

where we defined $\mathbf{U}_i = \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i$ and

$$\begin{aligned}
& R_i(\boldsymbol{\beta}, \boldsymbol{\theta}) \\
= \; & \sum_j X_{ij}(\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}) - \mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))^{\mathrm{T}}\boldsymbol{\theta}_j - \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
& + (\sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j} - m_i) \\
=: \; & R_{i1}(\boldsymbol{\beta}, \boldsymbol{\theta}) + R_{i2}(\boldsymbol{\beta}, \boldsymbol{\theta}).
\end{aligned}$$

Let $\mathbf{V} = \mathbf{U} - \mathbf{P}\mathbf{U}$ with the i-th row of $\mathbf{V}$ denoted by $\mathbf{V}_i^{\mathrm{T}} = \mathbf{U}_i^{\mathrm{T}} - \mathbf{P}_i^{\mathrm{T}}\mathbf{U}$. We further write

$$\begin{aligned}
& \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta})) \\
= \; & \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}((\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)) - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta})) \\
= \; & \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta})),
\end{aligned}$$

with $\boldsymbol{\eta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0 + (\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. We require the following lemma which is a refinement of Lemma 1.

Lemma 5.

$$\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C/\sqrt{n}, \|\boldsymbol{\eta}\| \le Cr_n} \Bigg| \sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta}))$$

$$-\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})) + \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i$$

$$\left. -E\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})) + E\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}))\right| = o_p(1).$$

Proof. Same as for Lemma 1, we assume $\tau = 1/2$ here for simplicity. We have

$$\begin{aligned}
R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \sum_j X_{ij}(\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}) - \mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))^{\mathrm{T}}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j}) \\
&\quad -\sum_j X_{ij}(\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_{0j} - \mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j} - \mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)) \\
&= \sum_j X_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j})(\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)) \\
&\quad -\sum_j X_{ij}\mathbf{B}^{(2)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{\theta}_{0j}(\mathbf{W}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))^2
\end{aligned}$$

It is easy to see that $|R_{i1}| \leq C\sqrt{K^3/n}r_n$ and $\sum_i R_{i1}^2 = O_p(r_n^2K^3 + 1/n) = O_p(r_n^2K^3)$.

For fixed $\boldsymbol{\eta},\boldsymbol{\beta}$, let $M_{ni}(\boldsymbol{\beta},\boldsymbol{\eta}) = \frac{1}{2}|e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})| - \frac{1}{2}|e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})| + (\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta}))(1/2 - I\{e_i \leq 0\})$, we have

$$\begin{aligned}
& |M_{ni}(\boldsymbol{\beta},\boldsymbol{\eta})| \\
\leq\ & |\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta})| I\{|e_i| \leq |\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta})| + |\mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} + R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})|\} \\
\leq\ & C(1/\sqrt{n} + \sqrt{K^3/n}r_n),
\end{aligned}$$

and

$$E|M_{ni}(\boldsymbol{\beta},\boldsymbol{\eta})|^2 \leq (1/n + K^3r_n^2/n)(\sqrt{K}r_n).$$

The rest of the proof is similar to the proof of Lemma 1. In particular, using a covering argument with Bernstein's inequality, we have that

$$\begin{aligned}
\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C/\sqrt{n}, \|\boldsymbol{\eta}\|\leq Cr_n} &\left|\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta}))\right. \\
&-\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})) + \sum_i (\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta}))\epsilon_i \\
&\left. -E\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})) + E\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}))\right|
\end{aligned}$$

is of order $O_p(\max\{K\log n(1/\sqrt{n}+\sqrt{K^3/n}r_n), \sqrt{(1+K^3r_n^2)(\sqrt{K}r_nK\log n}) = o_p(1)$.

Finally, using the above bounds for R_{i1} and similar arguments, we have $\sum_i R_{i1}(\boldsymbol{\beta},\boldsymbol{\theta})(1/2-I\{e_i\leq 0\}) = o_p(1)$ which completes the proof. $\square$

Lemma 6.

$$\sup_{\|\boldsymbol{\eta}\|\leq Cr_n, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C/\sqrt{n}} \left| \sum_i E\rho_\tau(e_i - \mathbf{\Pi}_i\boldsymbol{\eta} - \mathbf{V}_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})) \right.$$
$$\left. -\sum_i E\rho_\tau(e_i - \mathbf{\Pi}_i\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})) - \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\mathbf{V}_i\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right| = o_p(1).$$

Proof. By Knight's identity,

$$\begin{aligned}
&\sum_i E\rho_\tau(e_i - \mathbf{\Pi}_i\boldsymbol{\eta} - \mathbf{V}_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})) - \sum_i E\rho_\tau(e_i - \mathbf{\Pi}_i\boldsymbol{\eta} - R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})) \\
=& \int_{\mathbf{\Pi}_i\boldsymbol{\eta}+R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})}^{\mathbf{\Pi}_i\boldsymbol{\eta}+R_i(\boldsymbol{\beta},\boldsymbol{\theta})+\mathbf{V}_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)} F(t|\mathbf{X}_i,\mathbf{Z}_i) - F(0|\mathbf{X}_i,\mathbf{Z}_i)dt \\
=& \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}\left\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\mathbf{V}_i\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_{i1}^2 + 2R_{i1}\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2R_{i1}(\mathbf{\Pi}_i\boldsymbol{\eta}+R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}))\right. \\
&\left. +2(\mathbf{\Pi}_i\boldsymbol{\eta}+R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}))\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}(1+o_p(1)).
\end{aligned}$$

We have $\sum_i R_{i1}^2 = O(r_n^2K^3) = o_p(1)$, $(\sum_i R_{i1}\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))^2 = O_p(r_n^2K^3) = o_p(1)$, and $(\sum_i R_{i1}(\mathbf{\Pi}_i\boldsymbol{\eta}+R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})))^2 = (r_n^2K^3)(nr_n^2) = o_p(1)$. By the defined orthogonalization step, $\sum_i f(0|\mathbf{X}_i,\mathbf{Z}_i)\mathbf{\Pi}_i\mathbf{V}_i^{\mathrm{T}} = \sum_i f(0|\mathbf{X}_i,\mathbf{Z}_i)\mathbf{\Pi}_i(\mathbf{U}_i - \mathbf{\Pi}_i^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{U}) = \mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U} - \mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U} = \mathbf{0}$. Thus we only need to show

$$\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) = o_p(1),$$

with $R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}) = \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j} - m_i$.

Note that directly using $|R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})|\leq CK^{-d}$ shows that the above displayed equation is of order $O_p(\sqrt{n}K^{-d})\neq o_p(1)$ in general. So we need to use a different strategy based on finer analysis.

Write $\mathbf{V}_i^{\mathrm{T}} = (\mathbf{U}_i^{\mathrm{T}} - \mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}) + (\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}} - E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}]) + (E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}] - \mathbf{P}_i^{\mathrm{T}}\mathbf{U}))$, and we deal with each one of the three terms below.

By the approximation property of splines,

$$\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})(\mathbf{U}_i^{\mathrm{T}}-\mathbf{g}^{(1)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)=O_p(\sqrt{n}K^{-2d+1})=o_p(1).$$

Then, using the definition of projection, we have $E[f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}-E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}])]=0$. Thus by a simple variance calculation, we get

$$\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}-E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}])(\boldsymbol{\beta}-\boldsymbol{\beta}_0)=O_p(K^{-d})=o_p(1).$$

Finally, using (A4), we have $\|E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}]-\mathbf{P}_i^{\mathrm{T}}\mathbf{U}\|=O_p(K^{-d'}+K^{-d+1})$ and thus

$$\begin{aligned}&\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})(E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}_i^{\mathrm{T}}]-\mathbf{P}_i^{\mathrm{T}}\mathbf{U})\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\=\;&O_p(\sqrt{n}K^{-d-d'}+\sqrt{n}K^{-2d+1})=o_p(1).\end{aligned}$$

Thus $\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)R_i(\boldsymbol{\beta}_0,\boldsymbol{\theta})\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)=o_p(1)$ and the proof is complete. □.

Lemma 7.

$$\frac{1}{n}\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)\mathbf{V}_i\mathbf{V}_i^{\mathrm{T}}\to E[f(0|\mathbf{W},\mathbf{X})(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}-E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}])^{\otimes 2}]\ \textit{in probability},$$

$$\frac{1}{n}\sum_i \mathbf{V}_i\mathbf{V}_i^{\mathrm{T}}\to E[(\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}-E_{\mathcal{M}}[\mathbf{g}^{(1)\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}\mathbf{W}])^{\otimes 2}]\ \textit{in probability.}$$

Proof. The left hand side is $\mathbf{V}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{V}/n=\mathbf{U}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}^{\mathrm{T}})\boldsymbol{\Gamma}(\mathbf{I}-\mathbf{P})\mathbf{U}/n$ where the rows of $\mathbf{U}$ are $\mathbf{U}_i^{\mathrm{T}}=\sum_j Z_{ij}\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}\mathbf{W}_i^{\mathrm{T}}$. Let $\mathbf{U}^*$ be defined similarly as $\mathbf{U}$ with $\mathbf{B}^{(1)\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_{0j}$ replaced by $g_j^{(1)}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)$. By the approximation property of splines $\left\|(1/n)\mathbf{U}^{*T}(\mathbf{I}-\mathbf{P}^{\mathrm{T}})\boldsymbol{\Gamma}(\mathbf{I}-\mathbf{P})\mathbf{U}^*-(1/n)\mathbf{U}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}^{\mathrm{T}})\boldsymbol{\Gamma}(\mathbf{I}-\mathbf{P})\mathbf{U}\right\|_F=o_p(1)$ and then using the same arguments as in Lemma 1 of Wang et al. (2009). The second expression is proved in the same way. □

Proof of Theorem 2. Let $\widehat{\boldsymbol{\eta}}=\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0+(\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{U}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. By Lemmas 5, 6, and 7,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C/\sqrt{n}}\left|\sum_i\rho_\tau(e_i-\boldsymbol{\Pi}_i^{\mathrm{T}}\widehat{\boldsymbol{\eta}}-\mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}))\right.$$

$$-\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}} - R_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}})) + \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\epsilon_i$$
$$-\frac{n}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{\Phi}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\Big| = o_p(1). \tag{22}$$

Let $Q(\boldsymbol{\beta}) = \frac{n}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{\Phi}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\epsilon_i$ and define $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (1/n)\mathbf{\Phi}^{-1}\sum_i \mathbf{V}_i^{\mathrm{T}}\epsilon_i$. We have by central limit theorem

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{\Phi}^{-1}\mathbf{\Sigma}\mathbf{\Phi}^{-1}).$$

Note $\tilde{\boldsymbol{\beta}}$ is the minimizer of $Q(\boldsymbol{\beta})$, which has a quadratic form $(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{\Phi}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$ plus a term that is independent of $\boldsymbol{\beta}$. Define

$$\tilde{\tilde{\boldsymbol{\beta}}} := \underset{\|\boldsymbol{\beta}\|=1, \beta_1>0}{\arg\min} \ (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{\Phi}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}).$$

By Proposition 4.1 of Shapiro (1986) which works for overparametrized models (considering $\boldsymbol{\beta}$ as a function of $\boldsymbol{\beta}^{(-1)}$ and the parametrization using $\boldsymbol{\beta}$ is an overparametrization), we get that

$$\sqrt{n}(\tilde{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{\Phi}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{\Phi}\mathbf{J})^{-1})\mathbf{J}^{\mathrm{T}}).$$

Given any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$ and $\|\boldsymbol{\beta} - \tilde{\tilde{\boldsymbol{\beta}}}\| = \delta/\sqrt{n}$ for a sufficiently small $\delta > 0$, due to that Q being quadratic, we obtain

$$Q(\boldsymbol{\beta}) - Q(\tilde{\tilde{\boldsymbol{\beta}}}) \geq C\delta^2$$

and thus by (22),

$$P(\sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}} - \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}})) > \sum_i \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}} - \mathbf{V}_i^{\mathrm{T}}(\tilde{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\beta}_0) - R_i(\tilde{\tilde{\boldsymbol{\beta}}}, \hat{\boldsymbol{\theta}})) \to 1.$$

By the arbitrariness of δ, we get $\|\hat{\boldsymbol{\beta}} - \tilde{\tilde{\boldsymbol{\beta}}}\| = o_p(1/\sqrt{n})$ which finishes the proof. $\square$

Lemma 8. *Let $r_n = \sqrt{K/n} + K^{-d}$ and $r_n' = r_n/\sqrt{K}$. When $K^3 log^2 n/n \to 0$ and $K^{d+3/2} logn/n \to 0$, we have*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq Cr_n', \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq C/\sqrt{n}} \Big| \sum_{i=1}^n \rho_\tau(e_i - \mathbf{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta}, \boldsymbol{\theta}))$$

$$
\begin{aligned}
& -\sum_{i=1}^{n}\rho_\tau(e_i-\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)) \\
& +\sum_{i=1}^{n}\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\tau-I\{e_i\le 0\}) \\
& -E\sum_{i=1}^{n}\rho_\tau(e_i-\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\boldsymbol{\theta})) \\
& +E\sum_{i=1}^{n}\rho_\tau(e_i-\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0))\Bigg| \\
= \;& o_p(K^{-1}nr_n^2).
\end{aligned}
$$

Proof of Lemma 1. The proof is similar to that of Lemma 1 and Lemma 5. With abuse of notation, let $M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})=\frac{1}{2}|e_i-\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\boldsymbol{\theta})|-\frac{1}{2}|e_i-\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)|+(\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta}))(1/2-I\{e_i\le 0\})$, and $M_n(\boldsymbol{\theta})=\sum_{i=1}^n M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})$, where $\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})=R_i(\boldsymbol{\beta},\boldsymbol{\theta})-R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)$. We have by easy calculations $|\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})|\le C\sqrt{K^3/n}r_n'$ and $\sum_{i=1}^n\breve{R}_i^2(\boldsymbol{\beta},\boldsymbol{\theta})\le CK^3(r_n')^2$, which leads to

$$
\begin{aligned}
& |M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})| \\
\le\;& |\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})|\cdot I\{|e_i|\le|\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})|+|\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)|\} \\
\le\;& C\sqrt{K}r_n',
\end{aligned}
$$

$$
E|M_{ni}(\boldsymbol{\beta},\boldsymbol{\theta})|^2 \;\le\; C(\sqrt{K}r_n')((r_n')^2),
$$

and

$$
\sum_i\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})(1/2-I\{e_i\le 0\})=o_p((n/K)r_n^2).
$$

Using a fine covering as in Lemma 1 and Lemma 5, with Bernstein's inequality, the left hand side in the statement of the current lemma can be shown to be of order $O_p\left(\max\{K^{3/2}r_n'\log n,\sqrt{nK^{3/2}(r_n')^3\log n}\}\right)=o_p((n/K)r_n^2)$ by the more stringent conditions on K. □

Lemma 9.

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Cr'_n, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C/\sqrt{n}} \Bigg| \sum_i E\rho_\tau(e_i - \boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta}))$$
$$- \sum_i E\rho_\tau(e_i - \mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)) - \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$
$$- \sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0))\Bigg| = o_p((n/K)r_n^2).$$

Proof. The proof is similar to that of Lemma 6, with the difference being that since we do not use orthogonalization here some terms are no longer ignorable as appeared in the statement of the current lemma. Similar to the calculations in Lemma 6, we have

$$\begin{aligned}
& \sum_i E\rho_\tau(e_i - \boldsymbol{\Pi}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \mathbf{U}_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta})) - \sum_i E\rho_\tau(e_i - \boldsymbol{\Pi}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)) \\
= & \int_{\boldsymbol{\Pi}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)}^{\boldsymbol{\Pi}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\mathrm{U}_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+R_i(\boldsymbol{\beta},\boldsymbol{\theta})} F(t|\mathbf{W}_i,\mathbf{X}_i) - F(0|\mathbf{W}_i,\mathbf{X}_i)dt \\
= & \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}\Big\{(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \breve{R}_i^2(\boldsymbol{\beta},\boldsymbol{\theta}) + 2\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\
& + 2(\boldsymbol{\Pi}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta}))(\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0))\Big\}(1+o_p(1)).
\end{aligned}$$

Using $|\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})| \le C\sqrt{K^3/n}r'_n$ and $\sum_{i=1}^n \breve{R}_i^2(\boldsymbol{\beta},\boldsymbol{\theta}) \le CK^3(r'_n)^2$, all terms above involving $\breve{R}_i(\boldsymbol{\beta},\boldsymbol{\theta})$ are $o_p((n/K)r_n^2)$, which proves the Lemma. □

Proof of Theorem 3. Now define

$$\begin{aligned}
\boldsymbol{\theta}^* \quad := \quad & \arg\min_{\mathrm{rank}(\boldsymbol{\Theta})\le r} \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\
& + \sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0)) \\
& - \sum_i \boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\tau - I\{e_i \le 0\})
\end{aligned}$$

and

$$\boldsymbol{\theta}^{**} \quad := \quad \arg\min_{\boldsymbol{\theta}} \sum_i \frac{f(0|\mathbf{W}_i,\mathbf{X}_i)}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

$$
\begin{aligned}
&+\sum_i f(0|\mathbf{W}_i,\mathbf{X}_i)\mathbf{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+R_i(\boldsymbol{\beta},\boldsymbol{\theta}_0))\\
&-\sum_i \mathbf{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\tau-I\{e_i\le 0\})
\end{aligned}
$$

(the latter does not have rank constraint). We have obviously

$$
\boldsymbol{\theta}^{**}-\boldsymbol{\theta}_0=(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{\Pi}\boldsymbol{\epsilon}-\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{R}), \tag{23}
$$

where $\mathbf{\Gamma}=\mathrm{diag}\{f(0|\mathbf{W}_1,\mathbf{X}_1),\ldots,f(0|\mathbf{W}_n,\mathbf{X}_n)\}$, $\mathbf{R}=(R_1(\boldsymbol{\beta},\boldsymbol{\theta}_0),\ldots,R_n(\boldsymbol{\beta},\boldsymbol{\theta}_0))^{\mathrm{T}}$ and $\boldsymbol{\epsilon}=(\tau-I\{e_1\le 0\},\ldots,\tau-I\{e_n\le 0\})^{\mathrm{T}}$.

We note that the expression on the right hand side in the definition of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{**}$ can actually be written as

$$
\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{**})^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})(\boldsymbol{\theta}-\boldsymbol{\theta}^{**})+\text{terms not involving }\boldsymbol{\theta}.
$$

For a general $\mathbf{x}$, let

$$
\boldsymbol{\theta}^*(\mathbf{x})=\mathop{\arg\min}_{\mathrm{rank}(\boldsymbol{\Theta})\le r}\frac{1}{2}(\boldsymbol{\theta}-\mathbf{x})^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})(\boldsymbol{\theta}-\mathbf{x}).
$$

Then by our definition of $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^*$ can thus be regarded as a function of $\boldsymbol{\theta}^{**}$, denoted by $\boldsymbol{\theta}^*(\boldsymbol{\theta}^{**})$. Since $\boldsymbol{\Theta}_0$ has rank bounded by r, we can write $\boldsymbol{\Theta}_0=\mathbf{D}_0\mathbf{E}_0^{\mathrm{T}}$ for a $p\times r$ matrix $\mathbf{D}_0$ and a $K\times r$ matrix $\mathbf{E}_0$. The Jacobian of this parametrization of $\boldsymbol{\theta}_0$ is defined to be

$$
\mathbf{\Delta}=\left.\frac{\partial\boldsymbol{\theta}}{\partial(\mathrm{vec}^{\mathrm{T}}(\mathbf{D}),\mathrm{vec}^{\mathrm{T}}(\mathbf{E}))}\right|_{\mathbf{D}=\mathbf{D}_0,\mathbf{E}=\mathbf{E}_0}.
$$

By Proposition 3.1 in Shapiro (1986), the Jacobian matrix for the function $\boldsymbol{\theta}^*(.)$ is

$$
\mathbf{J}_{\boldsymbol{\theta}}=\left.\frac{\partial\boldsymbol{\theta}^*(\mathbf{x})}{\partial\mathbf{x}}\right|_{\mathbf{x}=\boldsymbol{\theta}^{**}}=\mathbf{\Delta}(\mathbf{\Delta}^{\mathrm{T}}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi}\mathbf{\Delta})^{-}\mathbf{\Delta}^{\mathrm{T}}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi},
$$

where $(.)^{-}$ denotes the Moore-Penrose inverse.

Next we show the asymptotic normality of $\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\theta}_j^{**}=(\mathbf{e}_j\otimes\mathbf{B}(x))^{\mathrm{T}}\boldsymbol{\theta}^{**}$, where $\mathbf{e}_j$ is the unit vector with j-th entry 1 and others 0. For the term $(\mathbf{e}_j\otimes\mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\boldsymbol{\epsilon}$, its obviously its conditional variance is $\tau(1-\tau)(\mathbf{e}_j\otimes\mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi})(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{e}_j\otimes$

$\mathbf{B}(x)) \asymp K/n$. One can verify that the Lindeberg-Feller condition holds and thus we get

$$\frac{(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\boldsymbol{\epsilon}}{\left(\tau(1-\tau)(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi})(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{e}_j \otimes \mathbf{B}(x))\right)^{1/2}} \xrightarrow{d} N(0,1),$$

using arguments similar to that used in Theorem 3.1 of Zhou et al. (1998). When $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(1/\sqrt{n})$, we have

$$\begin{aligned}
& (\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
= \; & O_p(\|(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\|\|\mathbf{U}\|\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|) \\
= \; & O_p(\frac{1}{\sqrt{n}} \cdot \sqrt{n} \cdot \frac{1}{\sqrt{n}}) = O_p(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Furthermore, write $\mathbf{R} = \mathbf{R}_1 + \mathbf{R}_2$ where $\mathbf{R}_1 = (R_{11}(\boldsymbol{\beta}, \boldsymbol{\theta}_0), \ldots, R_{n1}(\boldsymbol{\beta}, \boldsymbol{\theta}_0))^{\mathrm{T}}$ and $\mathbf{R}_2 = (R_{12}(\boldsymbol{\beta}, \boldsymbol{\theta}_0), \ldots, R_{n2}(\boldsymbol{\beta}, \boldsymbol{\theta}_0))^{\mathrm{T}}$. $\|\mathbf{R}_1\|^2 = O_p(1/n)$ is sufficiently small such that it directly leads to

$$\begin{aligned}
& (\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{R}_1 \\
= \; & O_p(n^{-3/2}) = o_p(\sqrt{\frac{K}{n}}).
\end{aligned}$$

On the other hand, by our definition of $\mathbf{\Theta}_0$, $\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{R}_2$ has mean zero, and thus

$$\begin{aligned}
& E[|(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{R}_2|^2] \\
= \; & O_p(\frac{1}{nK^{2d}}) = o_p(\frac{K}{n}).
\end{aligned}$$

Thus the dominating term in $(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\boldsymbol{\theta}^{**} - \boldsymbol{\theta}_0)$ is $(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{\Pi}\boldsymbol{\epsilon}$ and

$$\frac{(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\boldsymbol{\theta}^{**} - \boldsymbol{\theta}_0)}{\left(\tau(1-\tau)(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi})(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{e}_j \otimes \mathbf{B}(x))\right)^{1/2}} \to N(0,1).$$

Thinking $\boldsymbol{\theta}^*$ as a function of $\boldsymbol{\theta}^{**}$, delta method implies that we also have asymptotic normality for $\boldsymbol{\theta}^*$:

$$\frac{(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)}{\left(\tau(1-\tau)(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Pi})(\mathbf{\Pi}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Pi})^{-1}\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{T}}(\mathbf{e}_j \otimes \mathbf{B}(x))\right)^{1/2}} \to N(0,1).$$

When $Kn^{-1/(2d+1)} \to \infty$, the bias in estimating g_j is dominated by its standard deviation, and thus

$$\frac{(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}\boldsymbol{\theta}^* - \beta_j(x)}{\left(\tau(1-\tau)(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Pi})^{-1}(\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Pi})(\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Pi})^{-1}\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{T}}(\mathbf{e}_j \otimes \mathbf{B}(x))\right)^{1/2}} \to N(0,1).$$

Denote

$$\begin{aligned} Q(\boldsymbol{\beta},\boldsymbol{\theta}) \;=\; & -\sum_{i=1}^{n} \boldsymbol{\Pi}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\tau - I\{e_i \le 0\}) \\ & + E\sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) \\ & - E\sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j^*). \end{aligned}$$

For a small $\delta > 0$, using Lemma 8, we get

$$\begin{aligned} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \le \delta(n^{-1/2}+K^{-d-1/2}), \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le C/\sqrt{n}} & \left| \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) \right. \\ & \left. - \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j^*) - [Q(\boldsymbol{\beta},\boldsymbol{\theta}) - Q(\boldsymbol{\beta},\boldsymbol{\theta}^*)] \right| = o_p((n/K)r_n^2). \end{aligned}$$

Since $Q(\boldsymbol{\beta},\boldsymbol{\theta})$ is approximately quadratic, we have that when $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| = \delta(n^{-1/2} + K^{-d-1/2})$,

$$|Q(\boldsymbol{\beta},\boldsymbol{\theta}) - Q(\boldsymbol{\beta},\boldsymbol{\theta}^*)| \ge Cn\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 - o_p((n/K)r_n^2) > 0.$$

This yields

$$\begin{aligned} \lim_{n\to\infty} P\Bigg\{ & \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| = \delta(n^{-1/2}+K^{-d-1/2}), \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le C/\sqrt{n}} \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j) \\ & > \sum_{i=1}^{n} \rho_\tau(Y_i - \sum_j X_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}_j^*) \Bigg\} = 1. \end{aligned}$$

Thus

$$P\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \ge \delta(n^{-1/2} + K^{-d-1/2})\},$$

converges to zero and we deduce that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = o_p(n^{-1/2} + K^{-d-1/2})$, which implies $(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}\widehat{\boldsymbol{\theta}}$ has the same asymptotic distribution as $(\mathbf{e}_j \otimes \mathbf{B}(x))^{\mathrm{T}}\boldsymbol{\theta}^*$. $\square$

Proof of Theorem 4. In the proof, the true rank of the matrix $\boldsymbol{\Theta}_0$ is denoted by r_0, while we use r to denote a generic value for rank that can vary.

For any given r, let $\boldsymbol{\Theta}_r$ be the minimizer of

$$\min_{\mathrm{rank}(\boldsymbol{\Theta}) \le r} E[\rho_\tau(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)].$$

Denote the estimator of $\boldsymbol{\Theta}_r$ as $\widehat{\boldsymbol{\Theta}}_r$, which minimizes $\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f))$ also with the rank constraint. In the following, we consider two cases to finish the proof.

Case 1. ($r < r_0$, underfitted model) We first prove

$$\mathrm{E}[\rho_\tau(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}_r\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))] - \mathrm{E}[\rho_\tau(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}_0\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0))] \tag{24}$$

is bounded away from zero. By Knight's identity, we have

$$\begin{aligned}
& \rho_\tau(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}_r\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)) - \rho_\tau(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}_0\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)) \\
= & \mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)[I\{e \le \delta\} - \tau] + \textstyle\int_0^{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)}[I\{e \le \delta + t\} - I\{e \le \delta\}]dt, \\
= & \mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)[I\{e \le 0\} - \tau] + \mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)[I\{e \le \delta\} - I\{e \le 0\}] \\
& + \textstyle\int_0^{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)}[I\{e \le \delta + t\} - I\{e \le \delta\}]dt,
\end{aligned} \tag{25}$$

where $\delta = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}_0\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0) - m$ and $m = \mathbf{g}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}$ is the τ-th conditional quantile of Y given $\mathbf{W}$ and $\mathbf{X}$.

The first term in (25) obviously has mean zero. By taking an iterated expectation conditioning on $\mathbf{W}, \mathbf{X}$ first, the second term in (25) satisfies

$$E\{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)[I\{e \le \delta\} - I\{e \le 0\}]\} = O(\|\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0\|K^{-d}).$$

For the third term in (25), note that conditionally on $\mathbf{X}$ and $\mathbf{W}$, by (A2), we can find a neighborhood around zero on which the conditional density value of e is positive. Thus we have

$$\mathrm{E}\int_0^{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)}[I\{e \le \delta + t\} - I\{e \le \delta\}]dt$$

$$
\begin{aligned}
&= \mathrm{E}\int_0^{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r-\boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)}[F(\delta+t|\mathbf{W},\mathbf{X})-F(\delta|\mathbf{W},\mathbf{X})]dt \\
&\geq CE[\{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\Theta}_r-\boldsymbol{\Theta}_0)\mathbf{B}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^2] \geq C\|\boldsymbol{\Theta}_r-\boldsymbol{\Theta}_0\|^2.
\end{aligned}
$$

By assumption (B), when $r < r_0$, $\|\boldsymbol{\Theta}_r - \boldsymbol{\Theta}_0\|$ is bounded away from zero, and thus expectation of the third term in (25) dominates those of other terms and as a result we have (24).

By following the proof of Theorem 1, even with underfitted models, we still have $\|\widehat{\boldsymbol{\Theta}}_r - \boldsymbol{\Theta}_r\| = O_p(r_n)$, where $r_n = \sqrt{K/n} + K^{-d}$. We can also get, following the proof of Lemmas 1 and 2,

$$
\sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f)) - \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)) = O_p(nr_n^2). \quad (26)
$$

For the SIC, using (26), we can write

$$
\begin{aligned}
&\mathrm{SIC}(r) - \mathrm{SIC}(r_0) \\
= \;& \log\left(1 + \frac{\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f))/n - \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f))/n}{\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f))/n}\right) + O\left(\frac{K\log n}{n}\right) \\
= \;& \log\left(1 + \frac{\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n - \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n + O_p(r_n^2)}{\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n + O_p(r_n^2)}\right) \\
&+ O\left(\frac{K\log n}{n}\right).
\end{aligned}
$$

Applying the law of large number, we have that

$$
\textstyle\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n - \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n
$$

and

$$
\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n
$$

are both bounded away from zero, which leads to that

$$
P\{\mathrm{SIC}(r) > \mathrm{SIC}(r_0)\} \to 1 \quad \text{for} \quad \text{any} \quad r < r_0. \quad (27)
$$

Case 2. ($r > r_0$, overfitted model) By the same arguments as in (26), we have

$$
\mathrm{SIC}(r) - \mathrm{SIC}(r_0)
$$

$$
\begin{aligned}
= \quad & \log\left(1+\frac{\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\widehat{\mathbf{\Theta}}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f)))/n-\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\widehat{\mathbf{\Theta}}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f)))/n}{\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\widehat{\mathbf{\Theta}}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_f)))/n}\right) \\
& +(r-r_0)(p+K-r-r_0)\frac{\log n}{2n} \\
= \quad & \log\left(1+\frac{\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}_r\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n-\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n+O_p(r_n^2)}{\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}_{r_0}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))/n+O_p(r_n^2)}\right) \\
& +(r-r_0)(p+K-r-r_0)\frac{\log n}{2n}.
\end{aligned}
$$

In the overfitting case, we note that $\mathbf{\Theta}_r=\mathbf{\Theta}_{r_0}$, which implies

$$
\mathrm{SIC}(r)-\mathrm{SIC}(r_0)=O_p\left(\frac{K}{n}\right)+(r-r_0)(p+K-r-r_0)\frac{\log n}{2n},
$$

and thus

$$
P\left\{\mathrm{SIC}(r)>\mathrm{SIC}(r_0)\right\}\to 1 \quad \text{for} \quad \text{any} \quad r>r_0. \tag{28}
$$

$\square$

Proof of Theorem 5. First, we can show the existence of a r_n-consistent local minimizer for the penalized optimization problem (4). Using Lemma 1, similar to the proof of Theorem 1, we can show that for $\|\mathbf{\Theta}-\mathbf{\Theta}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|=Lr_n$ with $L>0$ sufficiently large,

$$
\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})>\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}_0\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0)
$$

with probability approaching one.

Considering the penalty terms, for $j\le q$, with $\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_{0j}\|\le Lr_n$, using the concrete form of the SCAD penalty, we have $p_\lambda(\|\boldsymbol{\theta}_j\|_{\mathbf{A}_j})=p_\lambda(\|\boldsymbol{\theta}_{0j}\|_{\mathbf{A}_j})$ since $\lambda=o(1)$ and both $\|\boldsymbol{\theta}_j\|$ and $\|\boldsymbol{\theta}_{0j}\|$ are bounded away from zero. On the other hand, when $j>q$, we have $p_\lambda(\|\boldsymbol{\theta}_j\|_{\mathbf{A}_j})\ge p_\lambda(\|\boldsymbol{\theta}_{0j}\|_{\mathbf{A}_j})=0$. Combining the two cases above, we get

$$
n\sum_j p_\lambda(\|\boldsymbol{\theta}_j\|_{\mathbf{A}_j})\ge n\sum_j p_\lambda(\|\boldsymbol{\theta}_{0j}\|_{\mathbf{A}_j}).
$$

Thus, we get that, uniformly for $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=Lr_n$ with L sufficiently large,

$$
\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}))+n\sum_j p_\lambda(\|\boldsymbol{\theta}_j\|_{\mathbf{A}_j})>\sum_i \rho_\tau(Y_i-\mathbf{X}_i^{\mathrm{T}}\mathbf{\Theta}_0\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta}_0))+n\sum_j p_\lambda(\|\boldsymbol{\theta}_{0j}\|_{\mathbf{A}_j}).
$$

This implies the existence of a r_n-consistent local minimizer.

The next step is to show that this r_n-consistent local minimizer, denoted as $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\theta}})$, satisfies part (i) of Theorem 5. We prove this fact by contradition. If (i) is not true, we can assume $\widetilde{\boldsymbol{\theta}}_{j^*} \neq 0$ for some $j^* > q$. We define $\widetilde{\boldsymbol{\theta}}^*$ to be the same as $\widetilde{\boldsymbol{\theta}}$, but we replace $\widetilde{\boldsymbol{\theta}}_{j^*}$ by $\widetilde{\boldsymbol{\theta}}^*_{j^*} = 0$. Due to the check loss function is convex, we have $\rho_\tau(x) - \rho_\tau(y) \geq (\tau - I\{y \leq 0\})(x-y)$, implying that

$$\begin{aligned}
& \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})) - \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}}^* \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})) \\
\geq\ & -\sum_i (\tau - I\{Y_i \leq \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})\}) X_{ij} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})^{\mathrm{T}} \widetilde{\boldsymbol{\theta}}_{j^*} \\
=\ & -\sum_i (\tau - I\{e_i \leq 0\}) X_{ij} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) \widetilde{\boldsymbol{\theta}}_{j^*} \\
& -\sum_i (I\{e_i \leq 0\} - I\{e_i \leq \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) - m_i\}) X_{ij} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})^{\mathrm{T}} \widetilde{\boldsymbol{\theta}}_{j^*}.
\end{aligned} \tag{29}$$

The first term above can be bounded as $\sum_i (\tau - I\{e_i \leq 0\}) X_{ij} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) \widetilde{\boldsymbol{\theta}}_{j^*} = O_p(\sqrt{nK}) \|\widetilde{\boldsymbol{\theta}}_{j^*}\|$. For the second term above, let M_n be any positive sequence diverging to infinity, we have, since $|\mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) - m_i| = O_p(\sqrt{K} r_n)$, $P(|\mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) - m_i| > M_n \sqrt{K} r_n) \to 0$, and

$$\begin{aligned}
& E\Big[\Big\| \sum_i (I\{e_i \leq 0\} - I\{e_i \leq \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) - m_i\}) X_{ij^*} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) \Big\|^2 \\
& \qquad I\{|\mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}} \mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) - m_i| \leq M_n \sqrt{K} r_n\} \Big] \\
\leq\ & E\left[\left(\sum_i |I\{e_i \leq M_n \sqrt{K} r_n\} - I\{e_i \leq -M_n \sqrt{K} r_n\}| \cdot \|\mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})\| \right)^2 \right] \\
=\ & E\left[\sum_i I\{-M_n \sqrt{K} r_n \leq e_i \leq M_n \sqrt{K} r_n\} \cdot \|\mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})\|^2 \right] \\
& + \sum_{i \neq i'} E\Big[I\{-M_n \sqrt{K} r_n \leq e_i \leq M_n \sqrt{K} r_n\} I\{-M_n \sqrt{K} r_n \leq e_{i'} \leq M_n \sqrt{K} r_n\} \\
& \qquad \|\mathbf{B}(\mathbf{W}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})\| \|\mathbf{B}(\mathbf{W}_{i'}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}})\| \Big] \\
\leq\ & C(n M_n \sqrt{K} r_n + n^2 M_n^2 K r_n^2) \|\widetilde{\boldsymbol{\theta}}_{j^*}\|^2,
\end{aligned}$$

and thus the second term of (29) is $O_p(n\sqrt{K}r_n)\|\widetilde{\boldsymbol{\theta}}_{j^*}\|$. Besides, considering the difference of the penalty, we have

$$n\sum_j p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|_{\mathbf{A}_j}) - n\sum_j p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|_{\mathbf{A}_j}) = np_\lambda(\|\widetilde{\boldsymbol{\theta}}_{j^*}\|_{\mathbf{A}_j}) = n\lambda\|\widetilde{\boldsymbol{\theta}}_{j^*}\|_{\mathbf{A}_j}, \qquad (30)$$

where the last equality used the fact that $p_\lambda(|x|) = \lambda|x|$ when $|x| < \lambda$. Putting together (29) and (30), we get

$$\sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})) + n\sum_j p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|_{\mathbf{A}_j}) > \sum_i \rho_\tau(Y_i - \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}\mathbf{B}(\mathbf{W}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})) + n\sum_j p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j^*\|_{\mathbf{A}_j})$$

with probability approaching one. This leads to a contradiction.

Finally, to show part (ii) of the theorem, we note that given part (i) holds, restricted to a r_n-neighborhood, the penalty $n\sum_{j=1}^q p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|_{\mathbf{A}_j})$ remains a constant. Thus the local minimizer without a penalty is also a local minimizer of the objective function with a penalty. Then asymptotic properties we want to prove directly follows from Theorem 1. □

Finally, since the proof of Theorem 6 is similar to that of Theorem 4, we choose to omit the details here.