Supplementary Materials for Interactive Q-learning for Quantiles

Kristin A. Linn[1], Eric B. Laber[2], Leonard A. Stefanski[2]

[1]Department of Biostatistics and Epidemiology

University of Pennsylvania, Philadelphia, PA 19104

[2]Department of Statistics

North Carolina State University, Raleigh, NC 27695

email: klinn@upenn.edu

December 22, 2015

1. PROOF OF THEOREM 2.1

The following conditions are restated from the main paper: (C1) consistency, so that $Y = Y^*(A_1, A_2)$; (C2) sequential ignorability (Robins, 2004), i.e., $A_t \perp\!\!\!\perp W \mid \boldsymbol{H}_t$ for $t = 1, 2$; and (C3) positivity, so that there exists $\epsilon > 0$ for which $\epsilon < \text{pr}(A_t = a_t | \boldsymbol{H}_t) < 1 - \epsilon$ with probability one for all a_t, $t = 1, 2$. Lemma 2.1 is useful in the proof of Theorem 2.1 below.

Lemma 2.1. *Assume pr$\{Y^*(\boldsymbol{\pi}) \leq y\}$ is continuous for all fixed $\boldsymbol{\pi}$. Then, pr$\{Y^*(\boldsymbol{\pi}^y) \leq y\}$ is continuous in y in a neighborhood of y^*_τ.*

Proof. Let $\epsilon > 0$ be fixed and arbitrary. Choose $\delta_1 > 0$, $\delta_2 > 0$, and $\delta_3 > 0$ such that

$$\begin{aligned}
\left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta_1}) \le y^*_\tau + \delta_1\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta_1}) \le y^*_\tau\right\}\right| &< \frac{\epsilon}{3} \\
\left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta_2}) \le y^*_\tau\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta_2}) \le y^*_\tau - \delta_2\right\}\right| &< \frac{\epsilon}{3} \\
\left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau + \delta_3\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau\right\}\right| &< \frac{\epsilon}{3},
\end{aligned}$$

and let $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Then,

$$\begin{aligned}
&\left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta}) \le y^*_\tau + \delta\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta}) \le y^*_\tau - \delta\right\}\right| \\
&\qquad \le \left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta}) \le y^*_\tau + \delta\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta}) \le y^*_\tau\right\}\right| \\
&\qquad + \left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta}) \le y^*_\tau\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta}) \le y^*_\tau - \delta\right\}\right| \\
&\qquad\qquad + \left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau + \delta\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau\right\}\right| < \epsilon,
\end{aligned}$$

where we have used the triangle inequality and the fact that

$$\begin{aligned}
&\left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau + \delta}) \le y^*_\tau\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau - \delta}) \le y^*_\tau\right\}\right| \\
&\qquad \le \left|\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau + \delta\right\} - \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau\right\}\right|.
\end{aligned}$$

■

Theorem 2.1. *Let $\epsilon > 0$ and $\tau \in (0,1)$ be arbitrary but fixed. Assume (C1)-(C3) and that the map $y \mapsto R(y; \boldsymbol{x}_1, a_1, \boldsymbol{x}_2, a_2)$ from $\mathbb{R}$ into $(0,1)$ is continuous and strictly increasing in a neighborhood of τ for all $\boldsymbol{x}_1, a_1, \boldsymbol{x}_2$, and a_2. Then, $\inf\{y : pr\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y\right\} \ge \tau\} = y^*_\tau$.*

Proof. Define $\tilde{y} = \inf \mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y\right\} \ge \tau$, and assume $\tilde{y} < y^*_\tau$. The assumption that $R(y; \boldsymbol{x}_1, a_1, \boldsymbol{x}_2, a_2)$ is continuous and strictly increasing for all $\boldsymbol{x}_1, a_1, \boldsymbol{x}_2$, and a_2 implies $\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}) \le y\right\}$ is continuous and strictly increasing for all fixed $\boldsymbol{\pi}$. Thus, $\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le \tilde{y}\right\} = \tau$. By Lemma 2.1, $\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y^*_\tau\right\} = \tau$. However, this implies $\mathrm{pr}\left\{Y^*(\boldsymbol{\pi}^{y^*_\tau}) \le y\right\}$ is not strictly increasing, which is a contradiction. Thus, $\tilde{y} = y^*_\tau$ and $\boldsymbol{\pi}^{y^*_\tau}$ is an optimal regime. ■

2. THRESHOLD INTERACTIVE Q-LEARNING WITH SECOND-STAGE HETEROSKEDASTICITY

Here we assume

$$Y = m(\boldsymbol{H}_2) + A_2 c(\boldsymbol{H}_2) + \eta(\boldsymbol{H}_2, A_2)\epsilon, \tag{1}$$

where we define $\eta(\boldsymbol{H}_2, A_2) = \exp\{r(\boldsymbol{H}_2) + A_2 s(\boldsymbol{H}_2)\}$ for functions r and s. In addition, $E(\epsilon) = 0$, $\mathrm{var}(\epsilon) = 1$, and ϵ is independent of $\boldsymbol{H}_2$ and A_2. Thus, the conditional variance of Y given $\boldsymbol{H}_2$ and A_2 is log-linear. Under model (1), the λ-optimal second-stage decision rule for a patient presenting with $\boldsymbol{h}_2$ is

$$\pi_{2,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_2) = \mathrm{sgn}\left[\frac{\lambda - m(\boldsymbol{h}_2) + c(\boldsymbol{h}_2)}{\exp\{r(\boldsymbol{h}_2) - s(\boldsymbol{h}_2)\}} - \frac{\lambda - m(\boldsymbol{h}_2) - c(\boldsymbol{h}_2)}{\exp\{r(\boldsymbol{h}_2) + s(\boldsymbol{h}_2)\}}\right]. \tag{2}$$

To see this, define

$$\begin{aligned}
\mathrm{pr}^{\pi_1,\pi_2}(Y > \lambda) &= E[E\{\mathrm{pr}^{\pi_1,\pi_2}(Y > \lambda \mid \boldsymbol{H}_2, a_2) \mid_{a_2=\pi_2(\boldsymbol{H}_2)} \mid \boldsymbol{H}_1, a_1\} \mid_{a_1=\pi_1(\boldsymbol{H}_1)}] \\
&= E\left\{E\left(\mathrm{pr}\left[\epsilon > \frac{\lambda - m(\boldsymbol{H}_2) - \pi_2(\boldsymbol{H}_2)c(\boldsymbol{H}_2)}{\exp\{r(\boldsymbol{H}_2) + \pi_2(\boldsymbol{H}_2)s(\boldsymbol{H}_2)\}}\right] \mid \boldsymbol{H}_1, \pi_1(\boldsymbol{H}_1)\right)\right\}.
\end{aligned}$$

To maximize the previous expression, choose $\pi_2(\boldsymbol{h}_2) \in \{-1, 1\}$ to minimize

$$\frac{\lambda - m(\boldsymbol{h}_2) - \pi_2(\boldsymbol{h}_2)c(\boldsymbol{h}_2)}{\exp\{r(\boldsymbol{h}_2) + \pi_2(\boldsymbol{h}_2)s(\boldsymbol{h}_2)\}},$$

leading to $\pi_{2,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_2)$ in (2). Define $G(\cdot,\cdot,\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ to be the joint conditional distribution of $\{m(\boldsymbol{h}_2), c(\boldsymbol{h}_2), r(\boldsymbol{h}_2), s(\boldsymbol{h}_2)\}$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$. Let $F_\epsilon(\cdot)$ denote the cumulative distribution function of ϵ. The first-stage λ-optimal decision rule is

$$\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \arg\min_{a_1} \int F_\epsilon\left(\frac{\lambda - t - \mathrm{sgn}\{K(t,u,v,w)\}u}{\exp[v + \mathrm{sgn}\{K(t,u,v,w)\}w]}\right) G(t,u,v,w \mid \boldsymbol{h}_1, a_1)\,dt\,du\,dv\,dw,$$

where

$$K(t,u,v,w) = \frac{\lambda - t + u}{\exp(v - w)} - \frac{\lambda - t - u}{\exp(v + w)}.$$

Thus, estimation of $\pi_{1,\lambda}^{\mathrm{TIQ}}$ involves specifying estimators for $F_\epsilon(\cdot)$ and the four-dimensional conditional density $G(\cdot,\cdot,\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$. Alternatively, a suitable transformation of the response may be employed to obtain constant variance at the second stage, and then the methods described in Section 2 of the main paper may be applied.

3. THRESHOLD INTERACTIVE Q-LEARNING WITH PATIENT-SPECIFIC THRESHOLDS

Denote the optimal second-stage rule for patient-specific threshold $\lambda(\boldsymbol{h}_t)$ by $\pi_{2,\lambda(\boldsymbol{h}_t)}^{\mathrm{TIQ}}(\boldsymbol{h}_2)$, where $t = 1$ or $t = 2$, depending on the scientific interest and trial design. Then, $\pi_{2,\lambda(\boldsymbol{h}_t)}^{\mathrm{TIQ}}(\boldsymbol{h}_2) = \pi_2^*(\boldsymbol{h}_2) = \mathrm{sgn}\{c(\boldsymbol{h}_2)\}$ whether $t = 1$ or 2. To see this, note for fixed π_1,

$$\mathrm{pr}^{\pi_1,\pi_2}\{Y > \lambda(\boldsymbol{H}_t)\} = E(E[\mathrm{pr}^{\pi_1,\pi_2}\{Y > \lambda(\boldsymbol{H}_t) \mid \boldsymbol{H}_2, a_2\}\,|_{a_2=\pi_2(\boldsymbol{H}_2)} \mid \boldsymbol{H}_1, a_1]\,|_{a_1=\pi_1(\boldsymbol{H}_1)}\,.$$

Because $\boldsymbol{H}_1 \subset \boldsymbol{H}_2$, conditioning on $\boldsymbol{H}_2$ reduces $\lambda(\boldsymbol{H}_t)$ to a constant whether $t = 1$ or 2. Thus, using the set-up in Section 2 of the main paper, the derivation of the optimal second-stage rule in that section applies, giving the result that $\pi_{2,\lambda(\boldsymbol{h}_t)}^{\mathrm{TIQ}}(\boldsymbol{h}_2) = \pi_2^*(\boldsymbol{h}_2) = \mathrm{sgn}\{c(\boldsymbol{h}_2)\}$.

When the threshold depends on the first-stage history, $\lambda(\boldsymbol{h}_1)$ replaces λ in Step TIQ.4 of the TIQ-learning algorithm in Section 2.1 of the main paper, and no additional modeling is needed. When the threshold depends on the second-stage history, the joint conditional distribution of $\{\lambda(\boldsymbol{H}_2),\ m(\boldsymbol{H}_2),\ c(\boldsymbol{H}_2)\}$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$ must be estimated. Let $G(\cdot,\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ denote this trivariate distribution and $\widehat{G}(\cdot,\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ an estimator. In this case, the estimated optimal first-stage decision rule is

$$\widehat{\pi}_{1,\lambda(\boldsymbol{h}_2)}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \arg\min_{a_1} \int \widehat{F}_\epsilon\left(t - u - |v|\right) \widehat{G}(t, u, v \mid \boldsymbol{h}_1, a_1)dtdudv.$$

Thus, the first-stage optimal treatment is based on the average of all possible future patient-specific thresholds, $\lambda(\boldsymbol{H}_2)$, given the observed first-stage history, $\boldsymbol{h}_1$.

4. QUANTILE INTERACTIVE Q-LEARNING OPTIMAL SECOND-STAGE DECISION RULE

We show the τ-optimal QIQ-learning second-stage rule is $\pi_{2,\tau}^{\mathrm{QIQ}}(\boldsymbol{h}_2) = \mathrm{sgn}\{c(\boldsymbol{h}_2)\}$ under the assumption of constant variance at the second-stage. Define the set $S^{\pi_1,\pi_2} \triangleq \{y : \mathrm{pr}^{\pi_1,\pi_2}(Y \leq y) \geq \tau\}$, so that $q^{\pi_1,\pi_2}(\tau) = \inf S^{\pi_1,\pi_2}$. In Section 2.1 of the main paper, we showed $\mathrm{pr}^{\pi_1,\pi_2}(Y \leq y) \geq \mathrm{pr}^{\pi_1,\pi_2^*}(Y \leq y)$ for arbitrary y, and hence for all fixed y, where we define $\pi_2^*(\boldsymbol{h}_2) = \mathrm{sgn}\{c(\boldsymbol{h}_2)\}$. It follows that $S^{\pi_1,\pi_2^*} \subset S^{\pi_1,\pi_2}$. Hence, $\inf S^{\pi_1,\pi_2^*} \geq \inf S^{\pi_1,\pi_2}$; equivalently, $q^{\pi_1,\pi_2^*}(\tau) \geq q^{\pi_1,\pi_2}(\tau)$. Thus, $\pi_{2,\tau}^{\mathrm{QIQ}}(\boldsymbol{h}_2) = \pi_2^*(\boldsymbol{h}_2) = \mathrm{sgn}\{c(\boldsymbol{h}_2)\}$ is optimal because this inequality holds for arbitrary π_1 and π_2.

5. PROOF OF LEMMA 3.1 IN SECTION 3

Lemma 3.1 from Section 3 of the main paper is restated below.

(A) $y < y_\tau^*$ implies $y < f(y) \leq y_\tau^*$;

(B) $f(y_\tau^{*-}) \triangleq \lim_{\delta \downarrow 0} f(y_\tau^* - \delta) = y_\tau^*$;

(C) $f(y_\tau^*) \leq y_\tau^*$ with strict inequality if there exists $\delta > 0$ such that $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}(Y \leq y_\tau^* - \delta) \geq \tau$;

(D) If $F_\epsilon(\cdot)$ is continuous and strictly increasing, then $f(y_\tau^*) = y_\tau^*$.

Proof. We showed below expression (8) of Section 3 in the main paper that $f(y) \leq y_\tau^*$ for all y. We prove the remainder of *(A)* by contradiction. Assume there exists a $y_0 < y_\tau^*$ such that $y_0 \geq f(y_0)$. It follows that

$$\tau \leq \mathrm{pr}^{\Gamma(\cdot,\, y_0),\, \pi_2^*}\{Y \leq f(y_0)\} \leq \mathrm{pr}^{\Gamma(\cdot,\, y_0),\, \pi_2^*}(Y \leq y_0)$$

because for the fixed regime $\pi = \{\Gamma(\cdot, y_0), \pi_2^*\}$, $\mathrm{pr}^{\Gamma(\cdot,\, y_0),\, \pi_2^*}\{Y \leq y\}$ is a distribution function and nondecreasing in y. However, we have a contradiction because by definition, y_τ^* is the smallest y satisfying $\mathrm{pr}^{\Gamma(\cdot,\, y),\, \pi_2^*}(Y \leq y) \geq \tau$.

Using (A) and the fact that for $\delta > 0$, $y_\tau^* - \delta < y_\tau^*$ implies $y_\tau^* - \delta < f(y_\tau^* - \delta)$, we see that $y_\tau^* - \delta < f(y_\tau^* - \delta) \leq y_\tau^*$. Letting $\delta \to 0$ proves *(B)*.

Given that $f(y) \leq y_\tau^*$ for all y, $f(y_\tau^*) \leq y_\tau^*$ and thus in light of *(B)* the inequality is strict when $f(y)$ is not left continuous at y_τ^*. If there exists $\delta > 0$ such that $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}(Y \leq y_\tau^* - \delta) \geq \tau$, then because $f(y_\tau^*)$ is the smallest $\tilde{y}$ for which $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}(Y \leq \tilde{y}) \geq \tau$ it must be that $f(y_\tau^*) \leq y_\tau^* - \delta < y_\tau^*$, proving *(C)*.

When $F_\epsilon(\cdot)$ is continuous and strictly increasing, $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}(Y \leq y)$ is also continuous and strictly increasing because it is an expectation of a continuous, strictly increasing function of y. It can be shown that for any fixed regime $\boldsymbol{\pi} = (\pi_1, \pi_2)$, $\mathrm{pr}^{\pi_1,\, \pi_2}(Y \leq y)$ continuous in y implies $\mathrm{pr}^{\Gamma(\cdot,\, y),\, \pi_2^*}(Y \leq y)$ is also continuous. Suppose toward a contradiction that $f(y_\tau^*) < y_\tau^*$. When $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}(Y \leq y)$ is continuous and strictly increasing, the Mean Value Theorem guarantees existence of exactly one point $\tilde{y} \in \mathbb{R}$ such that $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}\{Y \leq \tilde{y}\} = \tau$. By definition, $f(y_\tau^*)$ must be this point, and thus $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}\{Y \leq f(y_\tau^*)\} = \tau$. The assumption $f(y_\tau^*) < y_\tau^*$ implies $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}\{Y \leq y_\tau^*\} > \tau$. However, when $\mathrm{pr}^{\Gamma(\cdot,\, y),\, \pi_2^*}(Y \leq y)$ is continuous, $\mathrm{pr}^{\Gamma(\cdot,\, y_\tau^*),\, \pi_2^*}\{Y \leq y_\tau^*\} = \tau$ by the Mean Value Theorem and by the definition of y_τ^*. Thus, we have a contradiction and conclude that *(D)* holds. ■

6. QUANTILE INTERACTIVE Q-LEARNING TOY EXAMPLE: $f(y_\tau^*) \neq y_\tau^*$

Suppose all subjects have the same first-stage covariates, i.e., $\boldsymbol{H}_1 = \boldsymbol{h}_1$ with probability one. Fix $\tau = 0.5$ and let $p(y \mid \boldsymbol{h}_1, a_1)$ denote the conditional density of Y given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and

$A_1 = a_1$. Suppose

$$p(y \mid \boldsymbol{h}_1, 1) = \begin{cases} -2.5 \text{ with probability } 0.1 \\ -1.5 \text{ with probability } 0.2 \\ -0.5 \text{ with probability } 0.2 \\ 0.5 \text{ with probability } 0.2 \\ 1.5 \text{ with probability } 0.2 \\ 2.5 \text{ with probability } 0.1 \end{cases}$$

and

$$p(y \mid \boldsymbol{h}_1, -1) = \begin{cases} \text{Uniform}(-2, 0) \text{ with probability } 0.5 \\ 0 \text{ with probability } 0.5. \end{cases}$$

Then, $f(y_\tau^*) < y_\tau^*$ because $y_\tau^* = 0$ and $f(y_\tau^*) = -1$. Recall $y_\tau^* = \inf\{y : \text{pr}^{\Gamma(\cdot, y), \pi_2^*}(Y \leq y) \geq \tau\}$ by definition. Figure 1 provides plots of the cumulative distribution functions of Y when $A_1 = -1, 1$. In this example, $f(y_\tau^{*-}) = y_\tau^*$, where y_τ^{*-} denotes the left limit of y_τ^*.

7. PROOFS OF THEOREMS 3.2 AND 3.3

The following assumptions are used to establish consistency of the threshold exceedance probability and quantile that result from applying the estimated TIQ- and QIQ-learning optimal regimes, respectively.

A1. The method used to estimate $m(\cdot)$ and $c(\cdot)$ results in estimators $\widehat{m}(\boldsymbol{h}_2)$ and $\widehat{c}(\boldsymbol{h}_2)$ that converge in probability to $m(\boldsymbol{h}_2)$ and $c(\boldsymbol{h}_2)$, respectively, for each $\boldsymbol{h}_2$.

A2. $F_\epsilon(\cdot)$ is continuous, $\widehat{F}_\epsilon(\cdot)$ is a cumulative distribution function, and $\widehat{F}_\epsilon(y)$ converges in probability to $F_\epsilon(y)$ uniformly in y.

A3. For each fixed $\boldsymbol{h}_1$ and a_1, $\int |d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - dG(u, v \mid \boldsymbol{h}_1, a_1)|$ converges to zero in

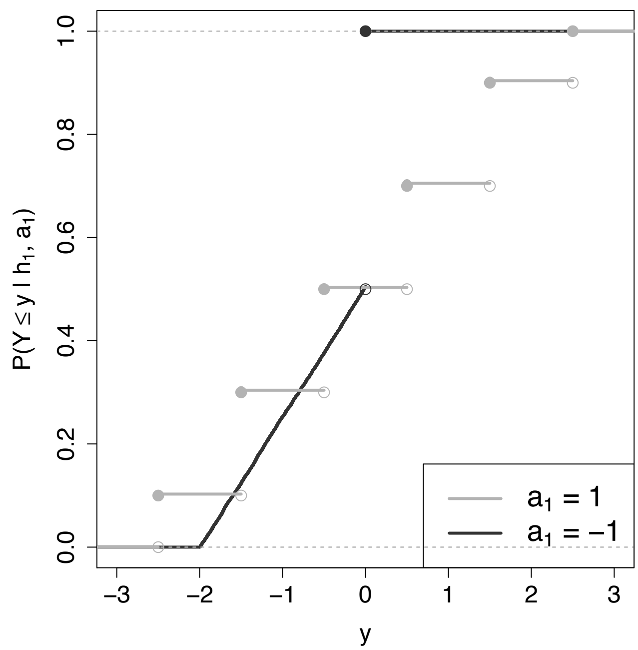

Figure 1: Cumulative distribution functions of Y given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = -1, 1$. The optimal $\tau = 0.5$ quantile is $y^*_\tau = 0$. However, if patients are treated with the treatment that minimizes $\mathrm{pr}(Y \le y^*_\tau \mid \boldsymbol{h}_1, a_1)$, namely $a_1 = 1$, the resulting quantile, $f(y^*_\tau) = -0.5$, is suboptimal.

probability.

A4. For each fixed a_1, $n^{-1}\sum_{i=1}^{n}\int |d\widehat{G}(u,v \mid \boldsymbol{H}_{1i}, a_1) - dG(u,v \mid \boldsymbol{H}_{1i}, a_1)|$ converges to zero in probability.

Theorem 3.2. *(Consistency of TIQ-learning) Assume A1–A3 and fix $\lambda \in \mathbb{R}$. Then, $pr^{\widehat{\boldsymbol{\pi}}_{\lambda}^{TIQ}}(Y > \lambda)$ converges in probability to $pr^{\boldsymbol{\pi}_{\lambda}^{TIQ}}(Y > \lambda)$, where $\widehat{\boldsymbol{\pi}}_{\lambda}^{TIQ} = (\widehat{\pi}_{1,\lambda}^{TIQ}, \widehat{\pi}_2^*)$.*

Theorem 3.3. *(Consistency of QIQ-learning) Assume A1–A4. Then, $q^{\widehat{\boldsymbol{\pi}}_{\tau}^{QIQ}}(\tau)$ converges in proability to y_τ^* for any fixed τ, where $\widehat{\boldsymbol{\pi}}_{\tau}^{QIQ} = (\widehat{\Gamma}(\cdot, \widehat{y}_\tau^*), \widehat{\pi}_2^*)$.*

Capital letters denote random variables and lower case letters denote observed realizations. Let $\mathcal{D} = \{\boldsymbol{X}_{1i}^{\intercal}, A_{1i}, \boldsymbol{X}_{2i}^{\intercal}, A_{2i}, Y_i\}_{i=1}^{n}$ denote the observed data, which are n independent and identically distributed realizations of the trajectory $(\boldsymbol{X}_1^{\intercal}, A_1, \boldsymbol{X}_2^{\intercal}, A_2, Y)^{\intercal}$. Let $(\boldsymbol{X}_1^{\intercal}, A_1, \boldsymbol{X}_2^{\intercal}, A_2, Y)^{\intercal}$ be a trajectory that is independent of $\mathcal{D}$ but identically distributed. Let $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $\boldsymbol{H}_2 = (\boldsymbol{X}_1^{\intercal}, A_1, \boldsymbol{X}_2^{\intercal})^{\intercal}$ denote the full patient histories available prior to treatment at stages one and two. When necessary, we use $\boldsymbol{H}_2^{A_1}$ and $\boldsymbol{H}_2^{\pi_1(\boldsymbol{H}_1)}$ to emphasize dependence of $\boldsymbol{H}_2$ on the first-stage treatment.

Using the set-up and assumptions described in Section 2, the optimal and estimated optimal second-stage rules for a patient presenting with $\boldsymbol{h}_2$ are $\pi_2^*(\boldsymbol{h}_2) = \operatorname{sgn}\{c(\boldsymbol{h}_2)\}$ and $\widehat{\pi}_2^*(\boldsymbol{h}_2) = \operatorname{sgn}\{\widehat{c}(\boldsymbol{h}_2)\}$. In addition, we use the following notation first introduced in Section 2.1:

$$\begin{aligned}
d(\boldsymbol{h}_1, y) &= \int F_\epsilon(y - u - |v|) dG(u, v \mid \boldsymbol{h}_1, -1) - \int F_\epsilon(y - u - |v|) dG(u, v \mid \boldsymbol{h}_1, 1), \\
\widehat{d}(\boldsymbol{h}_1, y) &= \int \widehat{F}_\epsilon(y - u - |v|) d\widehat{G}(u, v \mid \boldsymbol{h}_1, -1) - \int \widehat{F}_\epsilon(y - u - |v|) d\widehat{G}(u, v \mid \boldsymbol{h}_1, 1).
\end{aligned}$$

With this notation, the optimal and estimated optimal first-stage rules for TIQ-learning are $\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \operatorname{sgn}\{d(\boldsymbol{h}_1, \lambda)\}$ and $\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \operatorname{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\}$. We define $\operatorname{sgn}(0) = 1$. The following Lemmas are useful for the proofs of Theorems 3.2 and 3.3. In some of the Lemmas, we use Δ with or without a subscript to denote a difference of two quantities; this notation is used locally, and thus, Δ appears in multiple Lemmas representing different expressions.

Lemma 7.1. *If X_n converges to μ in probability, then $T_n = |sgn(X_n) - sgn(\mu)|\mathbb{1}_{|\mu|>0}$ converges to zero in probability, and $E(T_n)$ converges to zero as n converges to ∞.*

Proof. If $\mu = 0$, then $\mathrm{pr}(T_n = 0) = 1$ for all n. If $\mu > 0$, then $T_n = |\mathrm{sgn}(X_n) - 1|$ and $\mathrm{pr}(T_n > 0) = \mathrm{pr}(X_n < 0)$, which converges to zero. If $\mu < 0$, then $T_n = |\mathrm{sgn}(X_n) + 1|$ and $\mathrm{pr}(T_n > 0) = \mathrm{pr}(X_n > 0)$, which converges to zero. Because $0 \le T_n \le 2$ for all n, it follows that $E(T_n)$ converges to zero as n converges to ∞. ■

Lemma 7.2. *Assume A2 and A3. Then, for fixed $\boldsymbol{h}_1$, $\sup_y |\widehat{d}(\boldsymbol{h}_1, y) - d(\boldsymbol{h}_1, y)|$ converges to zero.*

Proof. By the triangle inequality,

$$\sup_y |\widehat{d}(\boldsymbol{h}_1, y) - d(\boldsymbol{h}_1, y)| \le \sup_y |\Delta(y; \boldsymbol{h}_1, -1)| + \sup_y |\Delta(y; \boldsymbol{h}_1, 1)|,$$

where $\Delta(y; \boldsymbol{h}_1, a_1) = \int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - \int F_\epsilon(y-u-|v|)dG(u, v \mid \boldsymbol{h}_1, a_1)$. Thus, we show $\sup_y |\Delta(y; \boldsymbol{h}_1, a_1)|$ converges in probability to zero for an arbitrary a_1. Applying the triangle inequality leads to the upper bound

$$\begin{aligned} \sup_y |\Delta(y; \boldsymbol{h}_1, a_1)| \le \sup_y \int \widehat{F}_\epsilon(y - u - |v|) \left| d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - dG(u, v \mid \boldsymbol{h}_1, a_1) \right| \\ + \sup_y \int \left| \widehat{F}_\epsilon(y - u - |v|) - F_\epsilon(y - u - |v|) \right| dG(u, v \mid \boldsymbol{h}_1, a_1). \quad (3) \end{aligned}$$

Add a term and subtract a term

An upper bound on the right-hand side of (3) is

$$\begin{aligned} \int \left| d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - dG(u, v \mid \boldsymbol{h}_1, a_1) \right| + \sup_w \left| \widehat{F}_\epsilon(w) - F_\epsilon(w) \right| \int dG(u, v \mid \boldsymbol{h}_1, a_1) \\ = \int \left| d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - dG(u, v \mid \boldsymbol{h}_1, a_1) \right| + \sup_w \left| \widehat{F}_\epsilon(w) - F_\epsilon(w) \right|, \quad (4) \end{aligned}$$

where we have used the fact that $\sup_w \widehat{F}_\epsilon(w) = 1$ and $\int dG(u, v \mid \boldsymbol{h}_1, a_1) = 1$. The first and second terms in (4) are $o_p(1)$ by assumptions A3 and A2. ■

Lemma 7.3. *Assume A1. Then,* $\sup_{\pi_1, y} \left| pr^{\pi_1, \widehat{\pi}_2^*}(Y \le y) - pr^{\pi_1, \pi_2^*}(Y \le y) \right|$ *converges to zero in probability.*

Proof. Define $\widehat{\Delta}_\epsilon(y; \boldsymbol{h}_2^{a_1}) = F_\epsilon\left[y - m(\boldsymbol{h}_2^{a_1}) - \text{sgn}\left\{\widehat{c}(\boldsymbol{h}_2^{a_1})\right\} c(\boldsymbol{h}_2^{a_1})\right] - F_\epsilon\left\{y - m(\boldsymbol{h}_2^{a_1}) - |c(\boldsymbol{h}_2^{a_1})|\right\}$ and $\widehat{\Delta}_c(\boldsymbol{h}_2^{a_1}) = |\text{sgn}\{\widehat{c}(\boldsymbol{h}_2^{a_1})\} - \text{sgn}\{c(\boldsymbol{h}_2^{a_1})\}| \, \mathbb{1}_{|c(\boldsymbol{h}_2^{a_1})|>0}$. Note that for each $\boldsymbol{h}_2^{a_1}$, $\left|\widehat{\Delta}_\epsilon(y; \boldsymbol{h}_2^{a_1})\right| \le \widehat{\Delta}_c(\boldsymbol{h}_2^{a_1})$; thus, using definitions given in Section 3,

$$
\begin{aligned}
&\sup_{\pi_1, y} \left| \text{pr}^{\pi_1, \widehat{\pi}_2^*}(Y \le y) - \text{pr}^{\pi_1, \pi_2^*}(Y \le y) \right| \\
&\qquad = \sup_{\pi_1, y} \left| \int \int \widehat{\Delta}_\epsilon\{y; \boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}\} dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\} dF_{\boldsymbol{h}_1}(\boldsymbol{h}_1) \right| \\
&\qquad \le \sup_{\pi_1, y} \int \int \left| \widehat{\Delta}_\epsilon\{y; \boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}\} \right| dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1) \\
&\qquad\qquad \le \sup_{\pi_1} \int \int \widehat{\Delta}_c\{\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}\} dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1),
\end{aligned}
$$

where we have used the fact that $\widehat{\Delta}_c\{\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}\}$ does not depend on y. Because $\pi_1(\cdot)$ has range $\{-1, 1\}$, an upper bound on the right-hand side above is

$$
\begin{aligned}
&\int \int \sum_{a_1 \in \{-1,1\}} \widehat{\Delta}_c(\boldsymbol{h}_2^{a_1}) dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, a_1\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1) \\
&\qquad = \sum_{a_1 \in \{-1,1\}} \int \int \widehat{\Delta}_c(\boldsymbol{h}_2^{a_1}) dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, a_1\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1) \\
&\qquad\qquad = \sum_{a_1 \in \{-1,1\}} E\left\{ \widehat{\Delta}_c(\boldsymbol{H}_2^{A_1}) \mid A_1 = a_1, \mathcal{D} \right\}, \quad (5)
\end{aligned}
$$

which does not depend on π_1. We claim the right-hand side of (5) is $o_p(1)$. To show this, note for each fixed a_1,

$$
E\left\{ \widehat{\Delta}_c(\boldsymbol{H}_2^{A_1}) \mid A_1 = a_1 \right\} = \int \int E\left\{ \widehat{\Delta}_c(\boldsymbol{h}_2^{a_1}) \right\} dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, a_1\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1),
$$

where $E\{\widehat{\Delta}_c(\boldsymbol{h}_2^{a_1})\}$ converges to zero by Lemma 7.1 for each $\boldsymbol{h}_2^{a_1}$. Because $0 \le E\{\widehat{\Delta}_c(\boldsymbol{h}_2^{a_1})\} \le 2$, applying the Dominated Convergence Theorem gives the result that $E\{\widehat{\Delta}_c(\boldsymbol{H}_2^{A_1}) \mid A_1 =$

$a_1\}$ converges to zero, which implies $E\{\widehat{\Delta}_c(\boldsymbol{H}_2^{A_1}) \mid A_1 = a_1, \mathcal{D}\}$ is $o_p(1)$ for each fixed a_1 by Lemma 7.1. Thus, the right hand side of (5) is $o_p(1)$. ■

Lemma 7.4. *Assume A2 and A3, and fix* $\lambda \in \mathbb{R}$. *Then,* $\left|pr^{\widehat{\pi}_{1,\lambda}^{TIQ},\pi_2^*}(Y \leq \lambda) - pr^{\pi_{1,\lambda}^{TIQ},\pi_2^*}(Y \leq \lambda)\right|$ *converges to zero in probability.*

Proof. Define $\widehat{\Delta}_G(\boldsymbol{h}_1; u, v) = dG\{u, v \mid \boldsymbol{h}_1, \widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1)\} - dG\{u, v \mid \boldsymbol{h}_1, \pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1)\}$, and note that $\widehat{\Delta}_G(\boldsymbol{h}_1; u, v) = \{\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) - \widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1)\}\{dG(u, v \mid \boldsymbol{h}_1, -1) - dG(u, v \mid \boldsymbol{h}_1, 1)\}/2$. Using the definitions given in Section 2.1,

$$\begin{aligned}\left|\mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \leq \lambda) - \mathrm{pr}^{\pi_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \leq \lambda)\right| &= \left|\int\int F_\epsilon(\lambda - u - |v|)\widehat{\Delta}_G(\boldsymbol{h}_1; u, v)dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1)\right| \\ &\leq \int |d(\boldsymbol{h}_1, \lambda)| \left|\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) - \widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1)\right| dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1). \quad (6)\end{aligned}$$

Substituting $\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \mathrm{sgn}\{d(\boldsymbol{h}_1, \lambda)\}$, $\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \mathrm{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\}$, and noting

$$|d(\boldsymbol{h}_1, \lambda)| \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{h}_1, \lambda)\}\right| \leq \mathbb{1}_{|d(\boldsymbol{h}_1,\lambda)|>0} \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{h}_1, \lambda)\}\right|,$$

an upper bound on the right-hand side of (6) is

$$\begin{aligned}&\int \mathbb{1}_{|d(\boldsymbol{h}_1,\lambda)|>0} \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{h}_1, \lambda)\}\right| dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1) \\ &\qquad = E\left[\mathbb{1}_{|d(\boldsymbol{H}_1,\lambda)|>0} \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{H}_1, \lambda)\}\right| \mid \mathcal{D}\right].\end{aligned}$$

We show the right-hand side is $o_p(1)$ by showing its expectation with respect to $\mathcal{D}$ converges to zero. Thus,

$$\begin{aligned}&E\left[\mathbb{1}_{|d(\boldsymbol{h}_1,\lambda)|>0} \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{H}_1, \lambda)\}\right|\right] \\ &\qquad = \int E\left[\mathbb{1}_{|d(\boldsymbol{h}_1,\lambda)|>0} \left|\mathrm{sgn}\{\widehat{d}(\boldsymbol{h}_1, \lambda)\} - \mathrm{sgn}\{d(\boldsymbol{h}_1, \lambda)\}\right|\right] dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1).\end{aligned}$$

The inside expectation converges to zero by Lemmas 7.1 and 7.2, and applying the Dominated

Convergence Theorem gives the result that the right-hand side above converges to zero. Thus, appealing to Lemma 7.1, we have shown $\left|\mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda) - \mathrm{pr}^{\pi_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda)\right|$ is bounded above by $E[\mathbb{1}_{|d(\boldsymbol{H}_1,\lambda)|>0}|\mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_1,\lambda)\} - \mathrm{sgn}\{d(\boldsymbol{H}_1,\lambda)\}| \mid \mathcal{D}]$ which is $o_p(1)$. ∎

Proof of Theorem 2.3. Fix $\lambda \in \mathbb{R}$. Define $\Delta(\lambda) = \mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\widehat{\pi}_2^*}(Y \le \lambda) - \mathrm{pr}^{\pi_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda)$. Then, by the triangle inequality,

$$
\begin{aligned}
|\Delta(\lambda)| \le &\left|\mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\widehat{\pi}_2^*}(Y \le \lambda) - \mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda)\right| \\
&+ \left|\mathrm{pr}^{\widehat{\pi}_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda) - \mathrm{pr}^{\pi_{1,\lambda}^{\mathrm{TIQ}},\pi_2^*}(Y \le \lambda)\right|. \qquad (7)
\end{aligned}
$$

The first term on the right-hand side of (7) is $o_p(1)$ by Lemma 7.3, and the second term on the right-hand side of (7) is $o_p(1)$ by Lemma 7.4. ∎

Lemma 7.5. *Assume A2 and A4. Then,* $\sup_y n^{-1}\sum_{i=1}^n |\widehat{d}(\boldsymbol{H}_{1i},y) - d(\boldsymbol{H}_{1i},y)|$ *converges to zero in probability.*

Proof. An upper bound on $\sup_y \frac{1}{n}\sum_{i=1}^n |\widehat{d}(\boldsymbol{H}_{1i},y) - d(\boldsymbol{H}_{1i},y)|$ is

$$
\sum_{a_1=1,-1} \sup_y \frac{1}{n}\sum_{i=1}^n \left| \int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},a_1) - \int F_\epsilon(y-u-|v|)dG(u,v \mid \boldsymbol{H}_{1i},a_1)\right|.
$$

By the triangle inequality, the previous expression is bounded above by

$$
\begin{aligned}
&\sum_{a_1=1,-1} \sup_y \frac{1}{n}\sum_{i=1}^n \int \left|\widehat{F}_\epsilon(y-u-|v|) - F_\epsilon(y-u-|v|)\right| d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},a_1) \\
&\quad + \sum_{a_1=1,-1} \sup_y \frac{1}{n}\sum_{i=1}^n \int F_\epsilon(y-u-|v|)\left|d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},a_1) - dG(u,v \mid \boldsymbol{H}_{1i},a_1)\right| \\
&\quad \le 2\sup_w \left|\widehat{F}_\epsilon(w) - F_\epsilon(w)\right| + \sum_{a_1=1,-1} \frac{1}{n}\sum_{i=1}^n \int \left|d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},a_1) - dG(u,v \mid \boldsymbol{H}_{1i},a_1)\right|.
\end{aligned}
$$

The term $\sup_w |\widehat{F}_\epsilon(w) - F_\epsilon(w)|$ is $o_p(1)$ by assumption A2, and for each a_1, $n^{-1}\sum_{i=1}^n \int |d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},a_1) - dG(u,v \mid \boldsymbol{H}_{1i},a_1)|$ is $o_p(1)$ by assumption A4. ∎

Lemma 7.6. *Assume A2 and A4. Then,* $\sup_y |\Delta(y)|$ *converges in probability to zero, where*

$$\Delta(y) = \frac{1}{n}\sum_{i=1}^{n}\int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, sgn\{\widehat{d}(\boldsymbol{H}_{1i},y)\}] \\ - \frac{1}{n}\sum_{i=1}^{n}\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1i}, sgn\{d(\boldsymbol{H}_{1i},y)\}]. \quad (8)$$

Proof. Writing $dG[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{d(\boldsymbol{H}_{1i},t)\}]$ as

$$\frac{1}{2}\{dG(u,v \mid \boldsymbol{H}_{1i},1) + dG(u,v \mid \boldsymbol{H}_{1i},-1)\} \\ - \frac{\mathrm{sgn}\{d(\boldsymbol{H}_{1i},y)\}}{2}\{dG(u,v \mid \boldsymbol{H}_{1i},-1) - dG(u,v \mid \boldsymbol{H}_{1i},1)\}$$

and $d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}]$ as

$$\frac{1}{2}\left\{d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},1) + d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},-1)\right\} \\ - \frac{\mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}}{2}\left\{d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},-1) - d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},1)\right\},$$

$|\Delta(y)|$ is bounded above by

$$\sup_y \frac{1}{n}\sum_{i=1}^{n}|\Delta_i(y)| + \sup_y \frac{1}{n}\sum_{i=1}^{n}\left||\widehat{d}(\boldsymbol{H}_{1i},y)| - |d(\boldsymbol{H}_{1i},y)|\right|, \quad (9)$$

where

$$\Delta_i(y) = \int \widehat{F}_\epsilon(y-u-|v|)\left\{d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},1) + d\widehat{G}(u,v \mid \boldsymbol{H}_{1i},-1)\right\} \\ - \int F_\epsilon(y-u-|v|)\{dG(u,v \mid \boldsymbol{H}_{1i},1) + dG(u,v \mid \boldsymbol{H}_{1i},-1)\}.$$

The term $\sup_y n^{-1}\sum_{i=1}^{n}\left||\widehat{d}(\boldsymbol{H}_{1i},y)| - |d(\boldsymbol{H}_{1i},y)|\right|$ in (9) is bounded above by $\sup_y n^{-1}\sum_{i=1}^{n}\left|\widehat{d}(\boldsymbol{H}_{1i},y) - d(\boldsymbol{H}_{1i},y)\right|$, which is $o_p(1)$ by Lemma 7.5. It can be shown the first term in (9) is bounded above by

$$2\sup_{w}\left|\widehat{F}_{\epsilon}(w)-F_{\epsilon}(w)\right|+\frac{1}{n}\sum_{i=1}^{n}\int\left|d\widehat{G}(u,v\mid \boldsymbol{H}_{1i},1)-dG(u,v\mid \boldsymbol{H}_{1i},1)\right|$$
$$+\frac{1}{n}\sum_{i=1}^{n}\int\left|d\widehat{G}(u,v\mid \boldsymbol{H}_{1i},-1)-dG(u,v\mid \boldsymbol{H}_{1i},-1)\right|,$$

which is $o_p(1)$ by assumptions A2 and A4. ■

Lemma 7.7. *For every fixed $\boldsymbol{h}_1$,*

$$\begin{aligned}\lim_{y\to\infty}\int F_{\epsilon}(y-u-|v|)dG[u,v\mid \boldsymbol{h}_1,a_1=sgn\{d(\boldsymbol{h}_1,y)\}] &= 1,\\ \lim_{y\to-\infty}\int F_{\epsilon}(y-u-|v|)dG[u,v\mid \boldsymbol{h}_1,a_1=sgn\{d(\boldsymbol{h}_1,y)\}] &= 0.\end{aligned}$$

Proof. For each fixed $\boldsymbol{h}_1$ and a_1,

$$\lim_{y\to\infty}\int F_{\epsilon}(y-u-|v|)dG(u,v\mid \boldsymbol{h}_1,a_1)=1,\qquad \lim_{y\to-\infty}\int F_{\epsilon}(y-u-|v|)dG(u,v\mid \boldsymbol{h}_1,a_1)=0,$$

because $\int F_{\epsilon}(y-u-|v|)dG(u,v\mid \boldsymbol{h}_1,a_1)$ is the conditional expectation of a distribution function in y, therefore permitting an exchange of the limit and integration by the dominated convergence theorem. Thus, even if the policy $\mathrm{sgn}\{d(\boldsymbol{h}_1,y)\}$ does not converge as $y\to\infty$ $(-\infty)$, $\lim_{y\to\infty(-\infty)}\int F_{\epsilon}(y-u-|v|)dG[u,v\mid \boldsymbol{h}_1,a_1=\mathrm{sgn}\{d(h_1,y)\}]$ must converge to 1 (0).
■

Lemma 7.8. *For every $\boldsymbol{h}_1$ in the domain of $\boldsymbol{H}_1$, $\int F_{\epsilon}(y-u-|v|)dG[u,v\mid \boldsymbol{h}_1,sgn\{d(\boldsymbol{h}_1,y)\}]$ is non-decreasing in y.*

Proof. We show for arbitrary $s,t\in\mathbb{R}$ such that $s>t$,

$$\int F_{\epsilon}(s-u-|v|)dG[u,v\mid \boldsymbol{h}_1,\mathrm{sgn}\{d(\boldsymbol{h}_1,s)\}]$$
$$-\int F_{\epsilon}(t-u-|v|)dG[u,v\mid \boldsymbol{h}_1,\mathrm{sgn}\{d(\boldsymbol{h}_1,t)\}] \quad (10)$$

is non-negative. Because $\int F_\epsilon(s-u-|v|)dG[u,v \mid \boldsymbol{h}_1, \mathrm{sgn}\{d(\boldsymbol{h}_1,s)\}]$ can be written as

$$\frac{1}{2}\left\{\int F_\epsilon(s-u-|v|)dG(u,v \mid \boldsymbol{h}_1,-1) + \int F_\epsilon(s-u-|v|)dG(u,v \mid \boldsymbol{h}_1,1) - |d(\boldsymbol{h}_1,s)|\right\},$$

(10) simplifies to

$$\begin{aligned}\frac{1}{2}&\left[\int \{F_\epsilon(s-u-|v|) - F_\epsilon(t-u-|v|)\}\, dG(u,v \mid \boldsymbol{h}_1,-1)\right]\\ &\qquad + \frac{1}{2}\left[\int \{F_\epsilon(s-u-|v|) - F_\epsilon(t-u-|v|)\}\, dG(u,v \mid \boldsymbol{h}_1,1)\right]\\ &\qquad\qquad - \frac{1}{2}\left\{|d(\boldsymbol{h}_1,s)| - |d(\boldsymbol{h}_1,t)|\right\}.\end{aligned}$$

The expression above is greater than or equal to zero. To see this, note that

$$\begin{aligned}|d(\boldsymbol{h}_1,s)| - |d(\boldsymbol{h}_1,t)| &\le ||d(\boldsymbol{h}_1,s)| - |d(\boldsymbol{h}_1,t)|| \le |d(\boldsymbol{h}_1,s) - d(\boldsymbol{h}_1,t)|\\ &\le \int \{F_\epsilon(s-u-|v|) - F_\epsilon(t-u-|v|)\}\, dG(u,v \mid \boldsymbol{h}_1,-1)\\ &\qquad + \int \{F_\epsilon(s-u-|v|) - F_\epsilon(t-u-|v|)\}\, dG(u,v \mid \boldsymbol{h}_1,1).\end{aligned}$$

■

Lemma 7.9. *Assume $F_\epsilon(\cdot)$ is continuous. For any fixed $\boldsymbol{h}_1$ in the domain of $\boldsymbol{H}_1$, $\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{h}_1, sgn\{d(\boldsymbol{h}_1,y)\}]$ is continuous in y.*

Proof. This follows immediately by writing $\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{h}_1, \mathrm{sgn}\{d(\boldsymbol{h}_1,y)\}]$ as

$$\frac{1}{2}\left\{\int F_\epsilon(y-u-|v|)dG(u,v \mid \boldsymbol{h}_1,-1) + \int F_\epsilon(y-u-|v|)dG(u,v \mid \boldsymbol{h}_1,1) - |d(\boldsymbol{h}_1,y)|\right\},$$

a linear combination of continuous functions. ■

Lemma 7.10. *Assume A2 and A4. Then,* $\sup_y |L_n(y) - L(y)|$ *converges in probability to* *zero, where*

$$\begin{aligned}
L_n(y) &= \frac{1}{n}\sum_{i=1}^{n}\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1i}, sgn\{d(\boldsymbol{H}_{1i},y)\}],\\
L(y) &= E\left(\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1}, sgn\{d(\boldsymbol{H}_{1},y)\}]\right).
\end{aligned}$$

Proof. The proof is similar to the proof of the Glivenko-Cantelli Theorem given in van der Vaart (2000). Let $\delta > 0$ be arbitrary. By the law of large numbers, $|L_n(y) - L(y)|$ converges to zero in probability for each fixed $y \in \mathbb{R}$. Using Lemmas 7.7, 7.8, and 7.9, it can be shown that $L_n(y)$ and $L(y)$ are both continuous distribution functions in y. Thus, there exists a partition, $-\infty = y_0 < y_1 < \cdots < y_k = \infty$ such that $L(y_i) - L(y_{i-1}) \leq \delta$. For $y_{i-1} \leq y < y_i$,

$$L_n(y_{i-1}) - L(y_{i-1}) - \delta \leq L_n(y) - L(y) \leq L_n(y_i) - L(y_i) + \delta.$$

Convergence of $L_n(y)$ to $L(y)$ is uniform on the finite set $y \in \{y_1, \ldots, y_{k-1}\}$, and thus, $\limsup_y |L_n(y) - L(y)| < \delta$ almost surely. Because δ is arbitrary, the result holds for each δ, which implies the limit superior is zero. ■

Lemma 7.11. *Assume A2 and A4. Then,* $\widehat{y}^*_\tau$ *converges in probability to* y^*_τ.

Proof. Define

$$\Delta(y) = \frac{1}{n}\sum_{i=1}^{n}\int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}] - E\left(\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1}, \mathrm{sgn}\{d(\boldsymbol{H}_{1},y)\}]\right).$$

By the triangle inequality, $\sup_y|\Delta(y)| \leq \sup_y|\Delta_1(y)| + \sup_y|\Delta_2(y)|$, where

$$\Delta_1(y) = \frac{1}{n}\sum_{i=1}^{n}\int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}] - \frac{1}{n}\sum_{i=1}^{n}\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{d(\boldsymbol{H}_{1i},y)\}],$$

$$\Delta_2(y) = \frac{1}{n}\sum_{i=1}^{n}\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_{1i}, A_1 = \mathrm{sgn}\{d(\boldsymbol{H}_{1i},y)\}]$$
$$- E\left(\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_1, A_1 = \mathrm{sgn}\{d(\boldsymbol{H}_1,y)\}]\right).$$

The terms $\sup_y |\Delta_1(y)|$ and $\sup_y |\Delta_2(y)|$ converge to zero in probability by Lemmas 7.6 and 7.10, respectively. Thus, $\frac{1}{n}\sum_{i=1}^{n}\int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}]$ converges uniformly to $E\left(\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_1, \mathrm{sgn}\{d(\boldsymbol{H}_1,y)\}]\right)$, which implies the infimums converge. That is, $\widehat{y}^*_\tau = \inf\left(y : \frac{1}{n}\sum_{i=1}^{n}\int \widehat{F}_\epsilon(y-u-|v|)d\widehat{G}[u,v \mid \boldsymbol{H}_{1i}, \mathrm{sgn}\{\widehat{d}(\boldsymbol{H}_{1i},y)\}] \geq \tau\right)$ converges in probability to $y^*_\tau = \inf\left\{y : E\left(\int F_\epsilon(y-u-|v|)dG[u,v \mid \boldsymbol{H}_1, \mathrm{sgn}\{d(\boldsymbol{H}_1,y)\}]\right) \geq \tau\right\}$. ■

Lemma 7.12. *Assume A2–A4. Let $\boldsymbol{h}_1$ be fixed and arbitrary. Then, $\left|\widehat{d}(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,y^*_\tau)\right|$ converges to zero in probability.*

Proof. By the triangle inequality,

$$\left|\widehat{d}(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,y^*_\tau)\right| \leq \left|\widehat{d}(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,\widehat{y}^*_\tau)\right| + |d(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,y^*_\tau)|$$
$$\leq \sup_y \left|\widehat{d}(\boldsymbol{h}_1,y) - d(\boldsymbol{h}_1,y)\right| + |d(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,y^*_\tau)|.$$

The right-hand side of the previous expression is $o_p(1)$ because $\sup_y |\widehat{d}(\boldsymbol{h}_1,y) - d(\boldsymbol{h}_1,y)|$ is $o_p(1)$ by Lemma 7.2. Note that continuity of $d(\boldsymbol{h}_1,y)$ is implied by assumption A2, and thus, $|d(\boldsymbol{h}_1,\widehat{y}^*_\tau) - d(\boldsymbol{h}_1,y^*_\tau)|$ is $o_p(1)$ by Lemma 7.11 and the continuous mapping theorem. ■

Proof of Theorem 2.4. Let $\epsilon > 0$ be arbitrary. Then, because $\sup_{\boldsymbol{h}_1} |d(\boldsymbol{h}_1,y) - d(\boldsymbol{h}_1,y^*_\tau)|$ is continuous in y, there exists a $\delta > 0$ such that

$$\sup_{\boldsymbol{h}_1, y\in[y^*_\tau-\delta, y^*_\tau+\delta]} |d(\boldsymbol{h}_1,y) - d(\boldsymbol{h}_1,y^*_\tau)| < \epsilon. \tag{11}$$

We begin by showing $\sup_{y\in[y^*_\tau-\delta,y^*_\tau+\delta]} |\Delta(y)|$ converges to zero in probability, where $\Delta(y) = \mathrm{pr}^{\mathrm{sgn}\{\widehat{d}(\cdot,\widehat{y}^*_\tau)\},\widehat{\pi}^*_2}(Y \leq y) - \mathrm{pr}^{\mathrm{sgn}\{d(\cdot,y^*_\tau)\},\pi^*_2}(Y \leq y)$. That is, $\Delta(y)$ is the difference in the distribution function at y when treatments are assigned according to the estimated optimal regime versus the true optimal regime. By the triangle inequality,

$$\sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta(y)| \leq \sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta_1(y)| + \sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta_2(y)|\,, \tag{12}$$

where we define the terms $\Delta_1(y) = \mathrm{pr}^{\mathrm{sgn}\{\widehat{d}(\cdot,\widehat{y}_\tau^*)\},\widehat{\pi}_2^*}(Y \leq y) - \mathrm{pr}^{\mathrm{sgn}\{\widehat{d}(\cdot,\widehat{y}_\tau^*)\},\pi_2^*}(Y \leq y)$ and $\Delta_2(y) = \mathrm{pr}^{\mathrm{sgn}\{\widehat{d}(\cdot,\widehat{y}_\tau^*)\},\pi_2^*}(Y \leq y) - \mathrm{pr}^{\mathrm{sgn}\{d(\cdot,y_\tau^*)\},\pi_2^*}(Y \leq y)$. Note that $\sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta_1(y)| \leq \sup_{\pi_1,y} |\mathrm{pr}^{\pi_1,\widehat{\pi}_2^*}(Y \leq y) - \mathrm{pr}^{\pi_1,\pi_2^*}(Y \leq y)|$, where the right-hand side is $o_p(1)$ by Lemma 7.3. It can be shown that

$$\begin{aligned}
\sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta_2(y)| &\leq E\left(\left|\mathrm{sgn}\left\{\widehat{d}(\boldsymbol{H}_1,\widehat{y}_\tau^*)\right\} - \mathrm{sgn}\{d(\boldsymbol{H}_1,y_\tau^*)\}\right| |d(\boldsymbol{H}_1,y_\tau^*)| \mid \mathcal{D}\right) \\
&+ \sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} E\left(\frac{1}{2}\left|\mathrm{sgn}\left\{\widehat{d}(\boldsymbol{H}_1,\widehat{y}_\tau^*)\right\} - \mathrm{sgn}\{d(\boldsymbol{H}_1,y_\tau^*)\}\right| |d(\boldsymbol{H}_1,y) - d(\boldsymbol{H}_1,y_\tau^*)| \mid \mathcal{D}\right) \\
&\leq o_p(1) + \sup_{\boldsymbol{h}_1,y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |d(\boldsymbol{h}_1,y) - d(\boldsymbol{h}_1,y_\tau^*)|\,,
\end{aligned}$$

where the $o_p(1)$ term is based on Lemmas 7.1 and 7.12. We have already established that the second term on the right hand side is bounded by ϵ. Since ϵ was arbitrary, we have shown that $\sup_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]} |\Delta_2(y)|$ is $o_p(1)$. Thus, for y in a neighborhood of y_τ^*, $\Delta(y)$ converges uniformly in probability to zero. Noting that $y_\tau^* = \inf\left\{y : \mathrm{pr}^{\pi_{1,\tau}^{\mathrm{QIQ}},\pi_2^*}(Y \leq y) \geq \tau\right\}$ and $\mathrm{pr}^{\pi_{1,\tau}^{\mathrm{QIQ}},\pi_2^*}(Y \leq y_\tau^*) = \tau$, conclude that the infimums converge in a neighborhood of y_τ^*. That is, $\inf_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]}\left\{\mathrm{pr}^{\widehat{\pi}_{1,\tau}^{\mathrm{QIQ}},\widehat{\pi}_2^*}(Y \leq y) \geq \tau\right\}$ converges in probability to $q^{\pi_{1,\tau}^{\mathrm{QIQ}},\pi_2^*}(\tau) = \inf_{y\in[y_\tau^*-\delta,y_\tau^*+\delta]}\left\{\mathrm{pr}^{\pi_{1,\tau}^{\mathrm{QIQ}},\pi_2^*}(Y \leq y) \geq \tau\right\} = \inf\left\{y : \mathrm{pr}^{\pi_{1,\tau}^{\mathrm{QIQ}},\pi_2^*}(Y \leq y) \geq \tau\right\}$. ■

8. TIQ-LEARNING UNDER A MORE GENERAL REGRESSION MODEL

Suppose that $Y = \zeta(\boldsymbol{H}_2, A_2, \epsilon)$ where ϵ is independent of $\boldsymbol{H}_2$, A_2, and $\boldsymbol{H}_2$ contains first-stage information, X_1 and A_1. For any $\boldsymbol{h}_2$ and a_2 write $\zeta_{\boldsymbol{h}_2,a_2}(u)$ as real-valued function on the domain of ϵ so that $\zeta_{\boldsymbol{h}_2,a_2}(u) = \zeta(\boldsymbol{h}_2, a_2, u)$. We assume that for almost all $\boldsymbol{h}_2$, a_2 the function $\zeta_{\boldsymbol{h}_2,a_2}(\cdot)$ is invertible. Special cases include: (i) the additive error model considered in the main body, $\zeta(\boldsymbol{h}_2, a_2, \epsilon) = m(\boldsymbol{h}_2) + a_2 c(\boldsymbol{h}_2) + \epsilon$; and (ii) a multiplicative error model, $\zeta(\boldsymbol{h}_2, a_2, \epsilon) = \epsilon\,[m(\boldsymbol{h}_2) + a_2 c(\boldsymbol{h}_2)]$ provided $\epsilon > 0$ with probability one and $\mathrm{pr}\,\{m(\boldsymbol{H}_2) + A_2 c(\boldsymbol{H}_2) = 0\} = 0$. Let $\boldsymbol{\pi} = (\pi_1, \pi_2)$ denote an arbitrary dynamic treatment regime of interest. Applying the same arguments as the main body, it can be shown that

$$\mathrm{pr}^{\pi_1,\,\pi_2}\,(Y \leq y) = \int\int F_\epsilon\left\{\zeta^{-1}_{\boldsymbol{h}_2,\pi_2(\boldsymbol{h}_2)}\,(y)\right\} dF_{\boldsymbol{H}_2|\boldsymbol{H}_1,A_1}\,\{\boldsymbol{h}_2|\boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\}\, dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1).$$

For a patient presenting with history $\boldsymbol{h}_2$, the optimal decision rule at the second stage is thus $\pi_2^*(\boldsymbol{h}_2) = \arg\max_{a_2} \zeta^{-1}_{\boldsymbol{h}_2,a_2}(y)$. Let $G\left\{\cdot,\cdot\middle|\boldsymbol{h}_1, a_1\right\}$ denote joint distribution of $\left\{\zeta^{-1}_{\boldsymbol{H}_2,1}(y), \zeta^{-1}_{\boldsymbol{H}_2,-1}(y)\right\}$. Then,

$$\mathrm{pr}^{\pi_1,\,\pi_2^*}\,(Y \leq y) = \int\int F_\epsilon\left\{(u+v)/2 + |u-v|/2\right\} dG\left\{u, v\middle|\boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\right\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1).$$

Therefore, the optimal first stage decision rule is

$$\pi_1^*(\boldsymbol{h}_1) = \arg\max_{a_1} \int F_\epsilon\left\{(u+v)/2 + |u-v|/2\right\} dG\left\{u, v\middle|\boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\right\}.$$

Estimation of the optimal dynamic treatment regime using TIQ-learning therefore consists on the following steps: (i) choose the form of the regression model $Y = \zeta(\boldsymbol{H}_2, A_2, \epsilon)$; (ii) construct an estimator $\widehat{\zeta}(\boldsymbol{h}_2, a_2, u)$ of $\zeta(\boldsymbol{h}_2, a_2, u)$ and estimator $\widehat{F}_\epsilon(u)$ of $F_\epsilon(u)$; (iii) use pairs $\left\{(\boldsymbol{H}_{1,i}, A_{1,i}, \widehat{\zeta}^{-1}_{\boldsymbol{H}_{2,i},1}(y), \widehat{\zeta}^{-1}_{\boldsymbol{H}_{2,i},1}(y))\right\}_{i=1}^{n}$ to construct an estimator of $\widehat{G}(\cdot,\cdot|\boldsymbol{h}_1, a_1)$ of $G(\cdot,\cdot|\boldsymbol{h}_1, a_1)$; and (iv) define the estimated optimal regime using TIQ to be $\widehat{\pi}_2^*(\boldsymbol{h}_2) = \arg\max_{a_2} \widehat{\zeta}^{-1}_{\boldsymbol{h}_2,a_2}(y)$ and $\widehat{\pi}_1(\boldsymbol{h}_1) = \arg\max_{a_1} \int \widehat{F}_\epsilon\left\{(u+v)/2 + |u-v|/2\right\} d\widehat{G}\left\{u, v\middle|\boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\right\}$.

Analogous derivations can be used to construct an estimator that optimizes a quantile of the outcome distribution under a general regression model for Y.

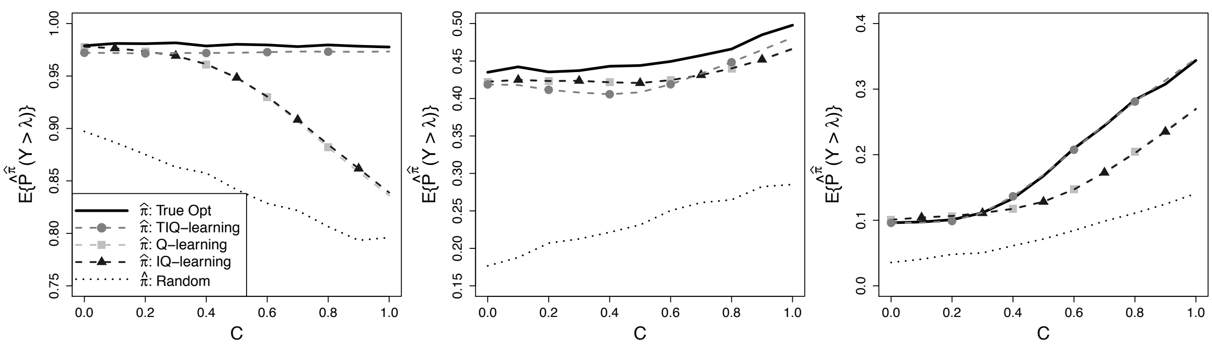

Figure 2: *Left to Right:* $\lambda = -2, 2, 4$. Solid black, true optimal threshold probabilities; dotted black, probabilites under randomization; dashed with circles/squares/triangles, probabilities under TIQ-, Q-, and Interactive Q-learning, respectively. Training set size of $n = 100$.

9. TIQ-LEARNING WITH BIVARIATE KERNEL DENSITY ESTIMATOR

As in the main paper, the data are generated using the model

$$\begin{aligned}
&\boldsymbol{X}_1 \sim \text{Norm}(\mathbf{1}_2, \boldsymbol{\Sigma}), && A_1, A_2 \sim \text{Unif}\{-1, 1\}^2, && \boldsymbol{H}_1 = (1, \boldsymbol{X}_1^{\intercal})^{\intercal}, \\
&\eta_{\boldsymbol{H}_1, A_1} = \exp\{\tfrac{C}{2}(\boldsymbol{H}_1^{\intercal}\boldsymbol{\gamma}_0 + A_1\boldsymbol{H}_1^{\intercal}\boldsymbol{\gamma}_1)\}, && \boldsymbol{\xi} \sim \text{Norm}(\mathbf{0}_2, \mathbf{I}_2), && \boldsymbol{X}_2 = \boldsymbol{B}_{A_1}\boldsymbol{X}_1 + \eta_{\boldsymbol{H}_1, A_1}\boldsymbol{\xi}, \\
&\boldsymbol{H}_2 = (1, \boldsymbol{X}_2^{\intercal})^{\intercal}, && \epsilon \sim \text{Norm}(0, 1), && Y = \boldsymbol{H}_2^{\intercal}\boldsymbol{\beta}_{2,0} + A_2\boldsymbol{H}_2^{\intercal}\boldsymbol{\beta}_{2,1} + \epsilon,
\end{aligned}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of 1s, $\mathbf{I}_q$ is the $q \times q$ identify matrix, and $C \in [0, 1]$ is a constant. The matrix $\boldsymbol{\Sigma}$ is a correlation matrix with off-diagonal $\rho = 0.5$. The 2×2 matrix $\boldsymbol{B}_{A_1}$ equals

$$\boldsymbol{B}_{A_1=1} = \begin{pmatrix} -0.1 & -0.1 \\ 0.1 & 0.1 \end{pmatrix}, \qquad \boldsymbol{B}_{A_1=-1} = \begin{pmatrix} 0.5, & -0.1 \\ -0.1 & 0.5 \end{pmatrix}.$$

The remaining parameters are $\boldsymbol{\gamma}_0 = (1, 0.5, 0)^{\intercal}$, $\boldsymbol{\gamma}_1 = (-1, -0.5, 0)^{\intercal}$, $\boldsymbol{\beta}_{2,0} = (0.25, -1, 0.5)^{\intercal}$, and $\boldsymbol{\beta}_{2,1} = (1, -0.5, -0.25)^{\intercal}$, which were chosen to ensure that the mean-optimal treatment produced a more variable response for some patients.

Results are based on $J = 1,000$ generated data sets; for each, we estimate the TIQ-, IQ-, and Q-learning policies and compare the results using a test set of size $N = 10,000$. We compare training sample sizes of $n = 100$ and $n = 250$. The normal scale model is used to estimate $F_{\epsilon}(\cdot)$, which is correctly specified for the generative model above. A bivariate

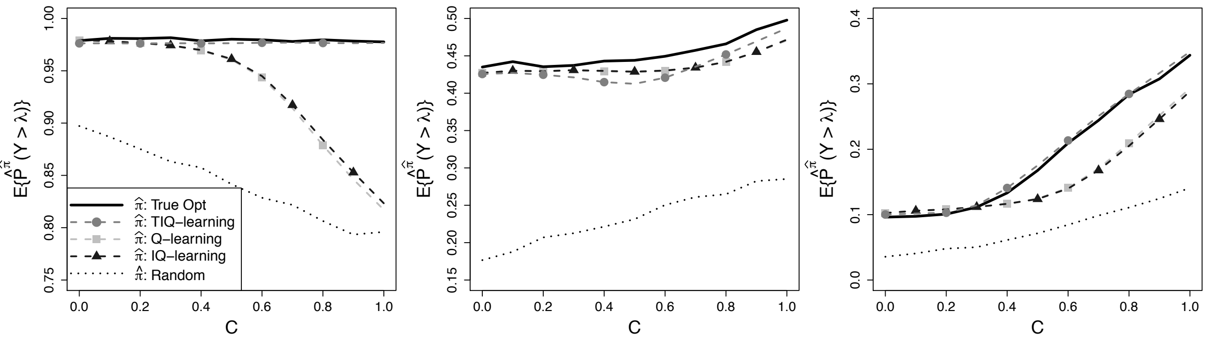

Figure 3: *Left to Right:* $\lambda = -2, 2, 4$. Solid black, true optimal threshold probabilities; dotted black, probabilites under randomization; dashed with circles/squares/triangles, probabilities under TIQ-, Q-, and Interactive Q-learning, respectively. Training set size of $n = 250$.

kernel density estimator is used to estimate $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$.

To study the performance of the TIQ-learning algorithm, we compare values of the cumulative distribution function of the final response when treatment is assigned according to the estimated TIQ-learning, IQ-learning, and Q-learning regimes. Define $\mathrm{pr}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ to be the true probability that Y exceeds λ given treatments are assigned according to $\widehat{\boldsymbol{\pi}}_j = (\widehat{\pi}_{1j}, \widehat{\pi}_{2j})$, the regime estimated from the j^{th} generated data set. For threshold values $\lambda = -2, 2, 4$, we estimate $\mathrm{pr}^{\boldsymbol{\pi}}(Y > \lambda)$ using $\sum_{j=1}^{J} \widehat{\mathrm{pr}}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)/J$, where $\widehat{\mathrm{pr}}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ is an estimate of $\mathrm{pr}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ obtained by calculating the proportion of test patients consistent with regime $\widehat{\boldsymbol{\pi}}_j$ whose observed Y values are greater than λ. Thus, our estimate is an average over training data sets and test set observations. In terms of the proportion of distribution mass above λ, results for $\lambda = -2$ and 4 in Figures 2 and 3 show a clear advantage of TIQ-learning for higher values of C, the degree of heteroskedasticity in the second-stage covariates $\boldsymbol{X}_2$. As anticipated by Remark 1 in Section 2.1 of the main paper, the methods perform similarly when $\lambda = 2$. Results appear similar for the sample sizes $n = 100$ and $n = 250$ considered here, suggesting good performance of the nonparametric bivariate kernel estimator for reasonable sample sizes.

REFERENCES

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer New York.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.