Interactive Q-learning for Probabilities and 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

July 15, 2014
Author’s Footnote:

Kristin A. Linn is Postdoctoral Fellow, Department of Biostatistics and Epidemiology, University of Pennsylvania. Mailing address: University of Pennsylvania School of Medicine, Center for Clinical Epidemiology and Biostatistics (CCEB), 211 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104; Eric B. Laber is Assistant Professor, Department of Statistics, North Carolina State University; Leonard A. Stefanski is Drexel Professor, Department of Statistics, North Carolina State University. Eric Laber acknowledges support from NIH grant P01 CA142538 and DNR grant PR-W-F14AF00171. Leonard Stefanski acknowledges support from NIH grants R01 CA085848 and P01 CA142538 and NSF grant DMS-0906421.
Abstract

A dynamic treatment regime is a sequence of decision rules in which each decision rule recommends treatment based on features of patient medical history such as past treatments and outcomes. Existing methods for estimating optimal dynamic treatment regimes from data optimize the mean of a response variable. However, the mean may not always be the most appropriate summary of performance. We derive estimators of decision rules for optimizing probabilities and quantiles computed with respect to the response distribution for two-stage, binary treatment settings. This enables estimation of dynamic treatment regimes that optimize the cumulative distribution function of the response at a prespecified point or a prespecified quantile of the response distribution such as the median. The proposed methods perform favorably in simulation experiments. We illustrate our approach with data from a sequentially randomized trial where the primary outcome is remission of depression symptoms.

KEYWORDS: Dynamic Treatment Regime; Personalized Medicine; Sequential Decision Making; Sequential Multiple Assignment Randomized Trial.
1. INTRODUCTION

A dynamic treatment regime operationalizes clinical decision making as a series of decision rules that dictate treatment over time. These rules take into account accrued patient medical history, including past treatments and outcomes. Each rule maps current patient characteristics to a recommended treatment, hence personalizing treatment to the individual. Typically, a dynamic treatment regime is estimated from data with the goal of optimizing the expected value of a clinical outcome, and the resulting regime is referred to as the estimated optimal regime. Current methods for estimating optimal dynamic treatment regimes include Q-learning (Watkins, 1989; Watkins and Dayan, 1992; Murphy, 2005a), regularized Q-learning (Moodie and Richardson, 2010; Chakraborty et al., 2010; Song et al., 2011; Goldberg et al., 2013), Interactive Q-learning (Laber et al., 2014), g-estimation (Robins, 2004), A-learning (Murphy, 2003), regret-regression (Henderson et al., 2010), and policy search methods (Orellana et al., 2010; Zhao et al., 2012; Zhang et al., 2012b,a, 2013). Extensions exist for estimating dynamic treatment regimes under censoring (Goldberg and Kosorok, 2012), high-dimensional data (McKeague and Qian, 2013), and missing data (Shortreed et al., 2014).

Despite many estimation methods, none are designed to handle functionals of the response distribution other than the mean, such as probabilities or quantiles. Using the potential outcomes framework (Rubin, 1974; Rosenbaum and Rubin, 1983), Zhang et al. (2012c) develop methods for estimating quantiles of the potential outcomes from observational data. However, they focus on comparing treatments at a single intervention time point rather than estimation of an optimal dynamic treatment regime.

The Q-learning algorithm is an approximate dynamic programming procedure that requires modeling nonsmooth, nonmonotone transformations of data; this leads to nonregular estimators and complicates the search for models that fit the data well (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2010; Song et al., 2011). Interactive Q-learning (IQ-learning), developed for the two-stage binary treatment setting, requires modeling only smooth, monotone transformations of the data, thereby reducing problems of model misspecification and nonregular inference (Laber et al., 2014). We extend the IQ-learning framework to handle optimization of functionals of the outcome distribution other than the expected value. In particular, we consider threshold-exceedance probabilities and quantiles of the response distribution.

Threshold-exceedance probabilities are relevant in clinical applications where the primary objective is remission or a specific target for symptom reduction. For example, consider a population of obese patients enrolled in a study to determine the effects of several treatment options for weight loss. The treatments of interest may include combinations of drugs, exercise programs, counseling, and meal plans (for example, see Berkowitz et al., 2010). Our method can be used to maximize the probability that patients achieve a weight below some prespecified, patient-specific threshold at the conclusion of the study. With adjustments to the method of maximizing probabilities, we also derive optimal decision rules for maximizing quantiles of the response distribution. Both frameworks can be used to study how the optimal regime changes as the target probability or quantile is varied. In addition, the quantile framework provides an analog of quantile regression in the dynamic treatment regime setting for constructing robust estimators; for example, it enables optimization of the median.
2. GENERALIZED INTERACTIVE Q-LEARNING

We assume data have been collected using a two-stage sequential multiple assignment randomized trial (Lavori and Dawson, 2000, 2004; Murphy, 2005b) with two treatments at each stage. This set-up facilitates a focused discussion of the proposed methods and is also useful in practice, as data in many sequentially randomized trials have this structure (Projects Using SMART, 2012; Laber, 2013). Although our developments are made under the assumption of treatment randomization, additional assumptions about the treatment assignment mechanism allow application of the proposed methods to observational data (Murphy, 2003; Robins, 2004; Moodie et al., 2012; Schulte et al., 2014).

In the two-stage, binary treatment setting, \( D = \{(X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_{i})\}_{i=1}^{n} \) represents the observed data which consist of time-ordered trajectories that are independent and identically distributed across \( n \) patients. Let \((X_1, A_1, X_2, A_2, Y)\) denote an observation for a single subject where: \( X_1 \in \mathbb{R}^{p_1} \) denotes a vector of baseline covariate information collected prior to randomization; \( A_1 \in \{-1,1\} \) denotes the first randomized treatment; \( X_2 \in \mathbb{R}^{p_2} \) denotes a vector of covariate information collected during the course of the first treatment but prior to the assignment of the second treatment; \( A_2 \in \{-1,1\} \) denotes the second randomized treatment; and \( Y \in \mathbb{R} \) is an outcome measured at the conclusion of stage two, coded so that higher values are more desirable. Capital letters denote random variables and lower case letters denote observed instances.

We consider the problem of optimizing distributional summaries of \( Y \) other than the mean. To simplify notation, define patient histories \( H_1 \) and \( H_2 \), where \( H_1 = X_1 \) is information collected prior to the first-stage treatment randomization, and \( H_2 = (H_1^\top, A_1, X_2^\top)^\top \) is all information available to the decision-maker prior to the second-stage treatment randomization. A regime \( \pi \) is a pair of decision rules \((\pi_1, \pi_2)\), where \( \pi_1 \) is a map from the support of \( H_1 \) to the space of possible first-stage treatments, and \( \pi_2 \) is a map from the support of \( H_2 \) to the space of possible second-stage treatments. That is, \( \pi_1 : \text{dom}(H_1) \mapsto \text{dom}(A_1) \), and \( \pi_2 : \text{dom}(H_2) \mapsto \text{dom}(A_2) \). Because \( A_2 \) is binary, there exist functions \( m \) and \( c \) such that \( E(Y \mid A_2, H_2) = m(H_2) + A_2c(H_2) \). We assume \( Y = E(Y \mid A_2, H_2) + \epsilon \) for mean-zero error term \( \epsilon \). Additionally, we assume \( \text{var}(\epsilon) < \infty \) and \( \epsilon \) is independent of \( A_2 \) and \( H_2 \).

Extension to the heteroskedastic setting where the variance of \( \epsilon \) depends on past covariates and treatments is provided in the supplementary material that appears after the references in this document.

2.1 Threshold Interactive Q-learning

Let \( pr^{\pi_1, \pi_2}(Y > \lambda) \) denote the probability that the outcome \( Y \) is greater than a predefined threshold \( \lambda \) under treatment assignment dictated by the regime \( \pi = (\pi_1, \pi_2) \). Threshold Interactive Q-learning (TIQ-learning) maximizes \( pr^{\pi_1, \pi_2}(Y > \lambda(H_i) \mid H_1) \) for all \( H_1 \) with respect to \( \pi = (\pi_1, \pi_2) \), where \( \lambda(H_i) \) is a threshold that depends on the \( t \)-th stage history, either \( t = 1 \) or 2. Here we consider a constant threshold, \( \lambda(H_i) = \lambda \); patient-specific thresholds are discussed in the supplementary materials.

Define the following: \( F_{H_1}(\cdot) \) is the distribution of \( H_1 \); \( F_{H_2 \mid H_1, A_1}(\cdot \mid h_1, a_1) \) is the conditional distribution of \( H_2 \) given \( H_1 = h_1 \) and \( A_1 = a_1 \); \( F_i(\cdot) \) is the distribution of
\( \epsilon \); and \( \mathbf{H}_{2}^{\pi_1(H_1)} \) is the second-stage history induced by treating according to decision rule \( \pi_1 \) at the first stage, i.e., \( \mathbf{H}_{2}^{\pi_1(H_1)} = \{ \mathbf{H}_1, \pi_1(\mathbf{H}_1), \mathbf{X}_2 \} \). Also define \( J^{\pi_1, \pi_2}(h_1, h_2, y) = F_c(y - m(h_2^{\pi_1(h_1)}) - \pi_2(h_2^{\pi_1(h_1)})c(h_2^{\pi_1(h_1)})) \), the distribution of \( \epsilon \) evaluated at \( y - m(h_2^{\pi_1(h_1)}) - \pi_2(h_2^{\pi_1(h_1)})c(h_2^{\pi_1(h_1)}) \), and

\[
\Pr^{\pi_1, \pi_2}(Y \leq y) = \int \int J^{\pi_1, \pi_2}(h_1, h_2, y) dF_{H_2 \mid H_1, A_1 \{ h_2 \mid h_1, \pi_1(h_1) \}} dF_{H_1}(h_1). \tag{1}
\]

The previous expression is the expected value of \( J^{\pi_1, \pi_2}(H_1, H_2, y) \).

Let \( \pi_2^*(h_2) = \text{sgn}\{c(h_2)\} \), where we define \( \text{sgn}(x) = 1_{x \geq 0} - 1_{x < 0} \). Then, noting that \( \pi_2(h_2^{\pi_1(h_1)})c(h_2^{\pi_1(h_1)}) \leq |c(h_2^{\pi_1(h_1)})| \) for all \( h_2^{\pi_1(h_1)} \),

\[
\Pr^{\pi_1, \pi_2}(Y \leq y) \geq \int \int J^{\pi_1, \pi_2^*}(h_1, h_2, y) dF_{H_2 \mid H_1, A_1 \{ h_2 \mid h_1, \pi_1(h_1) \}} dF_{H_1}(h_1), \tag{2}
\]

where \( J^{\pi_1, \pi_2^*}(h_1, h_2, y) = F_c(y - m(h_2^{\pi_1(h_1)}) - |c(h_2^{\pi_1(h_1)})|) \). Denote the right-hand side of (2) by \( \Pr^{\pi_1, \pi_2^*}(Y \leq y) \). Let \( G(\cdot, \cdot \mid h_1, a_1) \) denote the joint conditional distribution of \( m(H_2) \) and \( c(H_2) \) given \( H_1 = h_1 \) and \( A_1 = a_1 \). That is, \( G(u, v \mid h_1, a_1) = \Pr\{m(H_2) \leq u, c(H_2) \leq v \mid H_1 = h_1, A_1 = a_1 \} \). Then, using the fact that \( G(u, v \mid h_1, a_1) = \int \mathbb{1}_{m(h_2) \leq u} \mathbb{1}_{c(h_2) \leq v} dF_{H_2 \mid H_1, a_1} \{ h_2 \mid h_1, a_1 \} \), one can show by an interchange in the order of integration that

\[
\int \int F_c(y - u - |v|) dG\{u, v \mid h_1, \pi_1(h_1)\} dF_{H_1}(h_1) = \int \int J^{\pi_1, \pi_2^*}(h_1, h_2, y) dF_{H_2 \mid H_1, A_1 \{ h_2 \mid h_1, \pi_1(h_1) \}} dF_{H_1}(h_1),
\]

where the right-hand side is \( \Pr^{\pi_1, \pi_2^*}(Y \leq y) \) by definition. Define

\[
I\{y, F_c(\cdot), G(\cdot, \cdot \mid h_1, a_1)\} = \int F_c(y - u - |v|) dG(u, v \mid h_1, a_1). \tag{3}
\]

Then,

\[
\Pr^{\pi_1, \pi_2^*}(Y \leq y) = E(I[y, F_c(\cdot), G(\cdot, \cdot \mid H_1, \pi_1(H_1))]). \tag{4}
\]

The \( \lambda \)-optimal regime \( \pi_{\text{TIQ}}^{\lambda} = \{ \pi_{\text{TIQ}}^{\lambda_1, \lambda_2}, \pi_{\text{TIQ}}^{\lambda_2, \lambda} \} \) is determined by the requirement that \( \Pr^{\pi_{\text{TIQ}}^{\lambda}}(Y > \lambda) \geq \Pr^{\pi}(Y > \lambda) \) for all \( \pi = (\pi_1, \pi_2) \); or equivalently, that \( \Pr^{\pi_{\text{TIQ}}^{\lambda}}(Y \leq \lambda) \leq \Pr^{\pi}(Y \leq \lambda) \) for all \( \pi \). That is, the distribution of \( Y \) induced by regime \( \pi_{\text{TIQ}}^{\lambda} \) has at least as much mass above \( \lambda \) as the distribution of \( Y \) induced by any other regime. It follows from the lower bound on \( \Pr^{\pi_1, \pi_2^*}(Y \leq y) \) displayed in (2) that \( \pi_{\text{TIQ}}^{\lambda_2, \lambda}(h_2) = \pi_{\text{TIQ}}^{\lambda_2^*(h_2)} = \text{sgn}\{c(h_2)\} \) for all \( h_2 \), independent of \( \lambda \) and \( \pi_{\text{TIQ}}^{\lambda_1, \lambda} \). Henceforth, we denote \( \pi_{\text{TIQ}}^{\lambda_2, \lambda} \) by \( \pi_{\lambda_2}^* \). It follows from (4) that

\[
\Pr^{\pi_1, \pi_2^*}(Y > \lambda) = 1 - E\left[I[\lambda, F_c(\cdot), G(\cdot, \cdot \mid H_1, \pi_1(H_1))]|A_1\right] 
\leq 1 - E\left[\min_{A_1} I[\lambda, F_c(\cdot), G(\cdot, \cdot \mid H_1, A_1)]\right], \tag{5}
\]
thus showing that the $\lambda$-optimal first-stage rule is $\pi_{1,\lambda}^{\text{TQ}}(h_1) = \arg\min_{a_1} I(\lambda, F_\epsilon(\cdot), G(\cdot, \cdot | h_1, a_1))$.

Inequality (5) holds because $I(\lambda, F_\epsilon(\cdot), G(\cdot, \cdot | H_1, a_1))$ is minimized over $a_1$ for all $H_1$. Alternatively, defining

\[ d(h_1, \lambda) = I(\lambda, F_\epsilon(\cdot), G(\cdot, \cdot | h_1, -1)) - I(\lambda, F_\epsilon(\cdot), G(\cdot, \cdot | h_1, 1)), \]

the $\lambda$-optimal first-stage rule can be written as $\pi_{1,\lambda}^{\text{TQ}}(h_1) = \{d(h_1, \lambda)\}$.

We describe the general form of the TIQ-learning algorithm that can be used to estimate the $\lambda$-optimal regime in the case that $m$ and $c$ belong to parametric families, denoted $m(H_2; \beta_{2,0})$ and $c(H_2; \beta_{2,1})$. The exact algorithm depends on the choice of estimators for $F_\epsilon(\cdot)$ and $G(\cdot, \cdot | h_1, a_1)$. We discuss possible estimators in Sections 2.3 and 2.4 in practice the choice of estimators should be informed by the observed data. Define

\[ d(h_1, \lambda) = I(\lambda, \hat{F}_\epsilon(\cdot), \hat{G}(\cdot, \cdot | h_1, -1)) - I(\lambda, \hat{F}_\epsilon(\cdot), \hat{G}(\cdot, \cdot | h_1, 1)). \]

**The TIQ-learning algorithm:**

1. **TIQ.1** Estimate $m(H_2; \beta_{2,0})$ and $c(H_2; \beta_{2,1})$ by fitting the model $Y = m(H_2; \beta_{2,0}) + A_2c(H_2; \beta_{2,1}) + \epsilon$, e.g., using least squares. Given $h_2$, estimate $\pi_2^*$ using the plug-in estimator $\hat{\pi}_2^*(h_2) = \text{sgn}(c(h_2; \hat{\beta}_{2,1}))$.

2. **TIQ.2** Estimate $F_\epsilon(\cdot)$, the cumulative distribution function of $\epsilon$, using the residuals $\hat{\epsilon}^Y = Y - m(H_2; \beta_{2,0}) - A_2c(H_2; \hat{\beta}_{2,1})$ from TIQ.1. Let $\hat{F}_\epsilon(\cdot)$ denote this estimator.

3. **TIQ.3** Estimate $G(\cdot, \cdot | h_1, a_1)$, the joint conditional distribution of $m(H_2)$ and $c(H_2)$ given $H_1 = h_1$ and $A_1 = a_1$. Let $\hat{G}(\cdot, \cdot | h_1, a_1)$ denote this estimator.

4. **TIQ.4** Given $h_1$, estimate $\pi_{1,\lambda}^{\text{TQ}}$ using the plug-in estimator $\hat{\pi}_{1,\lambda}^{\text{TQ}}(h_1) = \{d(h_1, \lambda)\}$.

TIQ-learning requires more modeling than mean-optimal techniques such as Q-learning or IQ-learning. The Q-learning algorithm involves modeling $m$, $c$, and the conditional expectation of $m(H_2) + c(H_2)$ given $H_1$ and $A_1$. Interactive Q-learning requires modeling $m$, $c$, and the conditional expectation of $m(H_2)$ and the one-dimensional conditional density of $c(H_2)$ given $H_1$ and $A_1$ (Laber et al., 2014). In comparison, the TIQ-learning algorithm involves modeling $m$, $c$, the distribution function $F_\epsilon(\cdot)$, and the bivariate conditional density $G(\cdot, \cdot | h_1, a_1)$, which we discuss in Sections 2.3 and 2.4.

**Remark 2.1.** Let $\pi_1^M$ denote the first-stage decision rule of a mean optimal regime. Then, assuming the setup of Section 2 it can be shown

\[ \pi_1^M(h_1) = \arg\min_{a_1} \int (-u - |v|)dG(u, v | h_1, a_1) = \arg\min_{a_1} \int (\lambda - u - |v|)dG(u, v | h_1, a_1), \]

whereas $\pi_{1,\lambda}^{\text{TQ}}(h_1) = \arg\min_{a_1} \int F_\epsilon(\lambda - u - |v|)dG(u, v | h_1, a_1)$. If $F_\epsilon(\cdot)$ is approximately linear where the conditional distribution of $\lambda - m(H_2) - c(H_2)$ given $H_1 = h_1$ and $A_1 = a_1$ is concentrated, $\pi_1^M(h_1)$ and $\pi_{1,\lambda}^{\text{TQ}}(h_1)$ will likely agree. Thus, the difference between the mean
optimal and TIQ-learning optimal regimes can be compared empirically by computing
\[
\arg \min_{a_1} \int (-u - |v|) d\hat{G}(u, v \mid h_{1i}, a_1), \quad \arg \min_{a_1} \int \hat{F}_e(\lambda - u - |v|) d\hat{G}(u, v \mid h_{1i}, a_1),
\]
for each first-stage patient history \(h_{1i}, i = 1, \ldots, n\), and considering the proportion of patients for which these two rules differ.

### 2.2 Quantile Interactive Q-learning

Under some generative models, assigning treatment according to a mean-optimal regime leads to higher average outcomes at the expense of higher variability, negatively affecting patients with outcomes in the lower quantiles of the induced distribution of \(Y\). We demonstrate this using simulated examples in Section 3. Define the \(\tau^{th}\) quantile of the distribution of \(Y\) induced by regime \(\pi\) as \(q^\pi(\tau) \triangleq \inf\{y : \text{pr}\{\pi(Y \leq y) \geq \tau\}\} \). The goal of Quantile Interactive Q-learning (QIQ-learning) is to estimate a pair of decision rules, \(\pi_{QIQ} = \{\pi_{1,\tau}, \pi_{2,\tau}\}\), that maximize \(q^\pi(\tau)\) over \(\pi\) for a fixed, prespecified \(\tau\). QIQ-learning is similar to TIQ-learning, but the optimal first-stage rule is complicated by the inversion of the distribution function to obtain quantiles of \(Y\) under a given regime. Under the model assumptions of Section 2 the QIQ-learning second-stage optimal decision is the same as that of TIQ-learning, i.e., \(\pi^*_{QIQ}(h_2) = \pi^*_{2}(h_2) = \text{sgn}\{c(h_2)\}\), independent of \(\tau\) and \(\pi_{QIQ}^\tau\); see the supplementary materials for a derivation of this rule. We henceforth denote \(\pi_{QIQ}^\tau\) by \(\pi^*_2\).

Next we discuss the existence of an optimal first-stage decision rule, \(\pi_{1,\tau}\), which in turn motivates an algorithm for calculating it. Define
\[
\Gamma(h_1, y) \triangleq \text{sgn}\{d(h_1, y)\}, \quad \tag{7}
\]
\[
y^*_\tau \triangleq \inf\left\{y : \text{pr}\{\pi(Y \leq y) \geq \tau\}\right\}, \quad \tag{8}
\]
where \(d(h_1, y)\) is defined in [6]. Note that \(\Gamma(h_1, y) = \pi_{QIQ}^{1,\lambda}(h_1)|_{\lambda = y}\), and thus it is the optimal first-stage decision rule of TIQ-learning evaluated at \(\lambda = y\). We have introduced the new notation to emphasize the dependence on \(y\).

We show in Lemma 7 of the supplementary material that \(\lim_{y \to \infty(-\infty)} \text{pr}\{\Gamma(Y) = \pi^*_2(Y \leq y)\} = 1(0)\), so that \(y^*_\tau\) is defined for all \(\tau \in (0, 1)\). Furthermore, \(y^*_\tau\) is an upper bound on \(q^\pi_{\tau}(\tau)\), the quantile observed when treating according to \(\pi_1\) at stage one and the optimal rule at stage two. To see this, for each \(y \in \mathbb{R}\),
\[
\text{pr}\{\pi^*_1, \pi^*_2(Y \leq y)\} = E\{I[y, F_\epsilon(\cdot), G(\cdot \mid H_1, \pi_1(H_1)))]\}
\geq E\{I[y, F_\epsilon(\cdot), G(\cdot \mid H_1, \Gamma(H_1, y))]\}
= \text{pr}\{\pi(Y) = \pi^*_2(Y \leq y)\},
\]
where \(I(\cdot, \cdot, \cdot)\) is defined in [3]. The last equality follows because \(\Gamma(H_1, y)\) minimizes \(E\{I[y, F_\epsilon(\cdot), G(\cdot \mid H_1, a_1)]\}\) with respect to \(a_1\). Hence, \(\{y : \text{pr}\{\Gamma(Y) = \pi^*_2(Y \leq y)\} \geq \tau\} \subseteq \)
Lemma 2.2.

Our main results depend on the following lemma, proved in the supplementary material.

We now discuss conditions that guarantee existence of a \( \pi_1 \) such that \( q^{\pi_1, \pi_2^*}(\tau) = y^*_\tau \) and derive its form. Note that

\[
y^*_\tau \geq q^{\pi_1, \pi_2^*}(\tau) \text{ for all } \pi_1.
\]  

(9)

Thus, a first-stage decision rule \( \pi_1 \) is optimal if it induces a \( \tau \)-th quantile equal to the upper bound \( y^*_\tau \) when treatments are subsequently assigned according to \( \pi_2^* \), i.e., if \( q^{\pi_1, \pi_2^*}(\tau) = y^*_\tau \).

We now discuss conditions that guarantee existence of a \( \pi_1 \) such that \( q^{\pi_1, \pi_2^*}(\tau) = y^*_\tau \) and derive its form.

\[
f(y) \triangleq q^{\Gamma(\cdot, y), \pi_2^*}(\tau) = \inf \left\{ \tilde{y} : \text{pr}^{\Gamma(\cdot, y), \pi_2^*}(Y \leq \tilde{y}) \geq \tau \right\}
\]  

(10)

is the quantile obtained under regime \( \pi = \{ \Gamma(\cdot, y), \pi_2^* \} \). Thus, because it is a quantile and the bound in (9) applies, we have

\[
\text{pr}^{\Gamma(\cdot, y), \pi_2^*}(Y \leq f(y)) \geq \tau, \quad f(y) = q^{\Gamma(\cdot, y), \pi_2^*}(\tau) \leq y^*_\tau \text{ for all } y.
\]  

(11)

Our main results depend on the following lemma, proved in the supplementary material.

**Lemma 2.2.**

\[
\begin{align*}
(A) & \quad y < y^*_\tau \text{ implies } y < f(y) \leq y^*_\tau; \\
(B) & \quad f(y^*_\tau^-) \triangleq \lim_{\delta \downarrow 0} f(y^*_\tau - \delta) = y^*_\tau; \\
(C) & \quad f(y^*_\tau) \leq y^*_\tau \text{ with strict inequality if there exists } \delta > 0 \text{ such that } \\
& \quad \text{pr}^{\Gamma(\cdot, y^*_\tau), \pi_2^*}(Y \leq y^*_\tau - \delta) \geq \tau; \\
(D) & \quad \text{If } F_\epsilon(\cdot) \text{ is continuous and strictly increasing, then } f(y^*_\tau) = y^*_\tau.
\end{align*}
\]

It follows from (B) that \( f(y^*_\tau) = y^*_\tau \) if and only if \( f(y) \) is left continuous at \( y = y^*_\tau \), and part (D) is a sufficient condition guaranteeing left-continuity of \( f(y) \) at \( y^*_\tau \). In this case, the optimal first-stage rule is \( \pi_{1,\tau}^{Q IQ}(h_1) = \Gamma(h_1, y^*_\tau) \), i.e., \( q^{\Gamma(\cdot, y^*_\tau), \pi_2^*}(\tau) = y^*_\tau \). The condition stated in (D) is commonly satisfied, for example, when the density of \( \epsilon \) has positive support on the entire real line. If \( f(y) \) is not left continuous at \( y^*_\tau \), and thus \( f(y^*_\tau) < y^*_\tau \), in light of (18) we can always approach the optimal policy via a sequence of regimes of the form \( \{ \Gamma(\cdot, y^*_\tau - \delta_n), \pi_2^* \} \), where \( \delta_n \) decreases to 0. We discuss this further below. If the underlying distributions of the histories and \( Y \) were known, the algorithm below produces an optimal regime.

**Population-level algorithm to find \( \pi_{1,\tau}^{Q IQ} \):**

1. Compute \( y^*_\tau \) from [8] and \( f(y^*_\tau) \) from (10).
2. a. If \( f(y^*_\tau) = y^*_\tau \), \( \pi_{1,\tau}^{Q IQ}(h_1) = \Gamma(h_1, y^*_\tau) \) is optimal because it attains the optimal quantile \( y^*_\tau \).
   b. If \( f(y^*_\tau) < y^*_\tau \), \( \lim_{\delta \downarrow 0} \Gamma(h_1, y^*_\tau - \delta) \) is optimal.

In practice the generative model is not known, but the population-level algorithm suggests an estimator of \( \pi_{1,\tau}^{Q IQ} \). Assuming parametric models \( m(H_2; \beta_{2,0}) \) and \( c(H_2; \beta_{2,1}) \) for \( m(H_2) \)
and \(c(H_2)\), the following QIQ-learning algorithm can be used to estimate an optimal first-stage decision rule. The exact algorithm depends on the choice of estimators for \(F(\cdot)\) and \(G(\cdot, \cdot | h_1, a_1)\); several options are presented in Sections 2.3 and 2.4 but the choice should be data-driven.

**The QIQ-learning algorithm:**

QIQ.1 Follow TIQ.1 – TIQ.3 of the TIQ-learning algorithm in Section 2.1.

QIQ.2 With \(I(\cdot, \cdot, \cdot)\) as in (3) and first-stage patient histories \(h_{1i}\), estimate \(y^*_\tau\) using

\[
\hat{y}^*_\tau \triangleq \inf \left( y : \frac{1}{n} \sum_{i=1}^{n} I \left[ y, \hat{F}(\cdot), \hat{G}(\cdot, \cdot | h_{1i}, \hat{\Gamma}(h_{1i}, y)) \right] \geq \tau \right).
\]

QIQ.3 Estimate \(f(y^*_\tau)\) using

\[
\hat{f}(\hat{y}^*_\tau) \triangleq \inf \left( y : \frac{1}{n} \sum_{i=1}^{n} I \left[ y, \hat{F}(\cdot), \hat{G}(\cdot, \cdot | h_{1i}, \hat{\Gamma}(h_{1i}, \hat{y}^*_\tau)) \right] \geq \tau \right).
\]

QIQ.4

a. If \(\hat{f}(\hat{y}^*_\tau) = \hat{y}^*_\tau\), then \(\hat{\pi}^{QIQ}_{1,\tau}(h_1) = \hat{\Gamma}(h_1, \hat{y}^*_\tau)\) is an estimated optimal first-stage decision rule because it attains the estimated optimal quantile, \(\hat{y}^*_\tau\).

b. If \(\hat{f}(\hat{y}^*_\tau) < \hat{y}^*_\tau\), then the first-stage rule \(\hat{\pi}_1(h_1) = \hat{\Gamma}(h_1, \hat{y}^*_\tau - \delta), \delta > 0\), results in the estimated quantile \(\hat{f}(\hat{y}^*_\tau - \delta)\), which satisfies \(\hat{y}^*_\tau - \delta < \hat{f}(\hat{y}^*_\tau - \delta) \leq \hat{y}^*_\tau\). By choosing \(\delta\) arbitrarily small, this estimated quantile will be arbitrarily close to the estimated optimal quantile \(\hat{y}^*_\tau\).

To complete the TIQ- and QIQ-learning algorithms, we provide specific estimators \(F(\cdot)\) and \(G(\cdot, \cdot | h_1, a_1)\) in the next two sections. We suggest estimators that are likely to be useful in practice, but our list is not exhaustive. An advantage of TIQ- and QIQ-learning is that they involve modeling only smooth transformations of the data; these are standard, well-studied modeling problems in the statistics literature.

2.3 Working models for \(F(\cdot)\)

Both TIQ- and QIQ-learning require estimation of the distribution function of the second-stage error, \(\epsilon\). We suggest two estimators; the choice between them can be guided by inspection of the residuals from the second-stage regression.

**Normal Scale Model.**

The normal scale model for \(F(\cdot)\) is given by \(\hat{F}_N^*(z) \triangleq \Phi(z/\hat{\sigma}_\epsilon)\), where \(\hat{\sigma}_\epsilon\) is the estimated standard deviation of the second-stage residuals, \(\epsilon_i^Y \triangleq Y_i - m(H_{2i}; \hat{\beta}_{2,0}) - A_{2i}c(H_{2i}; \hat{\beta}_{2,1})\), where \(H_{2i}\) and \(A_{2i}\) denote the second-stage histories and treatments for patients \(i = 1, \ldots, n\). We assume constant variance across second-stage histories and treatment. However, variance modeling techniques (Carroll and Ruppert 1988) can be used to account for heteroskedastic variance for greater flexibility when necessary.

**Nonparametric Model.**
distribution of their standardized residuals. Define these standardized residuals as conditional mean and variance functions of 

\[ F_m(z) \]

In addition to modeling \( F_m(z) \), TIQ- and QIQ-learning require modeling the bivariate conditional density of \( m(H_2) \) and \( c(H_2) \) given \( H_1 \) and \( A_1 \). A useful strategy is to first model the conditional mean and variance functions of \( m(H_2) \) and \( c(H_2) \) and then estimate the joint distribution of their standardized residuals. Define these standardized residuals as

\[
e^m = \frac{m(H_2) - \mu_m(H_1, A_1)}{\sigma_m(H_1, A_1)}, \quad e^c = \frac{c(H_2) - \mu_c(H_1, A_1)}{\sigma_c(H_1, A_1)},
\]

where \( \mu_m(H_1, A_1) \triangleq E\{m(H_2) \mid H_1, A_1\} \) and \( \sigma_m^2(H_1, A_1) \triangleq E[(m(H_2) - \mu_m(H_1, A_1))^2 \mid H_1, A_1] \). The mean and variance functions of \( c(H_2) \) are defined similarly: \( \mu_c(H_1, A_1) \triangleq E\{c(H_2) \mid H_1, A_1\} \) and \( \sigma_c^2(H_1, A_1) \triangleq E[(c(H_2) - \mu_c(H_1, A_1))^2 \mid H_1, A_1] \). In simulations, we use parametric mean and variance models for \( \mu_m, \sigma_m^2, \mu_c, \) and \( \sigma_c^2 \), and we estimate the joint distribution of \( e^m \) and \( e^c \) using a Gaussian copula. Alternatively, the joint residual distribution could be modelled parametrically, for example, with a multivariate normal model; or nonparametrically, e.g., using a bivariate kernel density estimator (Silverman 1986, Ch. 4). Common exploratory analysis techniques can be used to interactively guide the choice of estimator for \( G(\cdot, \cdot \mid h_1, a_1) \). Using mean and variance modeling, the following steps would be substituted in Step TIQ.3 of the TIQ-learning algorithm.

### Mean and Variance Modeling.

1. Compute \( \hat{\theta}_m \triangleq \arg\min_{\theta_m} \sum_{i=1}^n \left\{ m(H_{2i}; \beta_{2,0}) - \mu_m(H_{1i}, A_{1i}; \theta_m) \right\}^2 \) and the resulting estimator \( \mu_m(H_1, A_1; \hat{\theta}_m) \) of the mean function \( \mu_m(H_1, A_1) \).

2. Use the estimated mean function from Step 3.1 to obtain

\[
\hat{\gamma}_m \triangleq \arg\min_{\gamma_m} \sum_{i=1}^n \left\{ \left( m(H_{2i}; \beta_{2,0}) - \mu_m(H_{1i}, A_{1i}; \hat{\theta}_m) \right)^2 - \sigma_m^2(H_{1i}, A_{1i}; \gamma_m) \right\}^2,
\]

so that \( \sigma_m^2(H_1, A_1; \hat{\gamma}_m) \) is an estimator of the variance function \( \sigma_m^2(H_1, A_1) \). In practice, we may specify a log-linear model for the natural log of the squared residuals from the mean fit in Step 3.1. Alternatively, a common variance across histories \( H_1 \) and treatment \( A_1 \) or within each level of \( A_1 \) could be estimated.

3. Repeat Steps 3.1 and 3.2 to obtain estimators of \( \mu_c(H_1, A_1; \hat{\theta}_c) \) and \( \sigma_c(H_1, A_1; \hat{\gamma}_c) \).

4. Plug in the estimated parametric mean and variance functions to obtain \( \hat{e}^m_i \) and \( \hat{e}^c_i \), for \( i = 1, \ldots, n \), defined as

\[
\hat{e}^m_i = \frac{m(H_{2i}; \beta_{2,0}) - \mu_m(H_{1i}, A_{1i}; \hat{\theta}_m)}{\sigma_m(H_{1i}, A_{1i}; \hat{\gamma}_m)}, \quad \hat{e}^c_i = \frac{c(H_{2i}; \beta_{2,0}) - \mu_c(H_{1i}, A_{1i}; \hat{\theta}_c)}{\sigma_c(H_{1i}, A_{1i}; \hat{\gamma}_c)}.
\]
Then, $\hat{\varepsilon}^m_i$ and $\hat{\varepsilon}_i^e$, $i = 1, ..., n$, can be used to estimate the joint distribution of the standardized residuals. Samples drawn from this distribution can be transformed back to samples from $\hat{G}(\cdot, | h_1, a_1)$ to estimate the integral $I \{ y, \hat{F}_e(\cdot), \hat{G}(\cdot, | h_1, a_1) \}$ with a Monte Carlo average.

2.5 Theoretical results

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 $\hat{m}(h_2)$ and $\hat{c}(h_2)$ that converge in probability to $m(h_2)$ and $c(h_2)$, respectively, for each $h_2$.

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

A3. For each fixed $h_1$ and $a_1$, $\int |d\hat{G}(u, v | h_1, a_1) - dG(u, v | h_1, a_1)|$ converges to zero in probability.

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

A5. $\Pr(|d(H_1, y^*_\tau)| = 0) = 0.$

In the simulation studies in Section 3 and data example in Section 4, we use linear working models for $m(\cdot)$ and $c(\cdot)$ which are estimated using least squares. Thus, A1 is satisfied under usual regularity conditions. When $\epsilon$ is continuous, assumption A2 can be satisfied by specifying $\hat{F}_\epsilon(\cdot)$ as the empirical distribution function. If for each fixed $h_1$ and $a_1$, $dG(\cdot, | h_1, a_1)$ is a density and $d\hat{G}(\cdot, | h_1, a_1)$ a pointwise consistent estimator, then A3 is satisfied (Glick, 1974). Assumption A5 states that all patients have a non-zero first-stage treatment effect at $y^*_\tau$. Theorem 12.3 is proved in the supplementary material.

Theorem 2.3. (Consistency of TIQ-learning) Assume A1–A3 and fix $\lambda \in \mathbb{R}$. Then, $\Pr(\hat{\pi}^{TIQ}_\lambda(Y > \lambda))$ converges in probability to $\Pr(\pi^{TIQ}_\lambda(Y > \lambda))$, where $\hat{\pi}^{TIQ}_\lambda = (\hat{\pi}^{TIQ}_{\lambda,1}, \hat{\pi}^{TIQ}_{\lambda,2}).$

Theorem 2.4. (Consistency of QIQ-learning) Assume A1–A5. Then, $\Pr(\hat{\pi}^{QIQ}_\tau(Y) > \lambda)$ converges in probability to $g^*_\tau$ for any fixed $\tau$, where $\hat{\pi}^{QIQ}_\tau = (\hat{\Gamma}(\cdot, y^*_\tau), \hat{\pi}^{QIQ}_{\tau,2}).$

3. MONTE CARLO RESULTS

We compare the performance of our estimators to the mean-optimal methods Q-learning and IQ-learning (Laber et al., 2014) for a range of data generative models. Gains are achieved in terms of the proportion of the distribution of $Y$ that exceeds the constant threshold $\lambda$ and the $\tau$th quantile for several values of $\lambda$ and $\tau$. The data is generated using the model

\[
X_1 \sim \text{Normal}_{p=2}(1_2, \Sigma), \quad A_t \in \{-1, 1\}, \quad t = 1, 2, \\
X_2 = B_{A_1}X_1 + \eta_{H_{1, A_1}}\xi, \quad H_1 = (1, X_1^\top)^\top, \\
\eta_{H_{1, A_1}} = \exp\{C(H_1^\top \gamma_0 + A_1 H_1^\top \gamma_1)/2\}, \quad \xi \sim \text{Normal}_{p=2}(0_2, I_{2 \times 2}), \\
C \in [0, 1], \quad H_2 = (1, X_2^\top)^\top, \\
Y = H_{2, 0}^\top \beta_{2, 0} + A_2 H_{2, 1}^\top \beta_{2, 1} + \epsilon, \quad \epsilon \sim \text{Normal}(0, 1),
\]
where $1_p$ is a $p \times 1$ vector of 1s. The matrix $\Sigma$ is a correlation matrix with off-diagonal $\rho = 0.5$. The $2 \times 2$ matrix $B_{A_1}$ takes values

$$B_{A_1=1} = \begin{pmatrix} -0.1 & -0.1 \\ 0.1 & 0.1 \end{pmatrix}, \quad B_{A_1=-1} = \begin{pmatrix} 0.5 & -0.1 \\ -0.1 & 0.5 \end{pmatrix}.$$ 

The remaining parameters are $\gamma_0 = (1, 0.5, 0)^\top$, $\gamma_1 = (-1, -0.5, 0)^\top$, $\beta_{2,0} = (0.25, -1, 0.5)^\top$, and $\beta_{2,1} = (1, -0.5, -0.25)^\top$, which were chosen to ensure that the mean-optimal treatment produced a more variable response for some patients.

### 3.1 TIQ-learning Simulation Results

Results are based on $J = 1,000$ generated data sets; for each we estimate the TIQ-, IQ-, and $Q$-learning policies using a training set of size $n = 250$ and compare the results using a test set of size $N = 10,000$. The normal scale model is used to estimate $F\phi(\cdot)$, which is correctly specified for the generative model above. The Gaussian copula model discussed in Section 2.4 is also correctly specified and is used as the estimator for $G(\cdot, \cdot \mid h_1, a_1)$.

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 $\text{pr}^{\hat{\pi}_j}(Y > \lambda)$ to be the true probability that $Y$ exceeds $\lambda$ given treatments are assigned according to $\hat{\pi}_j = (\hat{\pi}_{1j}, \hat{\pi}_{2j})$, the regime estimated from the $j$th generated data set. For threshold values $\lambda = 2, 2.4$, we estimate $\text{pr}^{\pi}(Y > \lambda)$ using $\sum_{j=1}^J \text{pr}^{\hat{\pi}_j}(Y > \lambda)/J$, where $\text{pr}^{\hat{\pi}_j}(Y > \lambda)$ is an estimate of $\text{pr}^{\hat{\pi}_j}(Y > \lambda)$ obtained by calculating the proportion of test patients consistent with regime $\hat{\pi}_j$ whose observed $Y$ values are greater than $\lambda$. Thus, our estimate is an average over training data sets and test set observations. In terms of the proportion of distribution mass above $\lambda$, results for $\lambda = -2$ and 4 in Figure 1 show a clear advantage of TIQ-learning for higher values of $C$, the degree of heteroskedasticity in the second-stage covariates $X_2$. As anticipated by Remark 1 in Section 2.1, all methods perform similarly when $\lambda = 2$.

Figure 2 illustrates how the optimal first-stage treatment for a test set of 1,000 individuals
changes as $\lambda$ varies. Results are shown for $C = 0.5$. The true optimal treatments displayed in the left plot show a distinct shift from treating most of the population with $A_1 = 1$ to $A_1 = -1$ as $\lambda$ increases from -4 to 4. The TIQ-learning estimated optimal treatments displayed in the middle plot are averaged over 100 Monte Carlo iterations and closely resemble the true policies on the left. Although the estimated $Q$-learning regime does not depend on $\lambda$, it is plotted for each $\lambda$ value to aid visual comparison. The first-stage treatments recommended by $Q$-learning differ the most from the true optimal treatments when $\lambda = 4$, corroborating the results for $C = 0.5$ in Figure 1.

3.2 QIQ-learning Simulations

To study the performance of the QIQ-learning algorithm, we compare quantiles of $Y$ when the population is treated according to the regimes estimated by QIQ-learning, IQ-learning, and $Q$-learning. A smaller test set of size $N = 5,000$ was used in this section to reduce computation time. Define $q_{\hat{\pi}_j}(\tau)$ to be the true $\tau$th quantile of the distribution of $Y$ given treatments are assigned according to $\hat{\pi}_j = (\hat{\pi}_{1j}, \hat{\pi}_{2j})$, the regime estimated from the $j$th generated data set. For $\tau = 0.1, 0.5, 0.75$, we estimate $q_{\pi}(\tau)$ using $\sum_{j=1}^J q_{\hat{\pi}_j}(\tau)/J$, where $q_{\hat{\pi}_j}(\tau)$ is an estimate of $q_{\hat{\pi}_j}(\tau)$ obtained by calculating the $\tau$th quantile of the subgroup of test patients consistent with regime $\hat{\pi}_j$. The generative model and all other parameter settings used here are the same as those in the previous section. For our generative model, the condition of Lemma 2.2 is satisfied, so the true optimal regime is attained asymptotically. The results in Figure 3 indicate that the lowest quantile, $\tau = 0.1$, suffers under the $Q$-learning regime as heterogeneity in the second-stage histories increases, represented by the scaling constant $C$. In contrast, quantiles of the QIQ-learning estimated regimes for $\tau = 0.1$ remain constant across the entire range of $C$. When $\tau = 0.5$, all methods perform similarly; for some $C$, IQ- and $Q$-learning outperform QIQ-learning. This is not surprising because all models used to generate the data were symmetric. Thus, maximizing the mean of $Y$ gives similar results to maximizing the median. We conjecture that in this scenario for $\tau = 0.5$
Figure 3: Left to Right: $\tau = 0.1, 0.5, 0.75$. Solid black, true optimal quantiles; dotten black, quantiles under randomization; dashed with circles/squares/triangles, quantiles under QIQ-, Q-, and IQ-learning, respectively.

the results for the three methods would converge as the sample size is increased.

4. STAR*D ANALYSIS

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Fava et al., 2003; Rush et al., 2004) is a four-stage Sequential Multiple Assignment Randomized Trial (Lavori and Dawson, 2004; Murphy, 2005a) studying personalized treatment strategies for patients with major depressive disorder. Depression is measured by the Quick Inventory of Depressive Symptomatology (QIDS) score, a one-number summary score that takes integer values 0 to 27. Lower scores indicate fewer depression symptoms. Remission is defined as QIDS $\leq 5$. Previous attempts to estimate optimal dynamic treatment regimes from this data have used the criteria, “maximize end-of-stage-two QIDS,” (see, for example, Schulte et al., 2014; Laber et al., 2014) a surrogate for the primary aim of helping patients achieve remission. We illustrate TIQ-learning by estimating an optimal regime that maximizes the probability of remission for each patient, directly corresponding to the primary clinical goal.

The first stage, which we will henceforth refer to as baseline, was non-randomized with each patient receiving Citalopram, a drug in the class of Selective Serotonin Reuptake Inhibitors (SSRIs). We use a subset of the STAR*D data from the first two randomized stages, and refer to the original trial levels 2 and 3 as “stage one” and “stage two.” Before each randomization, patients specified a preference to “switch” or “augment” their current treatment strategy and were then randomized to one of multiple options within their preferred category. In addition, patients who achieved remission in any stage exited the study. To keep our illustration of TIQ-learning concise, we restrict our attention to patients who participated in stages one and two who preferred the “switch” strategy at both stages. At stage one, our binary treatment variable is “SSRI,” which includes only Sertraline, versus “non-SSRI,” which includes both Bupropion and Venlafaxine. At stage two we compare Mirtazapine and Nortriptyline which are both non-SSRIs. In the patient subgroup considered in our analysis, treatments were randomized at both stages.

All measured QIDS scores are recoded as $27 - \text{QIDS}$ so that higher scores correspond to
fewer depression symptoms. After recoding, remission corresponds to QIDS > 21. Thus, TIQ-learning with \( \lambda = 21 \) maximizes the probability of remission for all patients. In general, QIDS was recorded during clinic visits at weeks 2, 4, 6, 9, and 12 in each stage, although some patients with inadequate response moved on to the next stage before completing all visits. We summarize longitudinal QIDS trajectories from the baseline stage and stage one by averaging over the total number of QIDS observations in the given stage. Variables used in our analysis are listed in Table 1. We describe all models used in the analysis below.

| Variable | Description |
|----------|-------------|
| qids0    | mean QIDS during the baseline stage. |
| slope0   | pre-randomization QIDS improvement; the difference between the final and initial baseline-stage QIDS scores, divided by time spent in the baseline stage. |
| qids1    | mean stage-one QIDS. |
| slope1   | first-stage QIDS improvement; the difference between the final and initial first-stage QIDS scores, divided by time spent in the first randomized stage. |
| A1       | First-stage treatment; 1=“SSRI” and -1=“non-SSRI.” |
| A2       | Second-stage treatment; 1=“NTP” for Nortriptyline and -1=“MIRT” for Mirtazapine. |
| Y        | 27 minus final QIDS score, measured at the end of stage two. |

At the second stage, we assume the linear working model \( Y = H_{2,0}^\top \beta_{2,0} + A_2 H_{2,1}^\top \beta_{2,1} + \epsilon \), where \( H_{2,0} = H_{2,1} = (1, \text{qids1}, \text{slope1}, \text{A1})^\top \), \( E(\epsilon) = 0 \), \( \text{var}(\epsilon) = \sigma^2 \), and \( \epsilon \) is independent of \( H_2 \) and \( A_2 \). We fit this model using least squares. A normal qq-plot of the residuals from the previous regression step indicates slight deviation from normality, so we use the empirical estimator of \( F_i(\cdot) \) given in Section 2.4. Next, we estimate the conditional mean and variance functions of \( m(H_2) \triangleq H_{2,0}^\top \beta_{2,0} \) and \( c(H_2) \triangleq H_{2,1}^\top \beta_{2,1} \) following steps described in Section 2.4. For the mean functions, we take \( H_{1,0} = H_{1,1} = (1, X_1^\top) \) with \( X_1 = (\text{qids0}, \text{slope0})^\top \) and use working models of the form \( E(k(H_2) \mid X_1, A_1) = H_{1,0}^\top \beta_{1,0} + A_1 H_{1,1}^\top \beta_{1,1} \). Exploratory analyses reveal little evidence of heteroskedasticity at the first-stage. Thus, we opt to estimate a constant residual variance for both terms following the mean modeling steps. After the mean and variance modeling steps, we use a Gaussian copula to estimate the joint conditional distribution of the standardized residuals of \( \{m(H_2), c(H_2)\} \) given \( H_1 \) and \( A_1 \), resulting in our estimate of \( G(\cdot, \cdot \mid h_1, a_1) \) which we denote by \( \hat{G}(\cdot, \cdot \mid h_1, a_1) \).

The estimated first-stage optimal rule is \( \hat{\pi}_{1,\text{TIQ}}(h_1) = \arg \min \int \hat{F}_i(21 - u - |v|) d\hat{G}(u, v \mid h_1, a_1) \). At stage two, \( \hat{\pi}_2(h_2) = \text{sgn}(-1.66 + 0.15 \times \text{qids1} - 4.03 \times \text{slope1} - 0.68 \times \text{A1}) \) is the estimated optimal treatment. Based on Remark 1 in Section 2.1, we compare the estimated first-stage treatment recommendations to those recommended by the mean-optimal rule, \( \arg \min_{a_1} \int (-u - |v|) d\hat{G}(u, v \mid h_1, a_1) \), for each observed \( h_1 \) in the data. Only one patient out of 132 is recommended differently. In addition, the difference in raw values of \( \int \hat{F}_i(21 -
$u - |v|)d\hat{G}(u, v \mid \mathbf{h}_1, a_1)$ for $a_1 = 1, -1$ as well as $\int(-u - |v|)d\hat{G}(u, v \mid \mathbf{h}_1, a_1)$ for $a_1 = 1, -1$ are the smallest for this particular patient. Thus, the treatment discrepancy is most likely due to a near-zero treatment effect for this patient.

Table 2: Estimated value of dynamic and non-dynamic regimes using the Adaptive Inverse Probability Weighted Estimator.

| Estimated Value |
|-----------------|
| TIQ-learning    | 0.24 |
| Q-learning      | 0.23 |
| (1, 1)          | 0.13 |
| (-1, 1)         | 0.24 |
| (1, -1)         | 0.07 |
| (-1, -1)        | 0.12 |

Comparing the results to Q-learning, which maximizes the expected value of $Y$, supports the claim that TIQ-learning and mean optimization are equivalent for this subset of the STAR*D data. The first step of Q-learning is to model the conditional expectation of $Y$ given $\mathbf{H}_2$ and $\mathbf{A}_2$ which is the same as the first step of TIQ-learning. Thus, we use the same model and estimated decision rule at stage two given in Step 1 of the TIQ-learning algorithm. Next, we model the conditional expectation of $\tilde{Y} = \mathbf{H}_2^\top_0 \beta_{2,0} + |\mathbf{H}_2^\top_1 \beta_{2,1}|$, where $\tilde{Y}$ is the predicted future optimal outcome at stage one. We specify the working model $E(\tilde{Y} \mid \mathbf{H}_1, \mathbf{A}_1) = \mathbf{H}_1^\top_0 \beta_{1,0} + \mathbf{A}_1 \mathbf{H}_1^\top_1 \beta_{1,1}$, where $\mathbf{H}_1^\top_0 = (1, \mathbf{X}_1^\top_0)$ and $\mathbf{X}_1 = (\text{qids0, slope0})^\top$. We fit the model using least squares. Then, the Q-learning estimated optimal first-stage rule is $\hat{\pi}_{1,\lambda}(\mathbf{h}_1) = \text{sgn}(-0.95 + 0.13 * \text{qids1} + 2.17 * \text{slope1})$. Q-learning recommends treatment differently at the first stage for only one of the 132 patients in the data. In addition, the estimated value of the TIQ- and Q-learning regimes are nearly the same and are displayed in Table 2. Also included in Table 2 are value estimates for four non-dynamic regimes that treat everyone according to the decision rules $\pi_1(\mathbf{h}_1) = a_1$ and $\pi_2(\mathbf{h}_2) = a_2$ for $a_1 \in \{-1, 1\}$ and $a_2 \in \{-1, 1\}$. We estimate the value using the Augmented Inverse Probability Weighted Estimator given in [Zhang et al.] (2013).

In summary, it appears that TIQ-learning and Q-learning perform similarly for this subset of the STAR*D data. This may be due to the lack of heteroskedasticity at the first stage. Thus, maximizing the end-of-stage-two QIDS using mean-optimal techniques seems appropriate and, in practice, equivalent to maximizing remission probabilities for each patient with TIQ-learning.

5. DISCUSSION

We have proposed modeling frameworks for estimating optimal dynamic treatment regimes in settings where a non-mean distributional summary is the intended outcome to optimize. Threshold Interactive Q-learning estimates a regime that maximizes the mass of the response
distribution that exceeds a constant or patient-dependent threshold, and Quantile Interactive Q-learning maximizes a prespecified quantile of the response distribution. An important application of TIQ-learning is to estimate a regime that maximizes the probability of achieving remission for each patient, where remission is defined in terms of an indicator of threshold exceedance.

Although our proposed methods are only suitable for the two-stage setting, this is an important development given that many completed and ongoing SMART studies have this structure [Projects Using SMART, 2012; Laber, 2013]. Here we considered binary treatments at both stages. In principle, the proposed methods may be extended to settings with more than two available treatments at each stage by modeling additional treatment contrasts. Formalization of this idea merits further research.

6. SUPPLEMENTARY MATERIALS
Supplementary materials follow the references section below and include discussions of modeling adjustments for heteroskedastic second-stage errors and patient-specific thresholds; a proof of Lemma 2.2 and toy example illustrating a setting where this lemma does not apply; and proofs of the theorems in Section 2.5.

REFERENCES
Berkowitz, R. I., Wadden, T. A., Gehrman, C. A., Bishop-Gilyard, C. T., Moore, R. H., Womble, L. G., Cronquist, J. L., Trumpikas, N. L., Katz, L. E. L., and Xanthopoulos, M. S. (2010). Meal Replacements in the Treatment of Adolescent Obesity: A Randomized Controlled Trial. *Obesity*, 19(6):1193–1199.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chakraborty, B., Murphy, S. A., and Strecher, V. J. (2010). Inference for Non-Regular Parameters in Optimal Dynamic Treatment Regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S. R., Lavori, P. W., Rosenbaum, J. F., Kupfer, D. J., and STAR*D Investigators Group (2003). Background and Rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study. *Psychiatric Clinics of North America*, 26(2):457 – 94.

Glick, N. (1974). Consistency Conditions for Probability Estimators and Integrals of Density Estimators. *Utilitas Mathematica*, 6:61–74.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with Censored Data. *Annals of statistics*, 40(1):529.

Goldberg, Y., Song, R., and Kosorok, M. R. (2013). Adaptive Q-learning. *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 150 – 62.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for Optimal Dynamic Treatment Regimes. *Biometrics*, 66(4):1192–1201.
Laber, E. B. (2013). Example SMARTs. http://www4.stat.ncsu.edu/ laber/smart.html.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014). Interactive Q-learning. *Biometrika*, accepted.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2010). Statistical Inference in Dynamic Treatment Regimes. *arXiv:1006.5831 [stat.ME]*.

Lavori, P. W. and Dawson, R. (2000). A Design for Testing Clinical Strategies: Biased Adaptive Within-Subject Randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38.

Lavori, P. W. and Dawson, R. (2004). Dynamic Treatment Regimes: Practical Design Considerations. *Clinical Trials*, 1(1):9–20.

McKeague, I. W. and Qian, M. (2013). Evaluation of Treatment Policies Based on Functional Predictors. *Statistica Sinica*, submitted.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for Estimating Optimal Dynamic Treatment Rules From Observational Data. *Canadian Journal of Statistics*, 40(4):629–645.

Moodie, E. E. M. and Richardson, T. S. (2010). Estimating Optimal Dynamic Regimes: Correcting Bias Under the Null. *Scandinavian Journal of Statistics*, 37(1):126 – 146.

Murphy, S. A. (2003). Optimal Dynamic Treatment Regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005a). A Generalization Error for Q-learning. *Journal of Machine Learning Research*, 6(7):1073 – 1097.

Murphy, S. A. (2005b). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*, 24(10):1455–1481.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content. *The International Journal of Biostatistics*, 6(2).

Projects Using SMART (2012). The Methodology Center at Pennsylvania State University. http://methodology.psu.edu/ra/adap-inter/projects.

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.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688.
7. THRESHOLD INTERACTIVE Q-LEARNING WITH SECOND-STAGE HETEROSKEDASTICITY

Here we assume

\[ Y = m(H_2) + A_2 c(H_2) + \eta(H_2, A_2) \epsilon, \]  

(16)

where we define \( \eta(H_2, A_2) = \exp\{r(H_2) + A_2 s(H_2)\} \) for functions \( r \) and \( s \). In addition, \( E(\epsilon) = 0, \text{var}(\epsilon) = 1 \), and \( \epsilon \) is independent of \( H_2 \) and \( A_2 \). Thus, the conditional variance
of $Y$ given $H_2$ and $A_2$ is log-linear. Under model [16], the $\lambda$-optimal second-stage decision rule for a patient presenting with $h_2$ is

$$
\pi_{2,\lambda}^{\text{TIQ}}(h_2) = \text{sgn} \left[ \frac{\lambda - m(h_2) + c(h_2)}{\exp \{ r(h_2) - s(h_2) \} - \exp \{ r(h_2) + s(h_2) \} } \right].
$$

To see this, define

$$
\text{pr}^{\pi_1,\pi_2}(Y > \lambda) = E[E\{\text{pr}^{\pi_1,\pi_2}(Y > \lambda \mid H_2, a_2) \mid a_2 = \pi_2(H_2) \mid H_1, a_1 \mid a_1 = \pi_1(H_1) \}]
\quad = \quad E \left\{ E \left[ \text{pr} \left[ \epsilon > \frac{\lambda - m(H_2) - \pi_2(H_2)c(H_2)}{\exp \{ r(H_2) + \pi_2(H_2)s(H_2) \} } \mid H_1, \pi_1(H_1) \right] \right] \right\}.
$$

To maximize the previous expression, choose $\pi_2(h_2) \in \{-1, 1\}$ to minimize

$$
\frac{\lambda - m(h_2) - \pi_2(h_2)c(h_2)}{\exp \{ r(h_2) + \pi_2(h_2)s(h_2) \}},
$$

leading to $\pi_{2,\lambda}^{\text{TIQ}}(h_2)$ in [17]. Define $G(\cdot, \cdot, \cdot \mid h_1, a_1)$ to be the joint conditional distribution of $\{m(h_2), c(h_2), r(h_2), s(h_2)\}$ given $H_1 = h_1$ and $A_1 = a_1$. Let $F_\epsilon(\cdot)$ denote the cumulative distribution function of $\epsilon$. The first-stage $\lambda$-optimal decision rule is

$$
\pi_{1,\lambda}^{\text{TIQ}}(h_1) = \arg \min_{\alpha_1} \int F_\epsilon \left( \frac{\lambda - t + \text{sgn}\{K(t, u, v, w)\}u}{\exp \{ v + \text{sgn}\{K(t, u, v, w)\}w \} } \right) G(t, u, v, w \mid h_1, a_1) \, dt \, du \, dv \, dw,
$$

where

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

Thus, estimation of $\pi_{1,\lambda}^{\text{TIQ}}$ involves specifying estimators for $F_\epsilon(\cdot)$ and the four-dimensional conditional density $G(\cdot, \cdot, \cdot \mid h_1, a_1)$. 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.

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

Denote the optimal second-stage rule for patient-specific threshold $\lambda(h_t)$ by $\pi_{2,\lambda(h_t)}^{\text{TIQ}}(h_2)$, where $t = 1$ or $t = 2$, depending on the scientific interest and trial design. Then, $\pi_{2,\lambda(h_t)}^{\text{TIQ}}(h_2) = \pi_{2}^{*}(h_2) = \text{sgn}\{c(h_2)\}$ whether $t = 1$ or $2$. To see this, note for fixed $\pi_1$,

$$
\text{pr}^{\pi_1,\pi_2}(Y > \lambda(H_t)) = E(E[\text{pr}^{\pi_1,\pi_2}(Y > \lambda(H_t) \mid H_2, a_2) \mid a_2 = \pi_2(H_2) \mid H_1, a_1] \mid a_1 = \pi_1(H_1)).
$$

Because $H_1 \subset H_2$, conditioning on $H_2$ reduces $\lambda(H_t)$ 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 $\pi_{2,\lambda(h_t)}^{\text{TIQ}}(h_2) = \pi_{2}^{*}(h_2) = \text{sgn}\{c(h_2)\}$.

When the threshold depends on the first-stage history, $\lambda(h_1)$ replaces $\lambda$ 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 \( \{\lambda(H_2), m(H_2), c(H_2)\} \) given \( H_1 = h_1 \) and \( A_1 = a_1 \) must be estimated. Let \( G(\cdot, \cdot, \cdot | h_1, a_1) \) denote this trivariate distribution and \( \hat{G}(\cdot, \cdot, \cdot | h_1, a_1) \) an estimator. In this case, the estimated optimal first-stage decision rule is

\[
\pi_{\hat{G}(\cdot, \cdot, \cdot | h_1, a_1)}^{\text{QIO}}(h_1) = \arg\min_{a_1} \int \hat{F}_t (t-u-|v|) \hat{G}(t,u,v \mid h_1,a_1) dt du dv.
\]

Thus, the first-stage optimal treatment is based on the average of all possible future patient-specific thresholds, \( \lambda(H_2) \), given the observed first-stage history, \( h_1 \).

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

We show the \( \tau \)-optimal QIQ-learning second-stage rule is \( \pi_{\tau,\text{QIO}}(h_2) = \text{sgn}\{c(h_2)\} \) under the assumption of constant variance at the second-stage. Define the set \( S_{\pi^1, \pi^2} \triangleq \{ y : \text{pr}^{\pi^1, \pi^2}(Y \leq y) \geq \tau \} \), so that \( q_{\pi^1, \pi^2}(\tau) = \inf S_{\pi^1, \pi^2} \). In Section 2.1 of the main paper, we showed \( \text{pr}^{\pi^1, \pi^2}(Y \leq y) \geq \text{pr}^{\pi^1, \pi^2}(Y \leq y) \) for arbitrary \( y \), and hence for all fixed \( y \), where we define \( \pi^*_2(h_2) = \text{sgn}\{c(h_2)\} \). It follows that \( S_{\pi^1, \pi^2} \subset S_{\pi^1, \pi^2} \text{.} \) 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, \( \pi_{\tau,\text{QIO}}(h_2) = \pi^*_2(h_2) = \text{sgn}\{c(h_2)\} \) is optimal because this inequality holds for arbitrary \( \pi_1 \) and \( \pi_2 \).

10. PROOF OF LEMMA 2.2 IN SECTION 2.2

Lemma 2.2 from Section 2.2 of the main paper is restated below.

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

(B) \( f(y^*_\tau - \delta) = \lim_{\delta \downarrow 0} f(y^*_\tau - \delta) = y^*_\tau \);

(C) \( f(y^*_\tau) \leq y^*_\tau \) with strict inequality if there exists \( \delta > 0 \) such that

\[
\text{pr}\Gamma(\cdot, y^*_\tau), \pi_2 (Y \leq y^*_\tau - \delta) > \tau;
\]

(D) If \( F_t(\cdot) \) is continuous and strictly increasing, then \( f(y^*_\tau) = y^*_\tau \).

Proof. We showed in expression (11) of Section 2.2 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 \( y_0 \geq f(y_0) \). Using (11) of Section 2.2 and the assumption that \( y_0 \geq f(y_0) \), it follows that

\[
\tau \leq \text{pr}\Gamma(\cdot, y_0), \pi^*_2 \{ Y \leq f(y_0) \} \leq \text{pr}\Gamma(\cdot, y_0), \pi^*_2 \{ Y \leq y_0 \}
\]

because for the fixed regime \( \pi = \{\Gamma(\cdot, y_0), \pi^*_2\} \), \( \text{pr}\Gamma(\cdot, y_0), \pi^*_2 \{ Y \leq y \} \) is a distribution function and nondecreasing in \( y \). However, we have a contradiction because by (8) in Section 2.2, \( y^*_\tau \) is the smallest \( y \) satisfying \( \text{pr}\Gamma(\cdot, y), \pi^*_2 (Y \leq y) > \tau \).

Using (12) in Section 2.2 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).

The second inequality in (11) from Section 2.2 shows that \( f(y^*_\tau) \leq y^*_\tau \) and thus in light of (B) the inequality is strict when \( f(y) \) is not left continuous at \( y^*_\tau \). If there exists \( \delta > 0 \) such that \( \text{pr}\Gamma(\cdot, y^*_\tau), \pi^*_2 (Y \leq y^*_\tau - \delta) > \tau \), then because \( f(y^*_\tau) \) is the smallest \( \tilde{y} \) for which \( \text{pr}\Gamma(\cdot, y^*_\tau), \pi^*_2 (Y \leq \tilde{y}) \geq \tau \) it must be that \( f(y^*_\tau) \leq y^*_\tau - \delta < y^*_\tau \), proving (C).
When $F_\epsilon(\cdot)$ is continuous and strictly increasing, $\Pr^{\Gamma(\cdot, y^*_\tau), \pi^*_2(Y \leq y)}$ 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 $\pi = (\pi_1, \pi_2)$, $\Pr^{\pi_1, \pi_2}(Y \leq y)$ continuous in $y$ implies $\Pr^{\Gamma(\cdot, y), \pi^*_2(Y \leq y)}$ is also continuous. Suppose toward a contradiction that $f(y^*_\tau) < y^*_\tau$. When $\Pr^{\Gamma(\cdot, y^*_\tau), \pi^*_2(Y \leq y)}$ is continuous and strictly increasing, the Mean Value Theorem guarantees existence of exactly one point $\tilde{y} \in \mathbb{R}$ such that $\Pr^{\Gamma(\cdot, y^*_\tau)}(\{Y \leq \tilde{y}\}) = \tau$. By definition, $f(y^*_\tau)$ must be this point, and thus $\Pr^{\Gamma(\cdot, y^*_\tau), \pi^*_2(Y \leq y^*_\tau)} = \tau$. The assumption $f(y^*_\tau) < y^*_\tau$ implies $\Pr^{\Gamma(\cdot, y^*_\tau), \pi^*_2(Y \leq y^*_\tau)} > \tau$. However, when $\Pr^{\Gamma(\cdot, y), \pi^*_2(Y \leq y)}$ is continuous, $\Pr^{\Gamma(\cdot, y^*_\tau), \pi^*_2(Y \leq y^*_\tau)} = \tau$ by the Mean Value Theorem and by the definition of $y^*_\tau$. Thus, we have a contradiction and conclude that (D) holds.

11. QUANTILE INTERACTIVE Q-LEARNING TOY EXAMPLE: $f(y^*_\tau) \neq y^*_\tau$

![Figure 4: Cumulative distribution functions of $Y$ given $H_1 = h_1$ 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 $\Pr(Y \leq y^*_\tau \mid h_1, a_1)$, namely $a_1 = 1$, the resulting quantile, $f(y^*_\tau) = -0.5$, is suboptimal.](image)

Suppose all subjects have the same first-stage covariates, i.e., $H_1 = h_1$ with probability one. Fix $\tau = 0.5$ and let $p(y \mid h_1, a_1)$ denote the conditional density of $Y$ given $H_1 = h_1$.
Then, \( f \) and \( A \) and \( A \). The method used to estimate optimal regimes, respectively.

\[
p(y \mid h_1, 1) = \begin{cases} 
-2.5 \text{ with probability 0.1} \\
-1.5 \text{ with probability 0.2} \\
-0.5 \text{ with probability 0.2} \\
0.5 \text{ with probability 0.2} \\
1.5 \text{ with probability 0.2} \\
2.5 \text{ with probability 0.1}
\end{cases}
\]

and

\[
p(y \mid h_1, -1) = \begin{cases} 
\text{Uniform}(-2, 0) \text{ with probability 0.5} \\
0 \text{ with probability 0.5}
\end{cases}
\]

Then, \( f(y^*_\tau) < y^*_\tau \) because \( y^*_\tau = 0 \) and \( f(y^*_\tau) = -1 \). Recall \( y^*_\tau = \inf\{y : \Pr^{\Gamma}(\cdot, y), \pi_2^*(Y \leq y) \geq \tau\} \) by definition. Figure 4 provides plots of the cumulative distribution functions of \( Y \) when \( A_1 = -1, 1 \). In this example, \( f(y^*_\tau^-) = y^*_\tau \), where \( y^*_\tau^- \) denotes the left limit of \( y^*_\tau \).

12. PROOFS OF THEOREMS 2.3 AND 2.4

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 \( \hat{m}(h_2) \) and \( \hat{c}(h_2) \) that converge in probability to \( m(h_2) \) and \( c(h_2) \), respectively, for each \( h_2 \).

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

A3. For each fixed \( h_1 \) and \( a_1 \), \( \int |d\hat{G}(u, v \mid h_1, a_1) - dG(u, v \mid h_1, a_1)| \) converges to zero in probability.

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

A5. \( \Pr\{|d(H_1, y^*_\tau)| = 0\} = 0 \).

**Theorem 12.3.** (Consistency of TIQ-learning) Assume A1–A3 and fix \( \lambda \in \mathbb{R} \). Then, \( \Pr^{\pi^\text{TIQ}}_\lambda (Y > \lambda) \) converges in probability to \( \Pr^{\pi^\text{TIQ}}_\lambda (Y > \lambda) \), where \( \pi^\text{TIQ}_\lambda = (\hat{\pi}_{1,\lambda}^\text{TIQ}, \hat{\pi}_2^\text{TIQ}) \).

**Theorem 12.4.** (Consistency of QIQ-learning) Assume A1–A5. Then, \( q^{\hat{\pi}^\text{QIQ}}_\tau(\tau) \) converges in probability to \( y^*_\tau \) for any fixed \( \tau \), where \( \hat{\pi}^\text{QIQ}_\tau = (\hat{\Gamma}(\cdot, y^*_\tau^-), \hat{\pi}_2^\text{QIQ}) \).

Capital letters denote random variables and lower case letters denote observed realizations. Let \( D = \{X_{1i}^I, A_{1i}, X_{2i}^I, A_{2i}, Y_i\}_{i=1}^n \) denote the observed data, which are \( n \) independent and identically distributed realizations of the trajectory \( (X_{1i}^I, A_1, X_{2i}^I, A_2, Y)^T \). Let \( (X_{11}^I, A_1, X_{21}^I, A_2, Y)^T \) be a trajectory that is independent of \( D \) but identically distributed. Let \( H_1 = X_1 \) and \( H_2 = (X_{11}^I, A_1, X_{21}^I)^T \) denote the full patient histories available prior to
treatment at stages one and two. When necessary, we use $H_2^{A_1}$ and $H_2^{\pi_1(H_1)}$ to emphasize dependence of $H_2$ 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 $h_2$ are $\pi_2^*(h_2) = \text{sgn}\{c(h_2)\}$ and $\hat{\pi}_2^*(h_2) = \text{sgn}\{\hat{c}(h_2)\}$. In addition, we use the following notation first introduced in Section 2.1:

$$d(h_1, y) = \int F_c(y - u - |v|)dG(u, v \mid h_1, -1) - \int F_c(y - u - |v|)dG(u, v \mid h_1, 1),$$

$$\hat{d}(h_1, y) = \int \hat{F}_c(y - u - |v|)d\hat{G}(u, v \mid h_1, -1) - \int \hat{F}_c(y - u - |v|)d\hat{G}(u, v \mid h_1, 1).$$

With this notation, the optimal and estimated optimal first-stage rules for TIQ-learning are $\pi_{1,t\lambda}^T(h_1) = \text{sgn}\{d(h_1, \lambda)\}$ and $\hat{\pi}_{1,t\lambda}^T(h_1) = \text{sgn}\{\hat{d}(h_1, \lambda)\}$. We define $\text{sgn}(0) = 1$. The following Lemmas are useful for the proofs of Theorems 2.3 and 2.4. In some of the Lemmas, we use $\Delta$ with or without a subscript to denote a difference of two quantities; this notation is used locally, and thus, $\Delta$ appears in multiple Lemmas representing different expressions.

**Lemma 12.5.** If $X_n$ converges to $\mu$ in probability, then $T_n = |\text{sgn}(X_n) - \text{sgn}(\mu)|I_{|\mu|>0}$ converges to zero in probability, and $E(T_n)$ converges to zero as $n$ converges to $\infty$.

**Proof.** If $\mu = 0$, then $\text{pr}(T_n = 0) = 1$ for all $n$. If $\mu > 0$, then $T_n = |\text{sgn}(X_n) - 1|$ and $\text{pr}(T_n > 0) = \text{pr}(X_n < 0)$, which converges to zero. If $\mu < 0$, then $T_n = |\text{sgn}(X_n) + 1|$ and $\text{pr}(T_n > 0) = \text{pr}(X_n > 0)$, which converges to zero. Because $0 \leq T_n \leq 2$ for all $n$, it follows that $E(T_n)$ converges to zero as $n$ converges to $\infty$. \hfill \square

**Lemma 12.6.** Assume A2 and A3. Then, for fixed $h_1$, $\sup_y |\hat{d}(h_1, y) - d(h_1, y)|$ converges to zero.

**Proof.** By the triangle inequality,

$$\sup_y |\hat{d}(h_1, y) - d(h_1, y)| \leq \sup_y |\Delta(y; h_1, -1)| + \sup_y |\Delta(y; h_1, 1)|,$$

where $\Delta(y; h_1, a_1) = \int \hat{F}_c(y - u - |v|)d\hat{G}(u, v \mid h_1, a_1) - \int F_c(y - u - |v|)dG(u, v \mid h_1, a_1)$. Thus, we show $\sup_y |\Delta(y; h_1, a_1)|$ converges in probability to zero for an arbitrary $a_1$. Applying the triangle inequality leads to the upper bound

$$\sup_y |\Delta(y; h_1, a_1)| \leq \sup_y \left| \int \hat{F}_c(y - u - |v|)d\hat{G}(u, v \mid h_1, a_1) - dG(u, v \mid h_1, a_1) \right|$$

$$+ \sup_y \int \left| \hat{F}_c(y - u - |v|) - F_c(y - u - |v|) \right| dG(u, v \mid h_1, a_1).$$

(18)

An upper bound on the right-hand side of (18) is

$$\int \left| d\hat{G}(u, v \mid h_1, a_1) - dG(u, v \mid h_1, a_1) \right| + \sup_w \left| \hat{F}_c(w) - F_c(w) \right| \int dG(u, v \mid h_1, a_1)$$

$$= \int \left| d\hat{G}(u, v \mid h_1, a_1) - dG(u, v \mid h_1, a_1) \right| + \sup_w \left| \hat{F}_c(w) - F_c(w) \right|,$$

(19)
where we have used the fact that \( \sup_w \hat{F}_\epsilon(w) = 1 \) and \( \int dG(u, v \mid h_1, a_1) = 1 \). The first and second terms in (19) are \( o_p(1) \) by assumptions A3 and A2.

**Lemma 12.7.** Assume A1. Then, \( \sup_{\pi_1, y} |pr^{\pi_1, \hat{\pi}_2}(Y \leq y) - pr^{\pi_1, \pi_2}(Y \leq y)| \) converges to zero in probability.

**Proof.** Define \( \hat{\Delta}_c(y; h_2^{a_1}) = F_2 [y - m(h_2^{a_1}) - \text{sgn}\{c(h_2^{a_1})\}c(h_2^{a_1}) - F_1 y - m(h_2^{a_1}) - |c(h_2^{a_1})|] \) and \( \Delta_c(h_2^{a_1}) = |\text{sgn}\{c(h_2^{a_1})\} - \text{sgn}\{c(h_2^{a_1})\}|1_{|c(h_2^{a_1})| > 0} \). Note that for each \( h_2^{a_1} \), \( \hat{\Delta}_c(y; h_2^{a_1}) \leq \Delta_c(h_2^{a_1}) \); thus, using definitions given in Section 2.2, we have

\[
\sup_{\pi_1, y} |pr^{\pi_1, \hat{\pi}_2}(Y \leq y) - pr^{\pi_1, \pi_2}(Y \leq y)| \\
= \sup_{\pi_1, y} \left| \int \int \hat{\Delta}_c(y; h_2^{\pi_1(h_1)}) dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, \pi_1(h_1)\} dF_{H_1}(h_1) \right| \\
\leq \sup_{\pi_1, y} \left| \int \int \hat{\Delta}_c(y; h_2^{\pi_1(h_1)}) dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, \pi_1(h_1)\} dF_{H_1}(h_1) \right| \\
\leq \sup_{\pi_1} \left| \int \int \hat{\Delta}_c(h_2^{\pi_1(h_1)}) dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, \pi_1(h_1)\} dF_{H_1}(h_1) \right|
\]

where we have used the fact that \( \hat{\Delta}_c(h_2^{\pi_1(h_1)}) \) does not depend on \( y \). Because \( \pi_1(\cdot) \) has range \( \{-1, 1\} \), an upper bound on the right-hand side above is

\[
\int \int \sum_{a_1 \in \{-1, 1\}} \hat{\Delta}_c(h_2^{a_1}) dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, a_1\} dF_{H_1}(h_1)
= \sum_{a_1 \in \{-1, 1\}} \int \int \hat{\Delta}_c(h_2^{a_1}) dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, a_1\} dF_{H_1}(h_1)
= \sum_{a_1 \in \{-1, 1\}} E \{\hat{\Delta}_c(H_2^{A_1}) \mid A_1 = a_1, D\}, \quad (20)
\]

which does not depend on \( \pi_1 \). We claim the right-hand side of (20) is \( o_p(1) \). To show this, note for each fixed \( a_1 \),

\[
E \{\hat{\Delta}_c(H_2^{A_1}) \mid A_1 = a_1\} = \int \int E \{\hat{\Delta}_c(h_2^{a_1})\} dF_{H_2 \mid H_1, A_1} \{h_2 \mid h_1, a_1\} dF_{H_1}(h_1),
\]

where \( E\{\hat{\Delta}_c(h_2^{a_1})\} \) converges to zero by Lemma 12.5 for each \( h_2^{a_1} \). Because \( 0 \leq E\{\hat{\Delta}_c(h_2^{a_1})\} \leq 2 \), applying the Dominated Convergence Theorem gives the result that \( E\{\hat{\Delta}_c(H_2^{A_1}) \mid A_1 = a_1\} \) converges to zero, which implies \( E\{\Delta_c(H_2^{A_1}) \mid A_1 = a_1, D\} \) is \( o_p(1) \) for each fixed \( a_1 \) by Lemma 12.5. Thus, the right hand side of (20) is \( o_p(1) \).

**Lemma 12.8.** Assume A2 and A3, and fix \( \lambda \in \mathbb{R} \). Then, \( \left| pr^{\pi_{1, \lambda}, \pi_2}(Y \leq \lambda) - pr^{\pi_{1, \lambda}, \pi_2}(Y \leq \lambda) \right| \) converges to zero in probability.
Proof. Define \( \widehat{\Delta}_G(h_1; u, v) = dG\{u, v \mid h_1, \pi_{1,\lambda}^{\text{TIQ}}(h_1)\} - dG\{u, v \mid h_1, \pi_{1,\lambda}^{\text{TIQ}}(h_1)\} \), and note that \( \widehat{\Delta}_G(h_1; u, v) = \{\pi_{1,\lambda}^{\text{TIQ}}(h_1) - \pi_{1,\lambda}^{\text{TIQ}}(h_1)\}\{dG(u, v \mid h_1, -1) - dG(u, v \mid h_1, 1)\}/2 \). Using the definitions given in Section 2.1,

\[
\left| \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) - \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) \right| = \int \int F_\lambda(\lambda - u - |v|)\widehat{\Delta}_G(h_1; u, v)dF_{H_1}(h_1)
\]

\[
\leq \int |d(h_1, \lambda)| \left| \pi_{1,\lambda}^{\text{TIQ}}(h_1) - \pi_{1,\lambda}^{\text{TIQ}}(h_1) \right| dF_{H_1}(h_1). \tag{21}
\]

Substituting \( \pi_{1,\lambda}^{\text{TIQ}}(h_1) = \text{sgn}\{d(h_1, \lambda)\} \), \( \pi_{1,\lambda}^{\text{TIQ}}(h_1) = \text{sgn}\{d(h_1, \lambda)\} \), and noting

\[
|d(h_1, \lambda)| \left| \text{sgn}\{d(h_1, \lambda)\} - \text{sgn}\{d(h_1, \lambda)\} \right| \leq 1_{|d(h_1, \lambda)| > 0} \left| \text{sgn}\{d(h_1, \lambda)\} - \text{sgn}\{d(h_1, \lambda)\} \right|,
\]

an upper bound on the right-hand side of (21) is

\[
\int 1_{|d(h_1, \lambda)| > 0} \left| \text{sgn}\{d(H_1, \lambda)\} - \text{sgn}\{d(H_1, \lambda)\} \right| dF_{H_1}(h_1)
\]

\[
= E \left[ 1_{|d(H_1, \lambda)| > 0} \left| \text{sgn}\{d(H_1, \lambda)\} - \text{sgn}\{d(H_1, \lambda)\} \right| \mid D \right].
\]

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

\[
E \left[ 1_{|d(h_1, \lambda)| > 0} \left| \text{sgn}\{d(H_1, \lambda)\} - \text{sgn}\{d(H_1, \lambda)\} \right| \mid D \right]
\]

\[
= \int E \left[ 1_{|d(h_1, \lambda)| > 0} \left| \text{sgn}\{d(h_1, \lambda)\} - \text{sgn}\{d(h_1, \lambda)\} \right| \right] dF_{H_1}(h_1).
\]

The inside expectation converges to zero by Lemma [12.5] and applying the Dominated Convergence Theorem gives the result that the right-hand side above converges to zero. Thus, appealing to Lemma [12.5], we have shown \( \left| \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) - \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) \right| \) is bounded above by \( E[1_{|d(H_1, \lambda)| > 0} |\text{sgn}\{d(H_1, \lambda)\} - \text{sgn}\{d(H_1, \lambda)\}| \mid D] \) which is \( o_p(1) \).

Proof of Theorem 2.3. Fix \( \lambda \in \mathbb{R} \). Define \( \Delta(\lambda) = \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) - \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) \).

Then, by the triangle inequality,

\[
|\Delta(\lambda)| \leq \left| \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) - \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) \right|
\]

\[
+ \left| \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) - \text{pr}_{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2}^*}(Y \leq \lambda) \right|. \tag{22}
\]

The first term on the right-hand side of (22) is \( o_p(1) \) by Lemma [12.7] and the second term on the right-hand side of (22) is \( o_p(1) \) by Lemma [12.8].

Lemma 12.9. Assume A2 and A4. Then, \( \sup_y n^{-1} \sum_{i=1}^{n} |\widehat{d}(H_{1i}, y) - d(H_{1i}, y)| \) converges to zero in probability.
Proof. An upper bound on \( \sup_y \frac{1}{n} \sum_{i=1}^n \left| \tilde{d}(H_{1i}, y) - d(H_{1i}, y) \right| \) is

\[
\sum_{a_1=1,-1} \sup_y \frac{1}{n} \sum_{i=1}^n \left| \int \hat{F}_e(y - u - |v|) d\hat{G}(u, v \mid H_{1i}, a_1) - \int F_e(y - u - |v|) dG(u, v \mid H_{1i}, a_1) \right|.
\]

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

\[
\sum_{a_1=1,-1} \sup_y \frac{1}{n} \sum_{i=1}^n \int \left| \hat{F}_e(y - u - |v|) - F_e(y - u - |v|) \right| d\hat{G}(u, v \mid H_{1i}, a_1)
\]
\[
+ \sum_{a_1=1,-1} \sup_y \frac{1}{n} \sum_{i=1}^n \int F_e(y - u - |v|) d\hat{G}(u, v \mid H_{1i}, a_1) - dG(u, v \mid H_{1i}, a_1)
\]
\[
\leq 2 \sup_w \left| \hat{F}_e(w) - F_e(w) \right| + \sum_{a_1=1,-1} \frac{1}{n} \sum_{i=1}^n \int \left| d\hat{G}(u, v \mid H_{1i}, a_1) - dG(u, v \mid H_{1i}, a_1) \right|.
\]

The term \( \sup_y \left| \hat{F}_e(w) - F_e(w) \right| \) is \( o_p(1) \) by assumption A2, and for each \( a_1, \)
\( n^{-1} \sum_{i=1}^n \int \left| d\hat{G}(u, v \mid H_{1i}, a_1) - dG(u, v \mid H_{1i}, a_1) \right| \) is \( o_p(1) \) by assumption A4.

Lemma 12.10. Assume A2 and A4. Then, \( \sup_y |\Delta(y)| \) converges in probability to zero, where

\[
\Delta(y) = \frac{1}{n} \sum_{i=1}^n \int \hat{F}_e(y - u - |v|) d\hat{G}[u, v \mid H_{1i}, sgn\{\tilde{d}(H_{1i}, y)\}]
\]
\[
- \frac{1}{n} \sum_{i=1}^n \int F_e(y - u - |v|) dG[u, v \mid H_{1i}, sgn\{\tilde{d}(H_{1i}, y)\}].
\]  (23)

Proof. Writing \( dG[u, v \mid H_{1i}, sgn\{\tilde{d}(H_{1i}, t)\}] \) as

\[
\frac{1}{2} \left\{ dG(u, v \mid H_{1i}, 1) + dG(u, v \mid H_{1i}, -1) \right\}
\]
\[
- \frac{sgn\{\tilde{d}(H_{1i}, y)\}}{2} \left\{ dG(u, v \mid H_{1i}, -1) - dG(u, v \mid H_{1i}, 1) \right\}
\]

and \( d\hat{G}[u, v \mid H_{1i}, sgn\{\tilde{d}(H_{1i}, y)\}] \) as

\[
\frac{1}{2} \left\{ d\hat{G}(u, v \mid H_{1i}, 1) + d\hat{G}(u, v \mid H_{1i}, -1) \right\}
\]
\[
- \frac{sgn\{\tilde{d}(H_{1i}, y)\}}{2} \left\{ d\hat{G}(u, v \mid H_{1i}, -1) - d\hat{G}(u, v \mid H_{1i}, 1) \right\}.
\]
Lemma 12.11. For every fixed $h_1$,

$$\lim_{y \to \infty} \int F_\epsilon(y - u - |v|) G[u, v \mid h_1, a_1 = sgn\{d(h_1, y)\}] = 1,$$

$$\lim_{y \to -\infty} \int F_\epsilon(y - u - |v|) G[u, v \mid h_1, a_1 = sgn\{d(h_1, y)\}] = 0.$$ 

Proof. For each fixed $h_1$ and $a_1$,

$$\lim_{y \to \infty} \int F_\epsilon(y - u - |v|) dG[u, v \mid h_1, a_1] = 1, \quad \lim_{y \to -\infty} \int F_\epsilon(y - u - |v|) dG[u, v \mid h_1, a_1] = 0,$$

because $\int F_\epsilon(y - u - |v|) dG[u, v \mid h_1, a_1]$ is the conditional expectation of a distribution function in $y$. Thus, even if the policy $sgn\{d(h_1, y)\}$ does not converge as $y \to \infty (-\infty)$, $\lim_{y \to \infty(-\infty)} \int F_\epsilon(y - u - |v|) G[u, v \mid h_1, a_1 = sgn\{d(h_1, y)\}]$ must converge to 1 (0). 

Lemma 12.12. For every $h_1$ in the domain of $H_1$, $\int F_\epsilon(y - u - |v|) dG[u, v \mid h_1, sgn\{d(h_1, y)\}]$ is non-decreasing in $y$. 

Proof. For every fixed $h_1$ and $a_1$,
Proof. We show for arbitrary $s, t \in \mathbb{R}$ such that $s > t$,
\[
\int F_\varepsilon(s - u - |v|)dG[u, v \mid h_1, \text{sgn}\{d(h_1, s)\}] - \int F_\varepsilon(t - u - |v|)dG[u, v \mid h_1, \text{sgn}\{d(h_1, t)\}] \tag{25}
\]
is non-negative. Because $\int F_\varepsilon(s - u - |v|)dG[u, v \mid h_1, \text{sgn}\{d(h_1, s)\}]$ can be written as
\[
\frac{1}{2} \left\{ \int F_\varepsilon(s - u - |v|)dG(u, v \mid h_1, -1) + \int F_\varepsilon(s - u - |v|)dG(u, v \mid h_1, 1) - |d(h_1, s)| \right\},
\]
(25) simplifies to
\[
\frac{1}{2} \left[ \int \{F_\varepsilon(s - u - |v|) - F_\varepsilon(t - u - |v|)\}dG(u, v \mid h_1, -1) \right] \\
+ \frac{1}{2} \left[ \int \{F_\varepsilon(s - u - |v|) - F_\varepsilon(t - u - |v|)\}dG(u, v \mid h_1, 1) \right] \\
- \frac{1}{2} \{ |d(h_1, s)| - |d(h_1, t)| \}.
\]
The expression above is greater than or equal to zero. To see this, note that
\[
|d(h_1, s)| - |d(h_1, t)| \leq ||d(h_1, s)| - |d(h_1, t)|| \leq |d(h_1, s) - d(h_1, t)| \\
\leq \int \{F_\varepsilon(s - u - |v|) - F_\varepsilon(t - u - |v|)\}dG(u, v \mid h_1, -1) \\
+ \int \{F_\varepsilon(s - u - |v|) - F_\varepsilon(t - u - |v|)\}dG(u, v \mid h_1, 1).
\]

Lemma 12.13. Assume $F_\varepsilon(\cdot)$ is continuous. For any fixed $h_1$ in the domain of $H_1$, $\int F_\varepsilon(y - u - |v|)dG[u, v \mid h_1, \text{sgn}\{d(h, y)\}]$ is continuous in $y$.

Proof. This follows immediately by writing $\int F_\varepsilon(y - u - |v|)dG[u, v \mid h_1, \text{sgn}\{d(h, y)\}]$ as
\[
\frac{1}{2} \left\{ \int F_\varepsilon(y - u - |v|)dG(u, v \mid h_1, -1) + \int F_\varepsilon(y - u - |v|)dG(u, v \mid h_1, 1) - |d(h_1, y)| \right\},
\]
a linear combination of continuous functions.

Lemma 12.14. Assume $A2$ and $A4$. Then, $\sup_y |L_n(y) - L(y)|$ converges in probability to
zero, where

\[ L_n(y) = \frac{1}{n} \sum_{i=1}^{n} \int F_\epsilon(y - u - |v|)dG[u, v \mid H_{1i}, \text{sgn}\{d(H_{1i}, y)\}] \]

\[ L(y) = E \left( \int F_\epsilon(y - u - |v|)dG[u, v \mid H_1, \text{sgn}\{d(H_1, y)\}] \right). \]

**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 12.11, 12.12, and 12.13, 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 \( \delta \) is arbitrary, the result holds for each \( \delta \), which implies the limit superior is zero.

**Lemma 12.15.** Assume \( A2 \) and \( A4 \). Then, \( \hat{g}_r^\tau \) converges in probability to \( y^* \).

**Proof.** Define

\[ \Delta(y) = \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_\epsilon(y - u - |v|)d\hat{G}[u, v \mid H_{1i}, \text{sgn}\{\hat{d}(H_{1i}, y)\}] \]

\[ - E \left( \int F_\epsilon(y - u - |v|)dG[u, v \mid H_1, \text{sgn}\{d(H_1, y)\}] \right). \]

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

\[ \Delta_1(y) = \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_\epsilon(y - u - |v|)d\hat{G}[u, v \mid H_{1i}, \text{sgn}\{\hat{d}(H_{1i}, y)\}] \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \int F_\epsilon(y - u - |v|)dG[u, v \mid H_{1i}, \text{sgn}\{d(H_{1i}, y)\}] \]

\[ \Delta_2(y) = \frac{1}{n} \sum_{i=1}^{n} \int F_\epsilon(y - u - |v|)dG[u, v \mid H_{1i}, A_1 = \text{sgn}\{d(H_{1i}, y)\}] \]

\[ - E \left( \int F_\epsilon(y - u - |v|)dG[u, v \mid H_1, A_1 = \text{sgn}\{d(H_1, y)\}] \right). \]

The terms \( \sup_y |\Delta_1(y)| \) and \( \sup_y |\Delta_2(y)| \) converge to zero in probability by Lemmas 12.10 and 12.14, respectively. Thus, \( \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_\epsilon(y - u - |v|)d\hat{G}[u, v \mid H_{1i}, \text{sgn}\{\hat{d}(H_{1i}, y)\}] \) converges
uniformly to \( E \left( \int F_i(y - u - |v|)dG[u, v \mid H_1, \text{sgn}\{d(H_1, y)\}] \right) \), which implies the infimums converge. That is, \( \hat{y}_r^* = \inf \left( y : \frac{1}{n} \sum_{i=1}^n \int \hat{F}_i(y - u - |v|)d\hat{G}[u, v \mid H_{1i}, \text{sgn}\{\hat{d}(H_{1i}, y)\}] \geq \tau \right) \) converges in probability to \( y_r^* = \inf \{ y : E \left( \int F_i(y - u - |v|)dG[u, v \mid H_1, \text{sgn}\{d(H_1, y)\}] \right) \geq \tau \}. \)

**Lemma 12.16.** Assume A2–A4. Let \( h_1 \) be fixed and arbitrary. Then, \( \left| \hat{d}(h_1, \hat{y}_r^*) - d(h_1, y_r^*) \right| \) converges to zero in probability.

**Proof.** By the triangle inequality,

\[
\left| \hat{d}(h_1, \hat{y}_r^*) - d(h_1, y_r^*) \right| \leq \left| \hat{d}(h_1, \hat{y}_r^*) - d(h_1, \hat{y}_r^*) \right| + \left| d(h_1, \hat{y}_r^*) - d(h_1, y_r^*) \right| \\
\leq \sup_y \left| \hat{d}(h_1, y) - d(h_1, y) \right| + \left| d(h_1, \hat{y}_r^*) - d(h_1, y_r^*) \right|.
\]

The right-hand side of the previous expression is \( o_p(1) \) because \( \sup_y \left| \hat{d}(h_1, y) - d(h_1, y) \right| \) is \( o_p(1) \) by Lemma 12.6. Note that continuity of \( d(h_1, y) \) is implied by assumption A2, and thus, \( \left| d(h_1, \hat{y}_r^*) - d(h_1, y_r^*) \right| \) is \( o_p(1) \) by Lemma 12.15 and the continuous mapping theorem. ■

**Proof of Theorem 2.4.** Choose \( \delta > 0 \) such that \( \text{pr}\{|d(H_1, y_r^*)| \leq \delta \} \leq \eta/2 \), which is possible by assumption A5. We begin by showing \( \sup_y |\Delta(y)| \) converges to zero in probability, where

\[
\Delta(y) = \text{pr}\{\text{sgn}\{\hat{d}(\cdot, \hat{y}_r^*)\}, \hat{\pi}_2^* (Y \leq y) - \text{pr}\{\text{sgn}\{d(\cdot, y_r^*)\}, \pi_2^* (Y \leq y) \}.
\]

By the triangle inequality,

\[
\sup_y |\Delta(y)| \leq \sup_y |\Delta_1(y)| + \sup_y |\Delta_2(y)|, \tag{26}
\]

where we define the terms

\[
\Delta_1(y) = \text{pr}\{\text{sgn}\{\hat{d}(\cdot, \hat{y}_r^*)\}, \hat{\pi}_2^* (Y \leq y) - \text{pr}\{\text{sgn}\{d(\cdot, y_r^*)\}, \pi_2^* (Y \leq y) \} \]

and

\[
\Delta_2(y) = \text{pr}\{\text{sgn}\{d(\cdot, y_r^*)\}, \pi_2^* (Y \leq y) - \text{pr}\{\text{sgn}\{d(\cdot, y_r^*)\}, \pi_2^* (Y \leq y) \} \}.
\]

Note that \( \sup_y |\Delta_1(y)| \leq \sup_{\pi_1, \pi_2} \{ \text{pr}\{\pi_1, \hat{\pi}_2^* (Y \leq y) - \text{pr}\{\pi_1, \pi_2^* (Y \leq y) \} \} \), where the right-hand side is \( o_p(1) \) by Lemma 12.7. It can be shown that

\[
\sup_y |\Delta_2(y)| \leq E \left( \sup_y |d(H_1, y)| \left[ \text{sgn}\left\{ \hat{d}(H_1, \hat{y}_r^*) \right\} - \text{sgn}\{d(H_1, y_r^*)\} \right] | D \right) \\
\leq E \left[ \left[ \text{sgn}\left\{ \hat{d}(H_1, \hat{y}_r^*) \right\} - \text{sgn}\{d(H_1, y_r^*)\} \right] | D \right].
\]

We write the right-hand side as

\[
E \left[ \mathbb{1}_{|d(H_1, y_r^*)| \leq \delta} \left| \text{sgn}\left\{ \hat{d}(H_1, \hat{y}_r^*) \right\} - \text{sgn}\{d(H_1, y_r^*)\} \right| | D \right] \\
+ E \left[ \mathbb{1}_{|d(H_1, y_r^*)| > \delta} \left| \text{sgn}\left\{ \hat{d}(H_1, \hat{y}_r^*) \right\} - \text{sgn}\{d(H_1, y_r^*)\} \right| | D \right] \\
\leq 2\text{pr}\{d(H_1, y_r^*) \leq \delta \} \\
+ E \left[ \mathbb{1}_{|d(H_1, y_r^*)| > \delta} \left| \text{sgn}\left\{ \hat{d}(H_1, \hat{y}_r^*) \right\} - \text{sgn}\{d(H_1, y_r^*)\} \right| | D \right]. \tag{27}
\]
By our choice of $\delta$, $2\Pr\{\lvert d(H_1, y^*_r) \rvert \leq \delta \} \leq \eta$. The second term on the right-hand side of (27) is bounded above by $E[\mathbb{1}_{d(H_1, y^*_r) > 0} \lvert \text{sgn}\{\hat{d}(H_1, \hat{y}^*_r)\} - \text{sgn}\{d(H_1, y^*_r)\} \rvert \lvert D]$. Taking the expectation of this upper bound with respect to $D$,

$$E \left[ \mathbb{1}_{d(H_1, y^*_r) > \delta} \lvert \text{sgn}\{\hat{d}(H_1, \hat{y}^*_r)\} - \text{sgn}\{d(H_1, y^*_r)\} \rvert \right] = \int E \left[ \mathbb{1}_{d(h_1, y^*_r) > \delta} \lvert \text{sgn}\{\hat{d}(h_1, \hat{y}^*_r)\} - \text{sgn}\{d(h_1, y^*_r)\} \rvert \right] dF_{H_1}(h_1),$$

where the inside expectation converges to zero by Lemmas 12.5 and 12.16. The Dominated Convergence Theorem applies, giving the result that $E[\mathbb{1}_{d(H_1, y^*_r) > 0} \lvert \text{sgn}\{\hat{d}(H_1, \hat{y}^*_r)\} - \text{sgn}\{d(H_1, y^*_r)\} \rvert \lvert D] \leq o_p(1)$.

We have shown (26) is $o_p(1)$, which implies $q^{\pi_{1,\tau}^Q, \pi_{2}^*}(\tau) = \inf \left\{ y : \Pr\{\hat{\pi}_{1,\tau}^Q, \hat{\pi}_{2}^*(Y \leq y) \geq \tau \} \right\}$ converges in probability to $q^{\pi_{1,\tau}^Q, \pi_{2}^*}(\tau) = \inf \left\{ y : \Pr\{\pi_{1,\tau}^Q, \pi_{2}^*(Y \leq y) \geq \tau \} \right\}$. 

\[\square\]