ABSTRACT

LINF, KRISTIN ASHLEY. Interactive Modeling Techniques for Non-smooth Functionals in Dynamic Treatment Regimes. (Under the direction of Dr. Eric Laber and Dr. Leonard Stefanski.)

Chronic illness treatment strategies must adapt to the evolving health status of the patient receiving treatment. Data-driven dynamic treatment regimes can offer guidance for clinicians and intervention scientists on how to treat patients over time in order to bring about the most favorable clinical outcome on average. Methods for estimating mean-optimal dynamic treatment regimes, such as $Q$-learning, typically require modeling nonsmooth, nonmonotone transformations of data. Thus, building models that fit the data well can be challenging and in some cases may result in a poor estimate of the optimal treatment regime. In Chapter 2, we propose Interactive $Q$-learning ($IQ$-learning) as an alternative to $Q$-learning that only requires modeling smooth, monotone transformations of the data. Consequently, the model building steps in the $IQ$-learning algorithm are amenable to standard modeling checks and residual diagnostics which leads to better-fitting models and higher-quality regimes in many data generative settings.

When the mean is not the most appropriate summary of performance, mean-optimal methods such as $IQ$- and $Q$-learning may not be well-suited to estimate a dynamic treatment regime from data. In two-stage, binary treatment settings, the modeling steps developed as part of $IQ$-learning suggest a more general approach to dynamic treatment regime estimation. In Chapter 3, we derive consistent estimators of regimes that optimize non-mean distributional summaries of the primary clinical outcome. We propose Threshold Interactive $Q$-learning ($TIQ$-learning) to optimize threshold exceedance probabilities of the outcome distribution, and Quantile Interactive $Q$-learning ($IQI$-learning) is proposed to optimize quantiles of the outcome distribution. We provide examples of generative models that induce different optimal regimes for mean and non-mean summaries of the outcome distribution.

In Chapter 4, we introduce Constrained $IQ$-learning to enable estimation of an optimal dynamic treatment regime within a class of regimes that satisfy a constraint on a competing outcome. The classic example of competing outcomes is the tradeoff between efficacy and side-effect burden. Thus, one application of Constrained $IQ$-learning is to estimate a regime that maximizes mean efficacy among a set of regimes that satisfy a pre-specified upper bound on the mean side-effect burden. We develop the method in the simple case of two stages, binary treatments at each stage, and two competing outcomes. We conclude with a discussion of future work regarding Constrained $IQ$-learning.
Interactive Modeling Techniques for Non-smooth Functionals in Dynamic Treatment Regimes

by
Kristin Ashley Linn

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics
Raleigh, North Carolina
2014

APPROVED BY:

Dr. Marie Davidian
Dr. Anastasios Tsiatis

Dr. Eric Laber
Co-chair of Advisory Committee

Dr. Leonard Stefanski
Co-chair of Advisory Committee
DEDICATION

To my family.
BIOGRAPHY

Kristin Ashley Linn was born March 17, 1986 in Raleigh, North Carolina to parents James and Dara Linn. She graduated from Green Hope High School in Morrisville, North Carolina in 2004 and attended the University of Michigan, Ann Arbor to pursue a career as a euphonium player. At Michigan she discovered an interest in statistics and graduated in 2008 with a Bachelor’s degree in Music Performance and a Minor in Statistics. Kristin joined the Department of Statistics at North Carolina State University as a graduate student in August 2009 and received her Master’s degree in Statistics in 2011. Under the direction of Drs. Laber and Stefanski, she will earn her Ph.D. in Statistics in 2014.
ACKNOWLEDGEMENTS

First, I would like to thank my advisors, Dr. Eric Laber and Dr. Leonard Stefanski. It has been a pleasure working with you, and I am grateful for your unwavering patience, respect, and humor. I have the highest regard for your broad knowledge of statistics. Just as importantly, you have demonstrated integrity and kindness in all our interactions. I could not have chosen two better mentors to get me started in statistical research, and this dissertation would not have been possible without your guidance.

I would not have made it this far in my career without professors at NC State, UNC, and Michigan who passionately shared their expert knowledge of statistics with me. I am especially thankful to Dr. Marie Davidian and Dr. Butch Tsiatis for first drawing my attention to dynamic treatment regimes through conference talks and a course on causal inference at NC State. I appreciate the time and constructive feedback they have provided as members of my committee. I offer much thanks to Dr. Spencer Muse for his mentorship and for starting me on the track to biostatistics research through his training grant, GM081057: Biostatistics Training in the Omics Era. In addition, the work in this dissertation was made possible by funding from the Statistical and Mathematical Sciences Institute and Graduate Assistance in Areas of National Need Fellowship. I also thank Dr. Jonathan Ocko for serving on my committee as the graduate school representative.

I offer many thanks to the Department of Statistics staff at NC State who ensured that I got paid for the last five years, provided frequent access to caffeine and calories, and guaranteed headache-free travel to conferences. A special thanks goes to Allison McCoy for always bringing a smile to my face. I also thank DGP’s Dr. Pam Arroway, Dr. Jackie Hughes-Oliver, Dr. Sujit Ghosh, Dr. John Monahan, Dr. Kim Weems, and Dr. Howard Bondell for recruiting me to NC State, identifying funding sources for me every year, and ensuring I met all the requirements of the program.

I am grateful to my E[X^5] family: Brad Turnbull, Kyle White, Joe Usset, Sidd Roy, Jason Osborne, and the band’s first-year predecessors. Playing music and writing lyrics with you was a highlight of the past five years, and I admire your talent and creativity. Thanks are due to Katy Perry for inspiring our first statistically-themed performance and my purple hair tradition.

I attribute much of my success to close friends Gina-Maria Pomann, Kasturi Talapatra, and Alana Unfried. Thank you for providing academic and emotional support throughout all stages of our program. In addition, Maggie Rahmoeller, Sarah Mikol, Stephanie Lepionka, Sarah Hale, Adam Suarez, Ander Wilson, and so many others made graduate school enjoyable and memorable. Many thanks to my fellow group members Kasturi Talapatra, Nick Meyer, Brad Ferguson, Fan Wu, and Yichi Zhang for providing support and constructive feedback on presentations and ideas about future research.

Most of all, I thank my family for their constant love and support. Mom and Dad, when I was young you instilled in me a love for math that has led to this exciting career in statistics. Your wisdom has guided me through many difficult times, notably the first and last years of graduate school. I am very
fortunate to have such kind, intelligent, and encouraging parents. Jacob, I admire your sense of humor and positivity, which always bring me much-needed laughter and stress relief. Being able to share the ups and downs of this journey with all of you and with our extended family has been a blessing.
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii

LIST OF FIGURES ............................................................... ix

**Chapter 1 Introduction** .................................................. 1
  1.1 Dynamic Treatment Regimes ........................................ 1
  1.2 Q-learning ............................................................. 2

**Chapter 2 Interactive Q-learning** .................................... 6
  2.1 Introduction .......................................................... 6
  2.2 Interactive Q-learning (IQ-learning) ................................ 7
    2.2.1 Location-Scale Working Models for $g(\cdot | \mathbf{h}_1, a_1)$ .... 9
    2.2.2 Asymptotic Theory ............................................ 12
  2.3 Monte Carlo Results ................................................ 15
    2.3.1 Power to detect a quadratic effect ......................... 20
  2.4 Application to STAR*D ............................................. 22

**Chapter 3 Interactive Q-learning for Probabilities and Quantiles** ........ 29
  3.1 Introduction ........................................................ 29
  3.2 Generalized Interactive Q-learning ................................ 30
    3.2.1 Two-stage Binary Treatment Models ......................... 30
    3.2.2 Threshold Interactive Q-learning .......................... 31
    3.2.3 Quantile Interactive Q-learning ............................ 34
    3.2.4 Working models for $F_\mathbf{E}(\cdot)$ ....................... 37
    3.2.5 Working models for $G(\cdot, | x_1, a_1)$ .................. 38
    3.2.6 Theoretical results ......................................... 39
    3.2.7 Proofs of Theoretical Results ............................... 40
  3.3 Monte Carlo Results ................................................ 51
    3.3.1 Threshold Interactive Q-learning simulations ............. 51
    3.3.2 Quantile Interactive Q-learning simulations .............. 53
  3.4 Application to STAR*D ............................................. 54

**Chapter 4 Constrained IQ-learning** .................................. 58
  4.1 Introduction ........................................................ 58
  4.2 Optimal Constrained Treatment Regimes .......................... 59
    4.2.1 Notation and Set-up ........................................ 59
    4.2.2 Estimating distributional summaries of $Y$ and $Z$ .......... 61
  4.3 Future work ........................................................ 66

References ................................................................. 68

Appendices ............................................................... 74
  Appendix A  Supplementary Material for IQ-learning ................ 75
LIST OF TABLES

Table 2.1  Variables comprising patient trajectories in the STAR*D data analysis. .................. 25
Table 2.2  Number of patients per treatment strategy by responder status. ......................... 26
Table 3.1  Variables used in the STAR*D analysis. ............................................................ 55
Table 3.2  Estimated value of dynamic and non-dynamic regimes using the Augmented Inverse
           Probability Weighted Estimator. ............................................................................. 56
LIST OF FIGURES

Figure 1.1 SMART design toy example with two randomized stages and two treatment options at each stage. Patients progress from left to right and are randomized to one of two treatment options just prior to Stage 1 and again at Stage 2. Randomizations are represented by gold circles; treatments are displayed in blue boxes. ............. 3

Figure 2.1 Scatterplots of $Y$ (left) and $\Delta(H_2)$ (right) against $X_1$ for $A_1 = -1, 1$ for 1,000 random samples from the toy model. Step Q3 of the Q-learning algorithm requires modeling the data in the left plot; note the nonlinearity and heteroscedasticity. Data in the right plot must be modeled for IQ-learning; note the common analysis of variance structure. ........................................... 7

Figure 2.2 Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p); R^2 = 0.6; p = 4; n = 250$. **Left to Right:** Average proportion of optimal value attained; integrated mean squared error of $Q_1$-function estimates; coverage of 95% confidence intervals for the first-stage $Q$-function, $Q_1(h_1, a_1)$. ................................................................. 17

Figure 2.3 Measures of performance of Q-learning vs. IQ-learning; elements of $\xi$ independent and identically distributed $t_5; R^2 = 0.6; p = 4; n = 250$. **Left to Right:** Average proportion of optimal value obtained; integrated mean squared error of $Q_1$-function estimates; coverage of 95% confidence intervals for the first-stage $Q$-function, $Q_1(h_1, a_1)$. ................................................................. 18

Figure 2.4 **Left:** Values of $X_1$ and $C$ for which Linear Q-learning agrees and disagrees with the true first-stage decision rule. **Right:** Average proportion of optimal value attained of IQ-learning and Q-learning regimes. ................................................................. 19

Figure 2.5 **Left:** Scatterplots of $Y$ against $X_1$ by treatment; the true quadratic relationship is masked by the nonsmooth transformation of data. **Right:** Scatterplots of the contrast $\Delta(H_2; \beta_2)$ against $X_1$ by treatment; the true quadratic relationship is clearly distinguishable. Red lines are cubic smoothing spline fits to the data. .............................. 20

Figure 2.6 **Left:** IQ-learning and Q-learning power to detect $X_1^2$; **Right:** IQ-learning and Q-learning power to detect $A_1 X_1^2$. ................................................................. 21

Figure 2.7 Results for IQ-learning and Q-learning estimators with linear first-stage model terms only. ................................................................. 22

Figure 2.8 Results for IQ-learning and Q-learning estimators with quadratic terms included in all first-stage models. ................................................................. 23

Figure 2.9 **Left:** Scatterplot of $Y$ against baseline QIDS score by first-stage treatment. **Middle:** Scatterplot of the estimated main effect function, $H_{2,0}^T \hat{\beta}_{2,0}$, against baseline QIDS score by first-stage treatment for non-responders. **Right:** Scatterplot of the estimated contrast function, $H_{2,1}^T \hat{\beta}_{2,1}$, against baseline QIDS score by first-stage treatment for non-responders. Cubic smoothing spline fits to the data for $A_1 = 1$ and $A_1 = -1$ are represented by solid gray and dashed black lines. ................................. 26
Figure 3.1 **Left to Right:** \( \lambda = -2, 2, 4 \). Solid black, true optimal threshold probabilities; dotted black, probabilities under randomization; dashed turquoise/purple/orange with circles/squares/triangles, probabilities under \( TIQ_-, Q_-, \) and \( IQ\)-learning.  

Figure 3.2 **Left:** True optimal first-stage treatments for 1,000 test set patients for threshold values \( \lambda = -4, -3, \ldots, 4 \), coded light blue when \( \pi_{1,Q}^{n}(h_1) = 1 \) and red otherwise. **Middle:** \( TIQ\)-learning estimated optimal first-stage treatments. **Right:** \( Q\)-learning estimated optimal first-stage treatments, plotted constant in \( \lambda \). 

Figure 3.3 **Left to Right:** \( \tau = 0.1, 0.5, 0.75 \). Solid black, true optimal quantiles; dotted black, quantiles under treatment randomization; dashed turquoise/purple/orange lines with circles/squares/triangles, quantiles under \( QIQ_-, Q_-, \) and \( IQ\)-learning.
Figure A.17 Measures of performance of Q-learning vs. IQ-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.6$; $p = 8$; $n = 500$. ........................................... 87

Figure A.18 Measures of performance of Q-learning vs. IQ-learning; components of $\xi$ independently drawn from Lognormal(0, 1); ................................................................. 88

Figure A.19 Measures of performance of Q-learning vs. IQ-learning; components of $\xi$ independently drawn from a mixture of Normal(4, 1) and Normal(0, 1), each with probability 0.5. ................................................................. 88

Figure B.1 Cumulative distribution functions of $Y$ given $x_1$ 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 $\Pr(Y \leq y^*_\tau \mid x_1, a_1)$, namely $A_1 = 1$, the resulting quantile, $f(y^*_\tau) = -0.5$, is suboptimal. ................................................................. 93
Chapter 1

Introduction

1.1 Dynamic Treatment Regimes

It is well-established that patients often respond differently to a given treatment, and therefore personalized medicine is a logical and necessary alternative to the “one-size-fits-all” approach to medical care. With the potential for better patient outcomes, reduced treatment burden, higher adherence, and lower cost, there is growing interest in personalized treatment strategies (Hamburg and Collins, 2010; Abraham, 2010; Chakraborty and Moodie, 2013). Chronic diseases and mental illnesses such as HIV/AIDS, cancer, diabetes, depression, schizophrenia, and drug and alcohol addiction typically require ongoing care with multiple interventions. In practice, clinicians and intervention scientists must adapt treatment recommendations in response to the uniquely evolving health status of each patient.

Dynamic treatment regimes (DTRs) formalize this treatment process as a sequence of decision rules. Each decision rule corresponds to a time at which a treatment decision must be made. Often treatment decision times are dictated by milestones in the disease progression or indications of non-response to current treatment. For example, if a patient fails to achieve an adequate response to a first-line therapy, a decision-maker might need to choose among the following strategies: switch the patient to a different treatment, increase the dose of the current treatment, or augment with an additional treatment. Decision rules map current and past patient information to a recommended treatment. One goal of research in the area of dynamic treatment regimes is to estimate high-quality decision rules from data, thus providing a mechanism for objective and evidence-based treatment decisions (Chakraborty and Moodie, 2013).

Data from Sequential Multiple Assignment Randomized Trials (SMARTs; Lavori and Dawson, 2004; Murphy, 2005b) can be used to estimate optimal DTRs. In Chapters 1 and 2, we say a DTR is optimal for a pre-specified outcome if, when applied to assign treatment to a population of interest, it yields the maximal expected outcome. In a SMART, subjects are randomized to treatment at each decision point or stage of the trial. Figure 1.1 contains a visual representation of a SMART design toy example where all subjects receive the same treatment at baseline (e.g., a first-line standard of care).
After some period of time in the baseline stage, patients are then randomized (represented by gold circles) at the start of the first stage to one of two treatment categories: “switch” or “augment” current treatment. After some period of time in the first stage, subjects are again randomized to either switch or augment their current treatment(s) in the second stage. In practice, many variations of this design have been implemented; for example, there are often more than two treatment strategies available at each stage. For ethical reasons, it is common to include an option at each stage for responders to continue their currently successful treatment.

Although it is possible to design a trial with additional stages, two stage SMARTs are common, as evidenced by many recently completed and ongoing trials. For a list of SMARTs that have finished or are in the field, see the website of The Methodology Center at Pennsylvania State University; Director: Dr. Linda Collins (2012, http://methodology.psu.edu/ra/adap-inter/projects) and Laber (2013). With additional randomizations beyond two stages, the number of patients assigned to each sequence of treatment strategies decreases, along with the power to estimate optimal decisions in the later stages. In principle, the sequential randomization scheme in SMARTs guarantees that there are no confounders that influence which types of subjects follow each of the possible treatment sequences. To keep our discussion focused, we work under the assumption of a two-stage SMART with randomized binary treatments at each stage. However, all the methods discussed here apply to observational data when additional assumptions are made on the treatment assignment mechanism (see, for example, Murphy, 2003; Moodie et al., 2012).

### 1.2 Q-learning

We assume data are collected from a two-stage randomized trial with binary treatments at each stage, resulting in \( n \) independent and identically distributed patient trajectories of the form \((X_1, A_1, X_2, A_2, Y)\), where \( X_1 \) and \( X_2 \) are vectors of dimension \( p_1 \) and \( p_2 \), respectively. The variables in the trajectory are: baseline covariates, \( X_1 \in \mathbb{R}^{p_1} \); first-stage randomized treatment, \( A_1 \in \{-1, 1\} \); covariates collected during stage one but prior to the second-stage treatment assignment, \( X_2 \in \mathbb{R}^{p_2} \); second-stage randomized treatment, \( A_2 \in \{-1, 1\} \); and the response, \( Y \in \mathbb{R} \), collected at the conclusion of the trial. We assume \( Y \) has been coded so that higher values indicate more desired clinical outcomes. To simplify notation, we group variables collected prior to each treatment randomization into a history vector \( H_t \), \( t = 1, 2 \). That is, \( H_1 = X_1 \) and \( H_2 = (X_1^\top, A_1, X_2^\top)^\top \).

A DTR is a pair of functions \( \pi = (\pi_1, \pi_2) \) where \( \pi_t \) maps the domain of \( H_t \) into the space of available treatments \( \{-1, 1\} \). Under \( \pi \), a patient presenting at time \( t \) with history \( H_t = h_t \) is assigned treatment \( \pi_t(h_t) \). In Chapters 1 and 2, our goal is to estimate a DTR that, when applied in a population of patients of interest, maximizes the expected response. Define the value of a fixed regime \( \pi \) as \( V^\pi = E^\pi Y \), where \( E^\pi \) denotes the expectation when treatment is assigned according to the policy \( \pi \). The
Figure 1.1: SMART design toy example with two randomized stages and two treatment options at each stage. Patients progress from left to right and are randomized to one of two treatment options just prior to Stage 1 and again at Stage 2. Randomizations are represented by gold circles; treatments are displayed in blue boxes.

optimal treatment regime, \( \pi^{opt} \), maximizes the value function:

\[
E^{\pi^{opt}} Y = \sup_\pi E^\pi Y.
\]

One way to estimate an optimal regime from data is to use the \( Q \)-learning algorithm. Denote the \( Q \)-learning-estimated optimal regime by \( \pi^{Q-opt} = (\pi_{Q-opt}^1, \pi_{Q-opt}^2) \).

\( Q \)-learning (Watkins, 1989; Watkins and Dayan, 1992; Murphy, 2005a) is an approximate dynamic programming algorithm that can be used to estimate an optimal DTR from observational or randomized study data. Define the \( Q \)-functions:

\[
\begin{align*}
Q_2(h_2, a_2) &= E(Y|H_2 = h_2, A_2 = a_2), \\
Q_1(h_1, a_1) &= E \left( \max_{a_2 \in \{1, -1\}} Q_2(H_2, a_2) \mid H_1 = h_1, A_1 = a_1 \right).
\end{align*}
\]

The \( Q \)-function at stage two measures the quality of assigning \( a_2 \) to a patient presenting with history \( h_2 \). Similarly, \( Q_1 \) measures the quality of assigning \( a_1 \) to a patient with \( h_1 \), assuming an optimal decision rule will be followed at stage two. Were the \( Q \)-functions known, dynamic programming (Bellman, 1957) gives the optimal solution, \( \pi_{opt}^t(h_t) = \arg \max_{a_t \in \{-1, 1\}} Q_t(h_t, a_t) \). Since the underlying distri-
bution of the patient histories is not known, the conditional expectations that define the $Q$-functions are unknown and must be approximated. $Q$-learning approximates the $Q$-functions with regression models; linear models are commonly chosen in practice because they are simple and interpretable. We consider linear models of the form:

$$Q_t(h_t, a_t; \beta_t) = h^T_{t,0} \beta_{t,0} + a_t h^T_{t,1} \beta_{t,1}, \quad t = 1, 2,$$

for unknown parameter vector $\beta_t \triangleq (\beta^T_{t,0}, \beta^T_{t,1})^T$, and where $h_{t,0}$ and $h_{t,1}$ each include an intercept and a subset of variables collected in $h_t$. The vectors $h_{t,0}$ and $h_{t,1}$ may or may not be the same summaries of the history $h_t$. The term $h^T_{t,0} \beta_{t,0}$ on the right-hand side of (1.1) contains main effects of patient information. The second term, $a_t h^T_{t,1} \beta_{t,1}$, contains the main effect of treatment $a_t$ and interactions of $a_t$ with elements of the $t^{th}$-stage history. Given this set-up, the $Q$-learning algorithm is given below.

**$Q$-learning Algorithm:**

Q1. (Modeling) Regress $Y$ on $H_2, A_2$ to obtain

$$\tilde{Q}_2(H_2, A_2) = Q_2(H_2, A_2; \beta_2) = H^T_{2,0} \beta_{2,0} + A_2 H^T_{2,1} \beta_{2,1}.$$  

Q2. (Maximization) Define $\tilde{Y} = \max_{a_2 \in \{-1, 1\}} Q_2(H_2, a_2, \beta_2).$ Thus, $\tilde{Y} = H^T_{2,0} \beta_{2,0} + |H^T_{2,1} \beta_{2,1}|$ is the predicted future outcome assuming the optimal decision is made at stage two.

Q3. (Modeling) Regress $\tilde{Y}$ on $H_1, A_1$ to obtain

$$\tilde{Q}_1(H_1, A_1) = Q_1(H_1, A_1; \beta_1) = H^T_{1,0} \beta_{1,0} + A_1 H^T_{1,1} \beta_{1,1}.$$  

The $t^{th}$-stage optimal decision rule then assigns the treatment $a_t$ that maximizes the estimated $Q_t$-function,

$$\hat{\pi}_t^{Q_{opt}}(h_t) = \arg \max_{a_t} Q_t(h_t, a_t; \beta_t).$$

In the version of $Q$-learning with linear models presented above, this can be written as

$$\hat{\pi}_t^{Q_{opt}}(h_t) = \text{sgn}(h^T_{t,1} \beta_{t,1}),$$

where $\text{sgn}(x) = 1_{x > 0} - 1_{x < 0}$. If $h^T_{t,1} \beta_{t,1} = 0$ for some $h_{t,1}$, then there is no unique optimal treatment at stage $t$ for a patient presenting with history $h_t$. To avoid dealing with non-uniqueness, we henceforth
define $\text{sgn}(x) = \mathbb{1}_{x \geq 0} - \mathbb{1}_{x < 0}$ so that treatment 1 is arbitrarily assigned whenever $h_{t,1}^\top \hat{\beta}_{t,1} = 0$ and hence the optimal treatment is not unique.

The first modeling step in the $Q$-learning algorithm is a standard multiple regression problem to which common model building and model checking techniques can be applied to find a parsimonious, well-fitting model. The absolute value in the definition of $\bar{Y}$ arises when $A_2$ is coded as $\{-1, 1\}$, since

$$\arg \max_{a_2} Q_2(H_2, a_2; \hat{\beta}_2) = \text{sgn}(H_{2,1}^\top \hat{\beta}_{2,1}).$$

The second modeling step ($Q^3$) requires modeling the conditional expectation of $\bar{Y}$ given $H_1$ and $A_1$. This can be written as

$$Q_1(H_1, A_1) = E(\bar{Y} | H_1, A_1)$$

$$= E(H_{2,0}^\top \beta_{2,0} + |H_{2,1}^\top \beta_{2,1}| \mid H_1, A_1).$$

An advantage of $Q$-learning is that it is easy to implement using standard statistical software. The method only requires modeling conditional expectations which can be accomplished by fitting a series of regressions. However, due to the absolute value function, $\bar{Y}$ is a nonsmooth, nonmonotone transformation of $H_2$. Thus, the linear model for expression (1.2) in step Q3 is generally misspecified. In addition, the nonsmooth, nonmonotone $\max$ operator in step Q2 leads to difficult nonregular inference for the parameters that index the first stage $Q$-function (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2010; Song et al., 2011; Chakraborty et al., 2013). In the next chapter, we develop an alternative to $Q$-learning, which we call Interactive $Q$-learning ($IQ$-learning), that addresses the applied problem of building good models for the first-stage $Q$-function and avoids model misspecification for a large class of generative models.
Chapter 2

Interactive $Q$-learning

2.1 Introduction

We illustrate the effect of the nonsmooth, nonmonotone transformation in Step Q2 of the $Q$-learning algorithm in Section 1.2 using data from the following generative model:

$$
X_1 \sim \text{Normal}(0, \sigma^2), \quad \xi \sim N(0, \tau^2),
X_2 = \zeta X_1 + \xi, \quad A_t \sim \text{Unif}\{-1, 1\}, \quad t = 1, 2,
\phi \sim N(0, \gamma^2), \quad Y = 1.25A_1A_2 + A_2X_2 - A_1X_1 + \phi, \quad (2.1)
$$

where $\sigma^2$, $\tau^2$, $\zeta$, and $\gamma^2$ are fixed parameters. In most applications, one expects $\zeta = \text{Cov}(X_1, X_2) \neq 0$ because $X_1$ and $X_2$ are often measurements of the same status variable in different stages. Treatments are randomly assigned at each stage as in a sequential multiple assignment randomized trial design (SMART; Murphy, 2005b; Lavori and Dawson, 2000, 2004). The linear working model in (1.1) for $Q_2(H_2, A_2)$ is correct, and thus the resulting predicted value $\tilde{Y}$ approximates the “true” fitted value, $\tilde{Y}_{\text{True}} = |1.25A_1 + X_2| - A_1X_1$. It follows that the regression in Step Q3 approximates the regression of $\tilde{Y}_{\text{True}}$ on $H_1 = X_1$ and $A_1$. Substituting for $X_2$ in $\tilde{Y}$ shows that $\tilde{Y}_{\text{True}} = |1.25A_1 + \zeta X_1 + \xi| - A_1X_1$ from which it is apparent that $E(\tilde{Y}_{\text{True}} \mid H_1, A_1)$ is linear in $X_1$ for fixed $A_1$ only in the practically unlikely case that $\zeta = 0$. Thus, correlation between $X_1$ and $X_2$ induces a nonlinear dependence of $\tilde{Y}$ on $X_1$.

The left-hand side of Figure 2.1 displays a scatterplot of $\tilde{Y}$ against $X_1$ for each value of $A_1$ based on 1,000 random draws from model (2.1) with $\zeta = 0.85$, $\sigma = 1$, and $\tau = \gamma = 1/\sqrt{2}$, using the $Q$-learning algorithm to calculate $\tilde{Y}$. The figure illustrates nonlinearity in the regression of $\tilde{Y}$ on $X_1$ and also heteroscedastic variation induced by the max operation in Step Q2. As this toy model makes clear, identifying the correct form of the regression model for $E(\tilde{Y}_{\text{True}} \mid H_1, A_1)$ and fitting it efficiently would be difficult in the realistic case that the data-generating model is unknown; one approach would
be to adopt nonparametric models for the $Q$-functions (e.g., Zhao et al., 2011; Moodie et al., 2013), however, some clinicians are wary of black-box approaches and it can be difficult to glean scientific knowledge from these models. Thus, common practice is to ignore the problem and settle for the best approximation afforded by fitting linear models. This problem is shared by variants of $Q$-learning as well (for example, $A$-learning, Murphy, 2003; Blatt et al., 2004; Robins, 2004; Schulte et al., 2014). In contrast, the right-hand side of Figure 2.1 shows the first-stage regression model that must be fit in our proposed method, $IQ$-learning, described next. It is a common analysis of covariance model in $X_1$ and $A_1$.

Figure 2.1: Scatterplots of $\tilde{Y}$ (left) and $\Delta(H_2)$ (right) against $X_1$ for $A_1 = -1, 1$ for 1,000 random samples from the toy model. Step Q3 of the $Q$-learning algorithm requires modeling the data in the left plot; note the nonlinearity and heteroscedasticity. Data in the right plot must be modeled for $IQ$-learning; note the common analysis of variance structure.

### 2.2 Interactive $Q$-learning ($IQ$-learning)

$IQ$-learning differs from $Q$-learning in the order in which the maximization step (Q2 in the $Q$-learning algorithm) is performed. We demonstrate how the maximization step can be delayed, enabling all modeling to be performed before this nonsmooth, nonmonotone transformation. This reordering of modeling and maximization steps facilitates the use of standard, interactive model building techniques because all terms to be modeled are linear, and hence smooth and monotone, transformations of the data. For a large class of generative models, $IQ$-learning more accurately estimates the first-stage $Q$-function, resulting
in a higher-quality estimated decision rule (Laber et al., 2013b). Another advantage of IQ-learning is that in many cases, conditional mean and variance modeling techniques (Carroll and Ruppert, 1988) offer a nice framework for the necessary modeling steps. These mean and variance models are interpretable, and the coefficients indexing them enjoy normal limit theory. Thus, they are better suited to inform clinical practice than the misspecified first-stage model in Q-learning whose indexing parameters are nonregular. However, the mean-variance modeling approach we advocate here is not necessary and other modeling techniques may be applied as needed. Indeed, a major advantage and motivation for IQ-learning is the ability for the seasoned applied statistician to build high-quality models using standard interactive techniques for model diagnosis and validation.

IQ- and Q-learning do not differ at step one (Q1 in the Q-learning algorithm from Section 1.2), which we refer to as the second-stage regression. Whereas Q-learning considers modeling $\max_{a_2 \in \{-1, 1\}} Q_2(H_2, a_2)$ directly, IQ-learning starts with the $Q_2$ main-effect and contrast functions:

$$m(H_2) = \frac{1}{2} \{Q_2(H_2, 1) + Q_2(H_2, -1)\},$$

$$\Delta(H_2) = \frac{1}{2} \{Q_2(H_2, 1) - Q_2(H_2, -1)\}.$$ 

The main-effect and contrast functions are linear, and hence smooth and monotone functions of $Q_2$. In addition, these functions are each one-dimensional summaries of $H_2$. Let $g(\cdot \mid h_1, a_1)$ denote the conditional distribution of the contrast $\Delta(H_2)$ given $H_1 = h_1$ and $A_1 = a_1$. With these definitions, $Q_1(h_1, a_1)$ defined in (1.2) can be written as

$$Q_1(h_1, a_1) = E \{m(H_2) \mid H_1 = h_1, A_1 = a_1\} + \int |z|g(z \mid h_1, a_1)dz. \quad (2.2)$$

The IQ-learning estimator of $Q_1(h_1, a_1)$ has the form

$$\hat{Q}^IQ_1(h_1, a_1) = \hat{L}(h_1, a_1) + \int |z|\hat{g}(\cdot \mid h_1, a_1)(z)dz, \quad (2.3)$$

where $\hat{L}(h_1, a_1)$ and $\hat{g}(\cdot \mid h_1, a_1)$ are estimators of $E \{m(H_2) \mid H_1 = h_1, A_1 = a_1\}$ and $g(\cdot \mid h_1, a_1)$.

Let $\hat{Q}_2(H_2, A_2)$ denote the estimator obtained in Step Q1 of the Q-learning algorithm. Define the estimated main-effect and contrast functions,

$$\hat{m}(H_2) = \frac{1}{2} \{\hat{Q}_2(H_2, 1) + \hat{Q}_2(H_2, -1)\},$$

$$\hat{\Delta}(H_2) = \frac{1}{2} \{\hat{Q}_2(H_2, 1) - \hat{Q}_2(H_2, -1)\}. \quad (2.4)$$

Then $\hat{L}(h_1, a_1)$ is obtained by modeling the regression of $\hat{m}(H_2)$ on $H_1$ and $A_1$ for which linear models are often adequate as no unusual transformations are involved. Obtaining $\hat{g}(\cdot \mid h_1, a_1)$ is ac-
accomplished by estimating the conditional distribution of \( \Delta(H_2) \) given \( H_1 \) and \( A_1 \). For this we exploit mean-variance function modeling as explained in Section 2.2.1. Thus, we have the following general form of the algorithm for IQ-learning.

**IQ-learning Algorithm:**

**IQ1.** Use Step Q1 of the Q-learning algorithm to obtain \( \hat{\beta}_2 \) and \( \hat{Q}_2^{IQ}(H_2, A_2) = Q_2(H_2, A_2; \hat{\beta}_2) \).

**IQ2.**

a. Regress the estimated main-effect function \( \hat{m}(H_2) \) from (2.4) on \( H_1 \) and \( A_1 \) to obtain an estimator \( \hat{L}(h_1, a_1) \) of \( E \{ m(H_2) | H_1 = h_1, A_1 = a_1 \} \).

b. Model the conditional distribution of the estimated contrast function \( \hat{\Delta}(H_2) \) from (2.4) given \( H_1 = h_1 \) and \( A_1 = a_1 \) to obtain an estimator \( \hat{g}(\cdot | h_1, a_1) \) of \( g(\cdot | h_1, a_1) \).

c. Combine the estimators from IQ2a and IQ2b to obtain
\[
\hat{Q}_1^{IQ}(h_1, a_1) = \hat{L}(h_1, a_1) + \int |z| \hat{g}(z | h_1, a_1) dz.
\]

**IQ3.** Define the IQ-learning estimated optimal treatment policy \( \hat{\pi}_t^{IQ-opt} = (\hat{\pi}_1^{IQ-opt}, \hat{\pi}_2^{IQ-opt}) \) so that
\[
\hat{\pi}_t^{IQ-opt}(h_t) = \arg \max_{a_t \in \{-1, 1\}} \hat{Q}_t^{IQ}(h_t, a_t).
\]

Completing our algorithm requires specific models for Steps IQ1, IQ2a, and IQ2b. As noted previously, Steps IQ1 and IQ2a are usually straightforward and linear models will often suffice. We now show how to accomplish the modeling in Step IQ2b efficiently and with sufficient flexibility for many applications by using mean-variance models.

**2.2.1 Location-Scale Working Models for** \( g(\cdot | h_1, a_1) \)

The IQ-learning algorithm requires estimating the one-dimensional conditional density \( g(\cdot | h_1, a_1) \). Mean-variance function modeling tools are well-studied (Carroll and Ruppert, 1988) and applicable in the IQ-learning setting. Henceforth, we consider mean-variance, location-scale estimators of \( g(\cdot | h_1, a_1) \) of the form
\[
\hat{g}(z | h_1, a_1) = \frac{1}{\hat{\sigma}(h_1, a_1)} \phi \left\{ \frac{z - \hat{\mu}(h_1, a_1)}{\hat{\sigma}(h_1, a_1)} \right\},
\]
where \( \hat{\mu}(h_1, a_1) \) and \( \hat{\sigma}^2(h_1, a_1) \) are estimators of
\[
\mu(h_1, a_1) = E \{ \Delta(H_2) | H_1 = h_1, A_1 = a_1 \},
\]
\[
\sigma^2(h_1, a_1) = E \left[ (\Delta(H_2) - \mu(h_1, a_1))^2 \right| H_1 = h_1, A_1 = a_1],
\]
respectively. In addition, $\hat{\phi}$ is an estimator of the density of the standardized residuals,

$$\frac{\{\Delta(H_2) - \mu(h_1, a_1)\}}{\sigma(h_1, a_1)},$$

say $\phi_{h_1, a_1}$, which we assume does not depend on the history $h_1$ or the treatment $a_1$. In other words, we assume that all of the dependence of $\Delta(H_2)$ on $(H_1, A_1)$ is captured by the conditional mean and variance functions. The great success of mean-variance function modeling (Carroll and Ruppert, 1988) suggests that this assumption is reasonable quite generally; however, we also note that substantial departures from the assumption can be investigated by stratifying on $h_1$ and $a_1$ and comparing higher-order moments, such as skewness and kurtosis, or nonparametric density estimates of the empirical residuals $\{\hat{\Delta}(H_2) - \hat{\mu}(H_1, A_1)\}/\hat{\sigma}(H_1, A_1)$ across the strata. We now describe two special cases of the estimator in (2.5).

Let $\phi$ denote the density of a standard normal random variable. A simple but useful estimator of $g(\cdot | h_1, a_1)$ is the normal location-scale model:

$$\hat{g}^N(z | h_1, a_1) = \frac{1}{\sigma(h_1, a_1)} \phi \left\{ \frac{z - \mu(h_1, a_1)}{\sigma(h_1, a_1)} \right\}, \quad (2.6)$$

which is a special case of (2.5) with $\hat{\phi} = \phi$. An advantage of this model is that $\int |z|\hat{g}^N(z | h_1, a_1)dz$ can be evaluated in closed form. In particular,

$$\int |z|\hat{g}^N(z | h_1, a_1)dz = \mu(h_1, a_1) \left[ 2\Phi \left\{ \frac{\mu(h_1, a_1)}{\sigma(h_1, a_1)} \right\} - 1 \right] + 2\sigma(h_1, a_1)\phi \left\{ \frac{\mu(h_1, a_1)}{\sigma(h_1, a_1)} \right\}, \quad (2.7)$$

where $\Phi$ is the standard normal cumulative distribution function. If the mean and variance functions are correctly specified and $\phi_{h_1, a_1} = \phi$, then the IQ-learning location-scale model is correct. As commonly implemented, Q-learning fits a misspecified model, and thus estimators are not consistent. The closed form expression in (2.7) makes it possible to study the bias in Q-learning when the true data-generating model is a normal mean-variance function model. Details are in Appendix A.

The normal location-scale model assumes that $\phi_{h_1, a_1} = \phi$, the standard normal density. Violations of normality can be diagnosed via examination of the observed standardized residuals,

$$\hat{e}_i = \{\hat{\Delta}(H_{2,i}) - \hat{\mu}(H_{1,i}, A_{1,i})\}/\hat{\sigma}(H_{1,i}, A_{1,i}),$$

for $i = 1, 2, \ldots, n$, where $n$ is the number of patients in the training data. When greater modeling flexibility is desired, the normality assumption can be dropped and the empirical distribution of the $\hat{e}_i$
used instead. Defining

\[
\tilde{G}(z \mid h_1, a_1) = \frac{1}{n} \sum_{i=1}^{n} 1_{\tilde{e}_i \leq z}, \quad \text{and} \quad \tilde{g}^E(z \mid h_1, a_1) dz = d\tilde{G}(z \mid h_1, a_1)
\]  

(2.8)

leads to the nonparametric location-scale estimator of \(\int |z|g(z \mid h_1, a_1) dz\),

\[
\int |z|\tilde{g}^E(z \mid h_1, a_1) dz = n^{-1} \sum_{i=1}^{n} |\tilde{\mu}(h_1, a_1) + \tilde{\sigma}(h_1, a_1) \tilde{e}_i|.
\]

We show in Section 2.2.2 that the nonparametric location-scale estimator is consistent and asymptotically normal under general conditions. Normal quantile-quantile plots of the standardized residuals may suggest whether the normal or non-parametric location-scale estimator is appropriate in a given data setting.

Data for estimating optimal sequential decision rules are typically expensive to collect, hence sample size is seldom large and parametric mean and variance function models are of greater utility than non-parametric models. Thus, we assume that \(\mu(h_1, a_1) = \mu(h_1, a_1; \theta)\) and \(\sigma(h_1, a_1) = \sigma(h_1, a_1; \gamma)\) for some \(\theta \in \mathbb{R}^{p_\mu}\) and \(\gamma \in \mathbb{R}^{p_\sigma}\). Similarly, we assume that \(L(h_1, a_1) = L(h_1, a_1; \alpha), \alpha \in \mathbb{R}^{p_L}\). For the results in Sections 2.3 we completed specification of the IQ-learning algorithm in Section 2.2 as follows. Steps IQ2a and IQ2b are amended to:

**IQ2a.** Set \(\tilde{L}(h_1, a_1) = L(h_1, a_1; \tilde{\alpha})\), where \(\tilde{\alpha} = \arg \min_{\alpha} \alpha \sum_{i=1}^{n} \{ \tilde{\mu}(H_{2,i}) - L(H_{1,i}, A_{1,i}; \alpha) \}^2\).

**IQ2bi.** Set \(\tilde{\mu}(h_1, a_1) = \mu(h_1, a_1; \tilde{\theta})\), where \(\tilde{\theta} = \arg \min_{\theta} \theta \sum_{i=1}^{n} \{ \tilde{\Delta}(H_{2,i}) - \mu(H_{1,i}, A_{1,i}; \theta) \}^2\).

**IQ2bii.** Set \(\tilde{\sigma}(h_1, a_1) = \sigma(h_1, a_1; \tilde{\gamma})\) where

\[
\tilde{\gamma} = \arg \min_{\gamma} \gamma \sum_{i=1}^{n} \left\{ \log |\tilde{\Delta}(H_{2,i}) - \mu(H_{1,i}, A_{1,i}; \tilde{\theta})| - \log \sigma(H_{1,i}, A_{1,i}; \gamma) \right\}^2.
\]

**IQ2biii.** Set \(\tilde{g}(\cdot \mid h_1, a_1)\) to either \(\tilde{g}^N(\cdot \mid h_1, a_1)\) in (2.6) or to \(\tilde{g}^E(\cdot \mid h_1, a_1)\) in (2.8).

We have used a simple model for the conditional variance in step IQ2bii; for a discussion of other conditional variance estimators and their asymptotic properties see Carroll and Ruppert (1988). It is helpful to note that the first-stage \(Q\)-function in IQ-learning depends only on the main-effect and contrast functions, which are one-dimensional summaries of \(H_2\). In contrast to IQ-learning, \(g\)-estimation (Robins et al., 1992) requires modeling the distribution of \(H_2\) given \(H_1\) and \(A_1\), a more challenging task when \(H_2\) and \(H_1\) are multidimensional or include continuous covariates. Thus, IQ-learning is an attractive alternative to \(g\)-estimation in many practical settings.
2.2.2 Asymptotic Theory

Asymptotic distribution theory for $\widehat{Q}_2^{lq}(h_2, a_2)$ is covered by standard results for linear regression, thus we address only the asymptotic distribution of $\widehat{Q}_1^{lq}(h_1, a_1)$ for the particular parametric estimators discussed in Section 2.2.1.

Define the population residuals

$$E(H_2, H_1, A_1; \theta, \gamma, \beta_2) = \{\Delta(H_2; \beta_2) - \mu(H_1, A_1; \theta)\}/\sigma(H_1, A_1; \gamma)$$

and the population parameters:

$$\beta_2^* = \arg\min_{\beta_2} E\{Y - Q_2(H_2, A_2; \beta_2)\}^2,$$

$$\theta^* = \arg\min_{\theta} E\{\Delta(H_2; \beta_2^*) - \mu(H_1, A_1; \theta)\}^2,$$

$$\gamma^* = \arg\min_{\gamma} E\{\log|\Delta(H_2; \beta_2^*) - \mu(H_1, A_1; \theta^*)| - \log\sigma(H_1, A_1; \gamma)\}^2,$$

$$\alpha^* = \arg\min_{\alpha} E\{m(H_2; \beta_2^*) - L(H_1, A_1; \alpha)\}^2.$$

Let $\hat{\theta}, \hat{\gamma}, \hat{\beta}_2$, and $\hat{\alpha}$ denote $n^{1/2}$-consistent estimators of their population analogs $\theta^*, \gamma^*, \beta_2^*$, and $\alpha^*$. For $x \in \mathbb{R}^p$, let $B_d(x)$ denote a ball of radius $d$ centered at $x$, and let $E_n$ denote the empirical expectation operator so that $E_n f = n^{-1} \sum_{i=1}^n f(H_{1,i}, A_{1,i}, H_{2,i}, A_{2,i}, Y_i)$. The asymptotic results are stated in terms of the seven centered statistics:

$$\Delta_L = L(h_1, a_1; \hat{\alpha}) - L(h_1, a_1; \alpha^*),$$

$$\Delta_\mu = \mu(h_1, a_1; \hat{\theta}) - \mu(h_1, a_1; \theta^*),$$

$$\Delta_\beta = \hat{\beta}_2 - \beta_2^*,$$

$$\Delta_\sigma = \sigma(h_1, a_1; \hat{\gamma}) - \sigma(h_1, a_1; \gamma^*),$$

$$\Delta_\theta = \hat{\theta} - \theta^*,$$

$$\Delta_\gamma = \hat{\gamma} - \gamma^*,$$

and

$$\Delta_\varepsilon = E_n\{[\mu(h_1, a_1; \theta^*) + \sigma(h_1, a_1; \gamma^*)E(H_2, H_1, A_1; \theta^*, \gamma^*, \beta_2^*)] - E([\mu(h_1, a_1; \theta^*) + \sigma(h_1, a_1; \gamma^*)E(H_2, H_1, A_1; \theta^*, \gamma^*, \beta_2^*)])\}.$$ 

The following assumptions are used to establish the limit theory for $lq$-learning.

(A1N) $n^{1/2} (\Delta_L, \Delta_\mu, \Delta_\sigma)$ is asymptotically Normal $\{0, \Sigma_N(h_1, a_1)\}$. 

12
Theorem 2.2.1

II and nonparametric location-scale estimators, respectively, so that learning estimators of the first-stage to be well-behaved.

Assumption (A3) is generally more difficult to verify, but its validity can be roughly assessed using the observed residuals \( \hat{e}_1, \ldots, \hat{e}_n \). We have not sought the most general assumptions, but rather a set of simple assumptions that illustrate what is needed for the IQ-learning estimator to be well-behaved.

The first result states the asymptotic normality of the normal and nonparametric location-scale IQ-learning estimators of the first-stage Q-function. Let \( \hat{Q}^{IQ,N}_1(h_1, a_1) \) and \( \hat{Q}^{IQ,E}_1(h_1, a_1) \) denote the normal and nonparametric location-scale estimators, respectively, so that

\[
\hat{Q}^{IQ,N}_1(h_1, a_1) = L(h_1, a_1; \hat{\alpha}) + \int |z|^g(z \mid h_1, a_1)dz,
\]

\[
\hat{Q}^{IQ,E}_1(h_1, a_1) = L(h_1, a_1; \hat{\alpha}) + \int |z|^g(z \mid h_1, a_1)dz.
\]

Define \( I(v, t, s) = v + s^{-1} \int |z|\phi(\{z - t\}/s)dz \). Let \( I^*(h_1, a_1) \) denote \( I\{L(h_1, a_1; \alpha^*), \mu(h_1, a_1; \theta^*), \sigma(h_1, a_1; \gamma^*)\} \). The following is proved in Appendix A.

**Theorem 2.2.1** (Asymptotic normality). Let \( h_1 \) and \( a_1 \) be fixed.

1. Assume (A1N). Then

\[
n^{1/2} \left[ \hat{Q}^{IQ,N}_1(h_1, a_1) - L(h_1, a_1; \alpha^*) - \frac{1}{\sigma(h_1, a_1; \gamma^*)} \int |z|\phi\left\{ \frac{z - \mu(h_1, a_1; \theta^*)}{\sigma(h_1, a_1; \gamma^*)} \right\} dz \right]
\]

converges in distribution to Normal\{0, \( \nabla I^*(h_1, a_1)\Sigma_N(h_1, a_1)\nabla I^*(h_1, a_1) \}.

13
2. Assume (A1E), (A2), and (A3). Then

\[ n^{1/2} \left( \hat{Q}^{1Q,E}(h_1, a_1) - L(h_1, a_1; \alpha^*) - \frac{1}{\sigma(h_1, a_1; \gamma^*)} \int |z| \kappa \left\{ \frac{z - \mu(h_1, a_1; \theta^*)}{\sigma(h_1, a_1; \gamma^*)} \right\} dz \right) \]

converges in distribution to

Normal[0, \{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^\top, 1\} \Sigma_E(h_1, a_1)\{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^\top, 1\}^\top].

Theorem 2.2.1 shows that both location-scale estimators \( \hat{Q}^N(h_1, a_1) \) and \( \hat{Q}^E(h_1, a_1) \) are asymptotically normal under the stated conditions, which do not require correct specification of the IQ-learning models. Under (C1) and (C2) below, the IQ-learning models are correctly specified, and consistency and asymptotic normality of the IQ-learning estimators follow.

(C1) \( \Delta(H_2; \beta_2^*) = \mu(H_1, A_1; \theta^*) + \sigma(H_1, A_1; \gamma^*) Z \), where \( Z \) denotes a standard normal random variable.

(C2) \( \Delta(H_2; \beta_2^*) = \mu(H_1, A_1; \theta^*) + \sigma(H_1, A_1; \gamma^*) W \), where \( W \) is a random variable with density function \( \kappa(\cdot) \).

The following results are direct consequences of Theorem 1 and we omit their proofs.

**Corollary 2.2.2.** Assume \( Q_2(h_2, a_2) = Q_2(h_2, a_2; \beta_2^*) \) and \( E\{m(H_2; \beta_2^*)|H_1 = h_1, A_1 = a_1\} = L(h_1, a_1; \alpha^*) \). Then

1. (C1) \( \implies Q_1(h_1, a_1) = L(h_1, a_1; \alpha^*) + \frac{1}{\sigma(h_1, a_1; \gamma^*)} \int |z| \phi \left\{ \frac{z - \mu(h_1, a_1; \theta^*)}{\sigma(h_1, a_1; \gamma^*)} \right\} dz \).

2. (C2) \( \implies Q_1(h_1, a_1) = L(h_1, a_1; \alpha^*) + \frac{1}{\sigma(h_1, a_1; \gamma^*)} \int |z| \kappa \left\{ \frac{z - \mu(h_1, a_1; \theta^*)}{\sigma(h_1, a_1; \gamma^*)} \right\} dz \).

**Theorem 2.2.3.** Assume (A1N) and the conditions of Theorem 2.2.2. Let \( h_1 \) and \( a_1 \) be fixed.

1. (C1) \( \implies n^{1/2} \left\{ \hat{Q}^{1Q,N}(h_1, a_1) - Q_1(h_1, a_1) \right\} \text{ converges in distribution to} \text{ Normal}[0, \nabla \Gamma^*(h_1, a_1)^\top \Sigma_N(h_1, a_1) \nabla \Gamma^*(h_1, a_1) ] \).

2. (A1E), (A2), (A3) and (C2) \( \implies n^{1/2} \left\{ \hat{Q}^{1Q,E}(h_1, a_1) - Q_1(h_1, a_1) \right\} \text{ converges in distribution to} \text{ Normal}[0, \{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^\top, 1\} \Sigma_E(h_1, a_1)\{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^\top, 1\}^\top]. \)

**Remark 2.2.4.** Theorem 2.2.3 can be used to construct asymptotically valid confidence intervals for the first-stage Q-function for fixed patient history \( h_1 \) and first-stage treatment \( a_1 \). This is a notoriously difficult task with Q-learning (see Laber et al., 2010). In practice, due to the complexity of the variance terms, the bootstrap may be preferred.
Remark 2.2.5. As noted in the introduction, IQ-learning does not alleviate the inherent nonregularity present in sequential decision making problems (see Robins, 2004; Laber et al., 2010; Chakraborty et al., 2010). However, IQ-learning is consistent for a nonregular scenario of interest, the so-called ‘global null’ in which there is no treatment effect for any patients at the second stage, i.e., $\Delta(H_2; \beta^*_2) = 0$ almost surely. To see this, note that assuming (A1N), (C1) holds with $\mu(H_1, A_1; \theta^*) = 0$ with $\sigma(H_1, A_1; \gamma^* \rightarrow 0$ almost surely. Part 1 of Corollary 1 depends only on part 1 of Theorem 1 and part 1 of Theorem 2. For the more complex case in which $0 < \Pr\{\Delta(H_2; \beta^*_2) = 0\} < 1$ we conjecture that using a mixture of normals to estimate $g(\cdot | h_1, a_1)$ may lead to improved small–sample performance.

2.3 Monte Carlo Results

We compare the small-sample performance of IQ- and Q-learning in terms of: (i) average value of the learned treatment regimes; (ii) integrated mean squared error of the first-stage Q-function; and (iii) coverage and width of 95% nonparametric bootstrap confidence intervals for the first-stage Q-function. A key advantage of IQ-learning is its compatibility with common model building steps that we illustrate with a study of the power to detect nonlinear effects in the first-stage Q-function using IQ-learning and Q-learning. Software implementing the IQ-learning estimators is available as part of the ‘iqLearn’ package on the comprehensive R network (cran.us.r-project.org/). The Monte Carlo results show that for the class of generative models we consider, IQ-learning generally performs favorably to Q-learning in terms of value, integrated mean squared error, coverage, and width of the confidence intervals. Simulations also show IQ-learning also has higher power to detect nonlinear effects.

Here we use data from the following class of generative models:

$$X_1 \sim \text{Normal}_p\{0.1, \Omega_{AR1}(0.5)\}, \quad A_t \sim \text{Uniform}\{-1, 1\}, \quad t = 1, 2,$$

$$X_2 = (1.5 - 0.5A_1)X_1 + \zeta_{A_1}, \quad Y = H_{2,0}^T\beta_{2,0} + A_2H_{2,1}^T\beta_{2,1} + \phi,$$

where $\{\Omega_{AR1}(0.5)\}_{i,j} = (0.5)^{|i-j|}, H_{2,0} = H_{2,1} = (1, X_1^T, A_1, A_1X_2^T)^T$, and $\zeta_{A_1} = (1.5 + 0.5A_1)^{1/2}$. Thus, the class is indexed by the dimension $p$, the distributions of $\xi$ and $\phi$, and the coefficient vectors $\beta_{2,0}$ and $\beta_{2,1}$. Here we fix $p = 4$; results for $p = 8$ are similar and are provided in Appendix A. We consider $\xi \sim \text{Normal}_p(0, I_p)$. We fix the main effect parameter $\beta_{2,0}$ and vary the second-stage treatment effect size by scaling $\beta_{2,1}$ as follows:

$$\beta_{2,0} = \frac{1_{2p+2}}{||1_{2p+2}||}, \quad \beta_{2,1} = C \frac{(-0.251_{p+1}, 1_{p+1})^T}{||(-0.251_{p+1}, 1_{p+1})||};$$

where $C$ ranges over a grid from 0 to 2, and $1_d$ denotes a $d$-dimensional vector of 1s. In addition, we fix the theoretical $R^2$ of the second-stage regression model at 0.6 by specifying $\phi \sim \text{Normal}\{0, \sigma^2_\phi(C)\}$.
and solving for the variance $\sigma^2_0(C)$ that yields the desired $R^2$. Additional simulations, provided in Appendix A, show results for $R^2 = 0.4$, 0.8 and non-normal error distributions for $\xi$.

We consider linear working models for the mean and log-variance functions

$$Q_2(h_2,a_2;\beta_2) = h_2^\top \beta_{2,0} + a_2 h_2^\top \beta_{2,1}, \quad Q_1(h_1,a_1;\beta_1) = h_1^\top \beta_{1,0} + a_1 h_1^\top \beta_{1,1},$$

$$L(h_1,a_1;\alpha) = h_1^\top \alpha_0 + a_1 h_1^\top \alpha_1, \quad \mu(h_1,a_1;\theta) = h_1^\top \theta_0 + a_1 h_1^\top \theta_1,$$

$$\log\{\sigma(h_1,a_1;\gamma)\} \propto h_1^\top \gamma_0 + a_1 h_1^\top \gamma_1,$$

where now $H_1 = (1, X_1^\top)$. In addition to Q-learning with linear working models, we include results using support vector regression using a Gaussian kernel (Zhao et al., 2011) to estimate both $Q$-functions.

We consider two versions of the IQ-learning estimator that differ in the estimation of $g(\cdot \mid h_1,a_1)$ and the model for $\sigma(h_1,a_1;\gamma)$: (i) normal estimator $\hat{g}^N(\cdot \mid h_1,a_1)$ of the residual distribution and a restricted variance model, $\log\{\sigma(h_1,a_1;\gamma)\} = \gamma_0 + a_1 \gamma_1$, that depends only on treatment (NormHomo); and (ii) nonparametric estimator $\hat{g}^E(\cdot \mid h_1,a_1)$ of the residual distribution with a log-linear variance model that depends on $h_1$ and $a_1$ (NonparHetero). When $\xi \sim \text{Normal}_p(0, I_p)$, both these estimators are correctly specified. Q-learning is always correctly specified at the second stage but only correctly specified at the first stage when $C = 0$ and hence $\beta_{2,1} = 0$.

Results are based on a training set of size $n = 250$ and $M = 2,000$ Monte Carlo data sets for each generative model. Additional results for $n = 500$ are provided in Appendix A. For NonparHetero, which is always correctly specified, the ‘true’ Q-functions and subsequent optimal regime are estimated using a test set of 10,000 observations. Recall that the value, $E^\pi Y$, of an arbitrary policy $\pi$ is the expected outcome if all patients are assigned treatment according to $\pi$, that is,

$$E^\pi Y = E(E[Y \mid H_2, a_2] \mid a_2 = \pi_2(H_2) \mid H_1, a_1)\mid a_1 = \pi_1(H_1)).$$

For a given training set of size $n$ and an algorithm that produces an estimated optimal policy, say $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$, we define the average value as $E(E^\pi Y)$ where the outer expectation is taken over all training sets of size $n$. We estimate the average value of both the IQ-learning estimators and the Q-learning estimators using a test set of size 10,000 to estimate the inner expectation and 2,000 Monte Carlo replications to estimate the outer expectation. We compare average values of the learned IQ- and Q-learning regimes to the value of the true optimal regime and present the proportion of optimal value obtained. For an estimator $\hat{Q}_1(h_1,a_1)$ of the first-stage $Q$-function, $Q_1(h_1,a_1)$, define the integrated mean squared error as $E[(\hat{Q}_1(H_1,A_1) - Q_1(H_1,A_1))^2]$ where the expectation is taken over the joint distribution of $(H_1,A_1)$ as well as the training data used to obtain $\hat{Q}_1(h_1,a_1)$.

Confidence intervals for $Q_1(h_1,a_1)$ based on IQ-learning and Q-learning estimators are formed by bootstrapping the respective estimators and taking percentiles. For example, if $\bar{l}$ and $\bar{u}$ denote the $100 \times \eta/2$ and $100 \times (1 - \eta/2)$ percentiles of the bootstrap distribution of $\hat{Q}^{IQ}(h_1,a_1)$ based on
Figure 2.2: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 4$; $n = 250$. **Left to Right:** Average proportion of optimal value attained; integrated mean squared error of $Q_1$-function estimates; coverage of 95% confidence intervals for $Q_1(h_1, a_1)$; width of 95% confidence intervals for the first-stage $Q$-function, $Q(h_1, a_1)$.

With 1,000 bootstrap resamples, then the $100 \times (1 - \eta)\%$ confidence interval is given by $\{2\hat{Q}^{IQ}(h_1, a_1) - \hat{u}, 2\hat{Q}^{IQ}(h_1, a_1) - \hat{l}\}$. Bootstrap intervals of this form are sometimes referred to as ‘hybrid bootstrap’ confidence intervals (Efron and Tibshirani, 1994). Coverage and width of the foregoing confidence intervals are estimated using 2,000 Monte Carlo replications with a new instance $(h_1, a_1)$ of $(H_1, A_1)$ drawn for each replication. Figure (2.2) displays the results from this simulation, where $\xi \sim \text{Normal}_p(0, I_p)$. Results with elements of $\xi$ generated independently from a $t$-distribution with five degrees of freedom are given in Figure 2.3.

The plots in Figure (2.2) indicate that the NormHomo and NonparHetero $IQ$-learning estimators perform favorably in comparison to both $Q$-learning estimators. Although some gains are achieved using the more flexible support vector regression version of $Q$-learning, the top left plot indicates that the learned regimes from $IQ$-learning obtain higher average value than the average value of the $Q$-learning regimes across most values of $C$. The top right plot shows that the $IQ$-learning estimators reduce integrated mean squared error the most, with greater reduction as the second-stage effects increase. $IQ$-learning also demonstrates a large improvement over linear $Q$-learning in terms of the coverage of...
Figure 2.3: Measures of performance of $Q$-learning vs. $IQ$-learning; elements of $\xi$ independent and identically distributed $t_5$; $R^2 = 0.6$; $p = 4$; $n = 250$. **Left to Right:** Average proportion of optimal value obtained; integrated mean squared error of $Q_1$-function estimates; coverage of 95% confidence intervals for $Q_1(h_1, a_1)$; width of 95% confidence intervals for $Q_1(h_1, a_1)$. 
95% confidence intervals for $\hat{Q}_1(h_1, a_1)$, as seen in the bottom left plot. The poor coverage of linear $Q$-learning is attributed to bias, whereas $IQ$-learning is consistent for the generative models in this section. Thus, $IQ$-learning estimators come close to achieving the nominal level. The average widths of the confidence intervals are similar for linear $Q$-learning and $IQ$-learning, as illustrated by the bottom right plot. Support vector regression $Q$-learning greatly improves coverage compared to linear $Q$-learning but at the expense of wider intervals. Results from the simulation where the elements of $\xi$ are generated from a $t_5$ distribution are included in Appendix A and appear similar to those in Figure (2.2), suggesting NormHomo is robust to slight misspecification of the residual distribution.

Next, we consider the generative model

$$
X_1 \sim \text{Normal}(-2, 1), \quad \xi \sim \text{Normal}(0, 1), \quad A_t \sim \text{Uniform}\{-1, 1\}, \quad t = 1, 2, \\
X_2 = X_1 + \xi, \quad \phi \sim \text{Normal}(0, 1), \quad Y = H_{2,0}^\intercal \beta_{2,0} + A_2 H_{2,1}^\intercal \beta_{2,1} + \phi,
$$

where $H_{2,0} = H_{2,1} = (1, X_1, A_1, X_1 A_1, X_2)^\intercal$, $\beta_{2,0} = (3, -1, .1, -1, -1)^\intercal$, and $\beta_{2,1} = C(-6, -2.5, 3, -2)^\intercal$. We use this example to illustrate a scenario where $IQ$-learning achieves a large gain in value over $Q$-learning with linear models. Since the predictor $X_1$ is univariate, we can visualize which patients are treated differently by $IQ$-learning compared to $Q$-learning.

The left plot in Figure (2.4) was obtained by deriving the true first stage $Q$-function, for which $IQ$-learning is consistent in this case, and comparing the true first-stage rule to the rule recommended by $Q$-learning. For each combination of $(X_1, C)$, the plot shows whether or not $Q$-learning makes the correct treatment decision. With this generative model, $Q$-learning assigns the wrong treatment to approximately half the population for a wide range of effect sizes. In contrast, with a sufficiently large sample size, $IQ$-learning treats all patients according to the optimal rule. Consequently, $IQ$-learning achieves higher average value than $Q$-learning, as displayed in the right plot of Figure (2.4).
are based on \( n = 250 \) training set samples and \( M = 1,000 \) Monte Carlo data sets. In this scenario, both IQ-learning and support vector regression \( Q \)-learning reach gains in optimal value attained of approximately 15% as the second-stage effect size grows.

### 2.3.1 Power to detect a quadratic effect

One strength of IQ-learning is that it enables practitioners to apply standard interactive model building techniques. We now consider a generative model with a univariate predictor \( X_1 \) and nonlinear relationship between \( X_1 \) and \( X_2 \). The new generative model is

\[
X_1 \sim \text{Normal}(1, 1), \quad A_t \sim \text{Uniform}\{-1, 1\}, \ t = 1, 2, \\
X_2 = X_1^2 + (1.5 - 0.5A_1)X_1 + \zeta_A, \xi, \quad \xi \sim \text{Normal}(0, 1), \\
\phi \sim \text{Normal}(0, 4), \quad Y = H_{2,0}^t\beta_{2,0} + A_2H_{2,1}^t\beta_{2,1} + \phi,
\]

where \( H_{2,0} = H_{2,1} = (1, X_2, A_1, A_1X_2)^T \) and \( \zeta_A = \sqrt{1.5 + 0.5A_1} \). Thus, the true \( Q \)-function depends on both \( X_1^2 \) and \( A_1X_1^2 \). As in the main paper, we fix \( \beta_{2,0} \) and scale \( \beta_{2,1} \). We specify the second-stage effects to be

\[
\beta_{2,0} = \frac{(3, -1.5, 4, -1)^T}{||(3, -1.5, 4, -1)^T||}, \quad \beta_{2,1} = C\frac{(2, -1, 2, -0.5)^T}{||(2, -1, 2, -0.5)^T||},
\]

for \( C \in (0, 2) \). Figure (2.5) illustrates how the quadratic effect of \( X_1 \) is masked by the absolute value operator in \( Q \)-learning. Alternatively, the quadratic relationship is clearly visible in the scatter plots of the contrast function \( \Delta(H_2; \beta_2) \) against \( X_1 \). The solid lines in Figure (2.5) are cubic smoothing splines fitted to the data using ordinary cross validation.

Figure (2.6) displays plots of the power to detect the quadratic effects \( X_1^2 \) and \( A_1X_1^2 \) as a function of
Figure 2.6: **Left:** IQ-learning and Q-learning power to detect $X_1^2$; **Right:** IQ-learning and Q-learning power to detect $A_1X_1^2$.

The second-stage effect size scaling constant $C$. The Q-learning curve represents the power to detect the quadratic terms in the regression of the pseudo outcome $\tilde{Y}$ on the first-stage history and treatment. The two identical IQ-learning curves represent the power to detect the quadratic effects in the regression of the contrast function on the first-stage information to obtain an estimate the contrast mean function. Results are based on $n = 250$ training samples, and the power was calculated by averaging over indicators from $M = 1,000$ Monte Carlo data sets of whether the estimated coefficients of $X_1^2$ and $A_1X_1^2$ were found to be significant by a $t$-test. When the treatment interaction effects are near zero, i.e., $C \approx 0$, Q-learning detects the nonlinear relationships because the pseudo outcome $\tilde{Y}$ is dominated by the linear main-effect term $\mathbf{H}_2^T \beta_{2,0}$, which is a function of both $X_1^2$ and $A_1X_1^2$. At first glance, IQ-learning appears to perform worse than Q-learning when the effect size is small. However, this is due to the fact that Figure (2.6) only displays results from the regression of the contrast function on first-stage information, and $C \approx 0$ implies $\Delta(\mathbf{H}_2; \beta_2) \approx 0$. Results from the regression of the main-effect term $\mathbf{H}_2^T \beta_{2,0}$ on first-stage information are not included in Figure (2.6).

The power of Q-learning to detect the quadratic terms decreases drastically as the second-stage treatment effects increase because the absolute value from the maximization operator masks the true underlying structure. We note that the parameters that index the first-stage Q-function are nonregular, so the $t$-tests for significance are invalid. In comparison, the first-stage IQ-learning coefficients are asymptotically normal. Thus, $t$-tests are approximately valid and they detect the quadratic relationships in the mean of the contrast function with increasing accuracy as the treatment effects grow larger.

In Figures (2.7) and (2.8), we provide results from the same model when: 1) all first-stage IQ- and Q-learning models include linear terms only, and 2) all first-stage IQ- and Q-learning models include a quadratic term. Although linear Q-learning outperforms both misspecified linear IQ-learning estimators in terms of IMSE, the average value of the NonparHetero estimator is comparable with Q-learning.
In addition, the correctly specified quadratic version of NonparHetero outperforms $Q$-learning with quadratic terms with respect to all four displayed measures of performance.

## 2.4 Application to STAR*D

Sequenced Treatment Alternatives to Relieve Depression (STAR*D; Fava et al., 2003; Rush et al., 2004) is a sequentially randomized study of major depressive disorder. A key feature of this trial is that patients experiencing early remission of symptoms were exempt from future randomization, complicating the analysis. We use a subset of the STAR*D data to illustrate how $IQ$-learning can be used to estimate an optimal dynamic treatment regime in the presence of responder-status dependent designs. There were four stages in the trial, but each patient received Citalopram in the first stage, and thus, there was no randomization. Since our aim is to demonstrate how to learn a regime using the $IQ$-learning machinery, we opt to perform a complete-case analysis and consider only the first two of three randomized stages. We refer to the second and third stages as stages one and two, respectively. At each stage, treatments can be categorized as “SSRI” or “non-SSRI,” where SSRI stands for Selective Serotonin Reuptake Inhibitor;
Figure 2.8: Results for IQ-learning and Q-learning estimators with quadratic terms included in all first-stage models.
this is the binary treatment variable in our analysis. The first-line treatment Citalopram given in the non-
randomized stage is in the class of SSRIs.

We use a measure of efficacy, the Quick Inventory of Depression Symptomatology (QIDS) score,
as the outcome (Rush et al., 2004); side-effects or other competing outcomes could be accommodated
using set-valued treatment regimes (Lizotte et al., 2012; Laber et al., 2013a) or composite outcomes
(Wang et al., 2012a). The QIDS score ranges from 0 to 27 with higher values corresponding to more
severe negative symptoms. To be consistent with our development, we recode these scores as $27 - \text{QIDS}$.

With this recoding, higher values correspond to better clinical outcomes. The QIDS score was recorded
at multiple time points throughout each stage, therefore, intermediate QIDS scores used as predictors
in our analysis are also recoded (see Rush et al., 2004 or Schulte et al., 2014 for more details about
the STAR*D trial design). At each stage, patients experiencing remission left the study. Remission was
declared as a QIDS score $\leq 5$, or, recoded QIDS $\geq 22$. Henceforth, we refer to patients who achieved
remission and left after stage one as “responders” and second-stage participants as “non-responders.”

The data we use here consists of $n = 795$ patient trajectories with complete information and does not
include 481 patients who dropped out for reasons other than remission. The variables composing each
trajectory are given in Table 2.1.

Of the 795 total patients, 329 were non-responders in the first stage and continued on to the second
stage. We define our primary outcome as $Y = RY_1 + (1 - R)(Y_1 + Y_2)/2$. That is, $Y$ is taken to be
$Y_1$ for patients who left the study after stage one, and $Y$ is taken to be the average of the QIDS scores
measured at the end of the first and second stages for non-responders. Our responder-status version of
$IQ$-learning is based on the $Q$-learning implementation described in Schulte et al. (2014). The first-stage
history vector contains all information available prior to the first-stage treatment randomization. Thus,
$H_1 = (X_{1,1}, X_{1,2})^\top$. The second-stage history is $H_2 = (X_{1,1}, X_{1,2}, A_1, Y_1, R, X_{2,1}, X_{2,2})^\top$, which
contains all information observed before the second-stage treatment assignment. The second-stage $Q$-
function is

$$Q_2(H_2, A_2) = E(Y|H_2 = h_2, A_2 = a_2) = RY_1 + (1 - R)\frac{Y_1 + E(Y_2|H_2 = h_2, A_2 = a_2)}{2},$$

where we have used the fact that $R$ and $Y_1$ are contained in $H_2$. Thus the first step in the $IQ$-learning
algorithm, and in $Q$-learning, is to specify and fit a model for $E(Y_2|H_2 = h_2, A_2 = a_2)$. Defining
the second-stage summary vectors as $H_{2,0} = H_{2,1} = (1, X_{2,1}, X_{2,2})^\top$, we consider a working model of
the form $E(Y_2|H_2 = h_2, A_2 = a_2) = h_{2,0}^\top \beta_{2,0} + a_2 h_{2,1}^\top \beta_{2,1}$, that we fit via least squares. Standard
regression diagnostics based on the residuals do not indicate any major departures from the usual linear
modeling assumptions.

In our subset of data, all non-responders who received a non-SSRI in stage one received a non-SSRI
in stage two. Table 2.2 provides the number of patients assigned to each treatment strategy. We define
Table 2.1: Variables comprising patient trajectories in the STAR*D data analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{1,1} \in [0, 27]$</td>
<td>27 minus the baseline patient QIDS score.</td>
</tr>
<tr>
<td>$X_{1,2} \in \mathbb{R}$</td>
<td>pre-randomization slope of patient QIDS score, computed by taking the difference between the recoded QIDS score at study entry and the beginning of the first randomized stage. This difference is then divided by the time between study entry and first randomization. Negative values are associated with symptom improvement.</td>
</tr>
<tr>
<td>$A_1 \in {-1, 1}$</td>
<td>initial treatment, coded so that $A_1 = 1$ corresponds to “SSRI” and $A_1 = -1$ corresponds to “non-SSRI.”</td>
</tr>
<tr>
<td>$Y_{1} \in [0, 27]$</td>
<td>27 minus the patient QIDS score measured at the end of the first stage.</td>
</tr>
<tr>
<td>$R \in {0, 1}$</td>
<td>first-stage responder indicator. $R = 1$ indicates remission in stage one and exit from the study.</td>
</tr>
<tr>
<td>$X_{2,1} \in [0, 27]$</td>
<td>27 minus the patient QIDS score measured just prior to the second randomization.</td>
</tr>
<tr>
<td>$X_{2,2} \in \mathbb{R}$</td>
<td>first-stage slope of patient QIDS score, computed as the difference between the recoded QIDS scores measured at the beginning and end of the first randomized stage. This difference is then divided by the time spent in the first randomized stage. Negative values are associated with symptom improvement.</td>
</tr>
<tr>
<td>$A_2 \in {-1, 1}$</td>
<td>second stage treatment, coded so that $A_2 = 1$ corresponds to “SSRI” and $A_2 = -1$ corresponds to “non-SSRI.”</td>
</tr>
<tr>
<td>$Y_{2} \in [0, 27]$</td>
<td>second-stage outcome, defined as 27 minus the end of second-stage patient QIDS score.</td>
</tr>
</tbody>
</table>

\[ \bar{Y} \text{ as follows:} \]

\[ \bar{Y} = \arg \max_{a_2} Q_2(H_2, a_2)R + \left\{ Q_2(H_2, -1)(1 - A_1) + \arg \max_{a_2} Q_2(H_2, a_2)(1 + A_1) \right\} (1 - R)/2. \]

Thus, $\bar{Y}$ is $Q_2(H_2, -1)$ for non-responders who received $A_1 = -1$ and $\arg \max_{a_2} Q_2(H_2, a_2)$ otherwise. With this working model, after substituting in $Q_2(H_2, a_2)$ and simplifying,

\[ Q_1(H_1, A_1) = E(RY_1|H_1, A_1) + \frac{1}{2} E\{(1 - R)Y_1|H_1, A_1\} \]

\[ + \frac{1}{2} E\{(1 - R)H_{2,0}^T \beta_{2,0} | H_1, A_1\} - \frac{1 - A_1}{4} E\{(1 - R)H_{2,1}^T \beta_{2,1} | H_1, A_1\} \]

\[ + \frac{1 + A_1}{4} E\{(1 - R)H_{2,1}^T \beta_{2,1} | H_1, A_1\}. \]
Table 2.2: Number of patients per treatment strategy by responder status.

<table>
<thead>
<tr>
<th>Treatment Sequence</th>
<th>Responders</th>
<th>Non-responders</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SSRI, NA)</td>
<td>319</td>
<td>NA</td>
</tr>
<tr>
<td>(non-SSRI, NA)</td>
<td>147</td>
<td>NA</td>
</tr>
<tr>
<td>(SSRI, SSRI)</td>
<td>NA</td>
<td>70</td>
</tr>
<tr>
<td>(SSRI, non-SSRI)</td>
<td>NA</td>
<td>120</td>
</tr>
<tr>
<td>(non-SSRI, SSRI)</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>(non-SSRI, non-SSRI)</td>
<td>NA</td>
<td>139</td>
</tr>
</tbody>
</table>

Figure 2.9: **Left:** Scatterplot of $\tilde{Y}$ against baseline QIDS score by first-stage treatment. **Middle:** Scatterplot of the estimated main effect function, $H_{2,0}^T \beta_{2,0}$, against baseline QIDS score by first-stage treatment for non-responders. **Right:** Scatterplot of the estimated contrast function, $H_{2,1}^T \beta_{2,1}$, against baseline QIDS score by first-stage treatment for non-responders. Cubic smoothing spline fits to the data for $A_1 = 1$ and $A_1 = -1$ are represented by solid gray and dashed black lines.

The Q-learning algorithm models $E(\tilde{Y} | H_1, A_1)$ directly. The left plot in Figure (2.9) shows a scatterplot of the pseudo-response $\tilde{Y}$ against baseline QIDS by first-stage treatment. Cubic smoothing spline fits to the data are indicated by solid gray and dashed black lines for $A_1 = 1$ and $A_1 = -1$, respectively. In the leftmost plot of Figure (2.9), these fits appear approximately linear; however, the variance appears non-constant across baseline QIDS, and there is clear separation between the responder and non-responder groups.

Notice that we can write the first three expectation terms in (2.9) as the following:

$$E(RY_1 | H_1, A_1) = E(Y_1 | H_1, A_1, R = 1) \Pr(R = 1 | H_1, A_1),$$

$$E \{(1 - R)Y_1 | H_1, A_1\} = E(Y_1 | H_1, A_1, R = 0) \Pr(R = 0 | H_1, A_1),$$

$$E \{(1 - R)H_{2,0}^T \beta_{2,0} | H_1, A_1\} = E(H_{2,0}^T \beta_{2,0} | H_1, A_1, R = 0) \Pr(R = 0 | H_1, A_1).$$
We estimate the right-hand conditional expectations by fitting three separate linear regressions, and we use logistic regression to estimate \( \text{pr}(R = r | H_1, A_1) \), \( r = 0, 1 \). In particular, we specify linear models of the form

\[
E(Y_i | H_1 = h_1, A_1 = a_1, R = r) = h_{1,0}^{T} \lambda_{r,0} + a_1 h_{1,1}^{T} \lambda_{r,1}
\]

for \( E(Y_i | H_1, A_1, R = r) \), where \( H_{1,0} = H_{1,1} = (1, X_{1,1}, X_{1,2})^T \). We posit the model

\[
E \left( H_{2,0}^{T} \beta_{2,0} | H_1 = h_1, A_1 = a_1, R = 0 \right) = h_{1,0}^{T} \alpha_0 + a_1 h_{1,1}^{T} \alpha_1
\]

for the main-effect term, which we fit with least squares using only the non-responder data. The middle and right plots in Figure (2.9) display scatterplots of the non-responder realizations of the main-effect term and contrast function, respectively, against baseline QIDS by first-stage treatment. Cubic smoothing spline fits to the data appear mostly linear. For the logistic regression, we fit the model

\[
\logit \{ \text{pr}(R = 1 | H_1 = h_1, A_1 = a_1) \} = h_{1,0}^{T} \delta_0 + a_1 h_{1,1}^{T} \delta_1.
\]

Finally, we must obtain estimates of \( E \{ (1 - R) | H_{2,1}^{T} \beta_{2,1} \mid H_1, A_1 \} \) and \( E \{ (1 - R) H_{2,1}^{T} \beta_{2,1} \mid H_1, A_1 \} \) in equation (2.9). Notice that

\[
E \{ (1 - R) | H_{2,1}^{T} \beta_{2,1} \mid H_1, A_1 \} = \text{pr}(R = 0 | H_1, A_1) \int g(z \mid H_1, A_1, R = 0) dz.
\]

We can use the IQ-learning machinery to obtain an estimate of \( g(\cdot \mid H_1, A_1, R = 0) \). The logistic regression previously described provides an estimate of \( \text{pr}(R = 0 | H_1, A_1) \). To estimate \( g(\cdot \mid H_1, A_1, R = 0) \), we first specify a model for the mean of the contrast function for non-responders. We posit the model

\[
E \left\{ \Delta(H_2; \tilde{\beta}_2) | H_1, A_1 \right\} = H_{1,0}^{T} \beta_0 + A_1 H_{1,1}^{T} \beta_1,
\]

(2.10)

where \( H_{1,0} = (1, X_{1,1}, X_{1,2})^T \) and \( H_{1,1} = (1, X_{1,1}, X_{1,2})^T \). This model is also used along with the logistic regression model to obtain an estimate of \( E \{ (1 - R) H_{2,1}^{T} \beta_{2,1} \mid H_1, A_1 \} = E(H_{2,1}^{T} \beta_{2,1} \mid H_1, A_1, R = 0) \text{pr}(R = 0 \mid H_1, A_1) \), the fourth expectation term in (2.9). We fit model (2.10) using least squares. Exploratory analysis suggests that a constant variance assumption is reasonable, therefore, we standardize the residuals of the mean fit using the sample standard deviation. A normal quantile-quantile plot of the standardized residuals suggests heavier tails than would be expected from a normal distribution. Thus, we opt to use the nonparametric density estimator described in Section 2.2.1.
Assembling the foregoing estimates of the four terms in equation (2.9) yields

\[
Q_1^{IQ}(h_1, a_1; \hat{\theta}_1) = (h_{1,0}^\top \hat{\lambda}_{1,0} + a_1 h_{1,1}^\top \hat{\lambda}_{1,1}) \text{expit}(h_{1,0}^\top \hat{\delta}_0 + a_1 h_{1,1}^\top \hat{\delta}_1)
\]
\[
+ \left\{1 - \text{expit}(h_{1,0}^\top \hat{\delta}_0 + a_1 h_{1,1}^\top \hat{\delta}_1)\right\}\left\{\frac{1}{2}(h_{1,0}^\top \hat{\lambda}_{0,0} + a_1 h_{1,1}^\top \hat{\lambda}_{0,1}) + \frac{1}{2}(h_{1,0}^\top \hat{\lambda}_{0,0} + a_1 h_{1,1}^\top \hat{\lambda}_{1,1})
\right.
\]
\[
- \frac{1}{4}(h_{1,0}^\top \hat{\beta}_0 + a_1 h_{1,1}^\top \hat{\beta}_1) + \frac{1}{4} \int |z| \hat{g}_E^E(z \mid h_1, a_1, R = 0) dz \right\},
\]

where \(\hat{g}_E^E(\cdot \mid h_1, a_1, R = 0)\) is the nonparametric density estimator described in Section 2.2.1 and
\[
\hat{\theta}_1 = (\hat{\lambda}_{0,0}^\top, \hat{\lambda}_{0,1}^\top, \hat{\lambda}_{1,0}^\top, \hat{\lambda}_{1,1}^\top, \hat{\delta}_0^\top, \hat{\delta}_1^\top, \hat{\alpha}_0^\top, \hat{\alpha}_1^\top, \hat{\beta}_0^\top, \hat{\beta}_1^\top)^\top.
\]

The first-stage rule estimated by \(Q\)-learning treats all training data patients with a SSRI. Roughly twelve percent of these patients are recommended a non-SSRI by the estimated first-stage \(IQ\)-learning rule. Broadly summarizing the \(IQ\)-learning rule, patients with very low recoded QIDS scores after the non-randomized stage should switch from Citalopram, an SSRI, to a non-SSRI. This rule is intuitive because it recommends a different strategy for patients who respond poorly to the initial SSRI; otherwise SSRI treatment strategies should continue.
Chapter 3

Interactive $Q$-learning for Probabilities and Quantiles

3.1 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 an estimated optimal regime. Current methods for estimating an optimal dynamic treatment regime 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., 2012), Interactive $Q$-learning (Laber et al., 2013b), $g$-estimation (Robins, 2004), $A$-learning (Murphy, 2003), and regret-regression (Henderson et al., 2010). Goldberg and Kosorok (2012) propose a version of $Q$-learning for censored data, and policy search methods optimize an estimator of the mean population outcome over a class of possible treatment regimes (Orellana et al., 2010; Zhao et al., 2012; Zhang et al., 2012b,a, 2013). Despite many existing estimation methods, none are specifically formulated to handle other functionals of the continuous response distribution, such as quantiles or expectations of indicator functions involving the response variable.

The $Q$-learning algorithm is an approximate dynamic programming procedure that requires modeling nonsmooth, nonmonotone transformations of data, resulting in nonregular estimators and challenges with regard to model building (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2010; Song et al., 2011). Interactive $Q$-learning ($IQ$-learning) was developed for the two-stage, binary treatment setting. $IQ$-learning requires modeling only smooth, monotone transformations of the data, reducing problems of model misspecification and nonregular inference (Laber et al., 2013b).
We extend the \( I\Omega \)-learning framework to handle optimization of more general features of the outcome distribution 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 on 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 pre-specified, 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, it provides an analog of quantile regression in the dynamic treatment regime setting for constructing robust estimators; for example, it enables optimization of the median response.

### 3.2 Generalized Interactive \( Q \)-learning

#### 3.2.1 Two-stage Binary Treatment Models

We assume data have been collected using a two-stage sequential multiple assignment randomized trial (SMART; 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. See the website of The Methodology Center at Pennsylvania State University; Director: Dr. Linda Collins (2012, http://methodology.psu.edu/ra/adap-inter/projects) or Laber (2013) for examples. The ideas presented here also apply to observational studies under certain causal conditions (Robins, 2004; Schulte et al., 2014). In the two-stage binary treatment setting, the training set \( D = \{(X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_i)\}_{i=1}^{n} \) comprises time-ordered trajectories that are independent and identically distributed across \( n \) patients. Let \((X_1, A_1, X_2, A_2, Y)\) represent observations for a single subject where: \( X_1 \in \mathbb{R}^{p_1} \) denotes baseline, pre-randomization covariate information; \( A_1 \in \{-1, 1\} \) denotes the first randomized treatment; \( X_2 \in \mathbb{R}^{p_2} \) denotes 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. A regime \( \pi \) is a pair of decision rules \( (\pi_1, \pi_2) \), where \( \pi_1 \) is a map from the support of \( X_1 \) to the space of possible first-stage treatments, and \( \pi_2 \) is a map from the support of \((X_1^T, A_1, X_2^T)^T\) to the space of possible second-stage treatments. That is, \( \pi_1 : \text{dom}(X_1) \rightarrow \mathcal{A}_1 = \{-1, 1\} \), and \( \pi_2 : \text{dom}\{(X_1^T, A_1, X_2^T)^T\} \rightarrow \mathcal{A}_2 \).
\[ A_2 = \{-1, 1\}. \] Because \( A_2 \) is binary, there exist functions \( m \) and \( c \) such that

\[
E(Y \mid A_2, X_2, A_1, X_1) = m(X_1, A_1, X_2) + A_2c(X_1, A_1, X_2).
\]

We assume a mean-zero additive error term \( \epsilon \) so that \( Y = E(Y \mid A_2, X_2, A_1, X_1) + \epsilon \); additionally, assume \( \text{var}(\epsilon) < \infty \) and \( \epsilon \) is independent of \((X_1, A_1, X_2, A_2)\). Extension to the heteroskedastic setting where the variance of \( Y \) depends on past covariates and treatments is provided in Appendix B.

### 3.2.2 Threshold Interactive Q-learning

Let \( \text{pr}^\pi(Y > \lambda) \) denote the probability that the outcome \( Y \) is greater than a pre-defined threshold \( \lambda \) under treatment assignment dictated by the regime \( \pi = (\pi_1, \pi_2) \). Threshold Interactive Q-learning (TIQ-learning) maximizes \( \text{pr}^\pi\{Y > \lambda(\tilde{X}) \mid X_1\} \) for all \( X_1 \) with respect to \( \pi \), where \( \lambda(\tilde{X}) \) is a threshold that depends on any subset of \( \tilde{X} = (X_1^T, A_1, X_2^T, A_2)^T \). Here we consider a constant threshold, \( \lambda(\tilde{X}) = \lambda \); the incorporation of patient-specific thresholds is discussed in Appendix B.

Define the following: \( F_{X_1}(\cdot) \) is the cumulative distribution function of vector \( X_1 \), \( F_{X_2 \mid x_1, a_1} (\cdot \mid x_1, a_1) \) is the conditional cumulative distribution function of vector \( X_2 \) given \( X_1 = x_1 \) and \( A_1 = a_1 \), and \( F_\epsilon (\cdot) \) is the cumulative distribution function of \( \epsilon \). Define

\[
J^{\pi_1, \pi_2}(x_1, x_2, y) = F_\epsilon[y - m\{x_1, \pi_1(x_1), x_2\} - \pi_2\{x_1, \pi_1(x_1), x_2\}c\{x_1, \pi_1(x_1), x_2\}] \]

and

\[
\text{pr}^{\pi_1, \pi_2}(Y \leq y) = \int \int J^{\pi_1, \pi_2}(x_1, x_2, y)dF_{X_2 \mid x_1, a_1}(x_2 \mid x_1, \pi_1(x_1))dF_{X_1}(x_1). \tag{3.1}
\]

In addition, define \( \pi_2(x_1, a_1, x_2) = \text{sgn}\{c(x_1, a_1, x_2)\} \), where \( \text{sgn}(x) = \mathbb{1}_{x \geq 0} - \mathbb{1}_{x < 0} \). Then, because \( \pi_2\{x_1, \pi_1(x_1), x_2\} \) takes values \(-1\) and \(1\),

\[
\pi_2\{x_1, \pi_1(x_1), x_2\}c\{x_1, \pi_1(x_1), x_2\} \leq |c\{x_1, \pi_1(x_1), x_2\}|
\]

for all \( (x_1, x_2) \). This implies

\[
\text{pr}^{\pi_1, \pi_2}(Y \leq y) \leq \int \int J^{\pi_1, \pi_2^2}(x_1, x_2, y)dF_{X_2 \mid x_1, a_1}(x_2 \mid x_1, \pi_1(x_1))dF_{X_1}(x_1), \tag{3.2}
\]

where \( J^{\pi_1, \pi_2^2}(x_1, x_2, y) = F_\epsilon[y - m\{x_1, \pi_1(x_1), x_2\} - |c\{x_1, \pi_1(x_1), x_2\}|] \). Denote the right-hand side of (3.2) by \( \text{pr}^{\pi_1, \pi_2^2}(Y \leq y) \). Let \( G(\cdot, \cdot \mid x_1, a_1) \) denote the joint conditional distribution of \( \{m(X_1, A_1, X_2), c(X_1, A_1, X_2)\} \) given \( X_1 = x_1 \) and \( A_1 = a_1 \). That is,

\[
G(u, v \mid x_1, a_1) = \text{pr}\{m(X_1, A_1, X_2) \leq u, c(X_1, A_1, X_2) \leq v \mid X_1 = x_1, A_1 = a_1\}. \]
Using the fact that

\[ G(u, v \mid x_1, a_1) = \int \mathbb{I}_{m(x_1, a_1, x_2) \leq u} \mathbb{I}_{c(x_1, a_1, x_2) \leq v} dF_{\cdot}(x_2 \mid x_1, A_1(x_2 \mid x_1, a_1)), \]

one can show by justifying an interchange in the order of integration that

\[
\int \int F_\epsilon(y - u - |v|)dG\{u, v \mid x_1, \pi_1(x_1)\}dF_{\cdot}(x_1) = \int \int J_{\pi_1, \pi_2}^2(x_1, x_2, y)dF_{\cdot}(x_2 \mid x_1, A_1(x_2 \mid x_1, \pi_1(x_1)))dF_{\cdot}(x_1),
\]

where the right-hand side is \( pr_{\pi_1, \pi_2}^2(Y \leq y) \) by definition. Define

\[ I\{y, F_\epsilon(\cdot), G(\cdot, \cdot) \mid x_1, a_1\} = \int F_\epsilon(y - u - |v|)dG\{u, v \mid x_1, a_1\}; \quad (3.3) \]

it follows that

\[ pr_{\pi_1, \pi_2}^2(Y \leq y) = E[I\{y, F_\epsilon(\cdot), G(\cdot, \cdot) \mid X_1, \pi_1(X_1)\}]. \quad (3.4) \]

The \( \lambda \)-optimal regime \( \pi_{\lambda}^{\text{TQ}} = \{\pi_{1, \lambda}^{\text{TQ}}, \pi_{2, \lambda}^{\text{TQ}}\} \) is determined by the requirement that \( pr_{\pi_{1, \lambda}^{\text{TQ}}}(Y > \lambda) \geq pr_{\pi}(Y > \lambda) \) for all \( \pi \); or equivalently, that \( pr_{\pi_{1, \lambda}^{\text{TQ}}}(Y \leq \lambda) \leq pr_{\pi}(Y \leq \lambda) \) for all \( \pi \). That is, the distribution of \( Y \) induced by regime \( \pi_{\lambda}^{\text{TQ}} \) 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 \( pr_{\pi_1, \pi_2}^2(Y \leq y) \) displayed in (3.2) that

\[ \pi_{2, \lambda}^{\text{TQ}}(x_1, a_1, x_2) = \pi_{2}^\ast(x_1, a_1, x_2) = \text{sgn}\{c(x_1, a_1, x_2)\} \]

for all \( (x_1, a_1, x_2) \), independent of \( \lambda \) and \( \pi_{1, \lambda}^{\text{TQ}} \). Henceforth, we denote \( \pi_{2, \lambda}^{\text{TQ}} \) by \( \pi_{2}^\ast \). It follows from (3.4) that

\[
pr_{\pi_1, \pi_2}^\ast(Y > \lambda) = 1 - E[I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot) \mid X_1, \pi_1(X_1)\}] \leq 1 - E\left[ \min_{a_1} I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot) \mid X_1, a_1\} \right], \quad (3.5)
\]

thus showing that the \( \lambda \)-optimal first-stage rule is

\[
\pi_{1, \lambda}^{\text{TQ}}(x_1) = \arg\min_{a_1} I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot) \mid x_1, a_1\} = \text{sgn}\{d(x_1, \lambda)\}, \quad (3.6)
\]

where

\[ d(x_1, \lambda) = I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot) \mid x_1, -1\} - I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot) \mid x_1, 1\}. \quad (3.7) \]
Thus, if \( a_1 = -1 \) minimizes \( I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot \mid x_1, a_1)\} \), \( d(x_1, \lambda) \) will be negative, and if \( a_1 = 1 \) minimizes \( I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot \mid x_1, a_1)\} \), \( d(x_1, \lambda) \) will be positive. Inequality (3.5) holds because for all \( \mathbf{X}_1, I\{\lambda, F_\epsilon(\cdot), G(\cdot, \cdot \mid \mathbf{X}_1, a_1)\} \) is minimized over \( a_1 \).

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 \( m(\mathbf{H}_2; \beta_{2,0}) \) and \( c(\mathbf{H}_2; \beta_{2,1}) \), respectively, where \( \mathbf{H}_2 \) is shorthand for the set of random variables \( \{\mathbf{X}_1, A_1, \mathbf{X}_2\} \), which is the full history available prior to the second-stage treatment assignment. The exact algorithm depends on the choice of estimators \( \widehat{F}_\epsilon(\cdot) \) and \( \widehat{G}(\cdot, \cdot \mid x_1, a_1) \) for \( F_\epsilon(\cdot) \) and \( G(\cdot, \cdot \mid x_1, a_1) \). We discuss possible estimators in Sections 3.2.4 and 3.2.5; in practice the choice of estimators should be informed by the observed data. Let \( \widehat{d}(x_1, \lambda) = I\{\lambda, \widehat{F}_\epsilon(\cdot), \widehat{G}(\cdot, \cdot \mid x_1, -1)\} - I\{\lambda, \widehat{F}_\epsilon(\cdot), \widehat{G}(\cdot, \cdot \mid x_1, 1)\} \) denote an estimator of (3.7).

**The Threshold Interactive Q-learning Algorithm**

**TIQ1.** Estimate \( m(\mathbf{H}_2; \beta_{2,0}) \) and \( c(\mathbf{H}_2; \beta_{2,1}) \) by fitting the model

\[
Y = m(\mathbf{H}_2; \beta_{2,0}) + A_2 c(\mathbf{H}_2; \beta_{2,1}) + \epsilon,
\]

e.g., using least squares. The estimated optimal second-stage rule is \( \widehat{\pi}_2^\ast(h_2) = \text{sgn}\{c(h_2; \widehat{\beta}_{2,1})\} \).

**TIQ2.** Estimate \( F_\epsilon(\cdot) \), the distribution function of \( \epsilon \), using the residuals

\[
\widehat{\epsilon}^\ast = Y - m(\mathbf{H}_2; \widehat{\beta}_{2,0}) - A_2 c(\mathbf{H}_2; \widehat{\beta}_{2,1})
\]

from Step TIQ1. Let \( \widehat{F}_\epsilon(\cdot) \) denote this estimator.

**TIQ3.** Estimate \( G(\cdot, \cdot \mid x_1, a_1) \), the joint conditional distribution of \( m(\mathbf{H}_2) \) and \( c(\mathbf{H}_2) \) given \( \mathbf{X}_1 = x_1 \) and \( A_1 = a_1 \). Let \( \widehat{G}(\cdot, \cdot \mid x_1, a_1) \) denote this estimator. In practice, one would use estimated versions of \( m(\mathbf{H}_2) \) and \( c(\mathbf{H}_2) \), i.e., \( m(\mathbf{H}_2; \widehat{\beta}_{2,0}) \) and \( c(\mathbf{H}_2; \widehat{\beta}_{2,1}) \) to obtain \( \widehat{G}(\cdot, \cdot \mid x_1, a_1) \).

**TIQ4.** For a given \( x_1 \), estimate \( \pi_{1,\lambda}^{\text{TIQ}}(x_1) \) from (3.6) using the plug-in estimator

\[
\pi_{1,\lambda}^{\text{TIQ}}(x_1) = \text{sgn}\{\widehat{d}(x_1, \lambda)\}.
\]

TIQ-learning requires more modeling than mean-optimal techniques such as Q-learning or IQ-learning. The Q-learning algorithm involves modeling \( m(\cdot), c(\cdot) \), and the conditional expectation of \( m(\mathbf{H}_2) + |c(\mathbf{H}_2)| \) given \( (\mathbf{H}_1, A_1) \). IQ-learning requires modeling \( m(\cdot), c(\cdot) \), the conditional expectation of \( m(\mathbf{H}_2) \) given \( (\mathbf{H}_1, A_1) \), and the one-dimensional conditional density of \( c(\mathbf{H}_2) \) given \( (\mathbf{H}_1, A_1) \) (see Chapter 2 or Laber et al., 2013b). In comparison, the TIQ-learning algorithm involves modeling
\( m, c \), the distribution function \( F_c(\cdot) \), and the bivariate conditional density \( G(\cdot, \cdot \mid x_1, a_1) \), which we discuss in Sections 3.2.4 and 3.2.5.

**Remark 3.2.1.** Let \( \pi^M_1 \) denote the first-stage decision rule of a mean optimal regime. Then, assuming the setup of section 3.2.1, it can be shown \( \pi^M_1(x_1) = \arg \min_{a_1} \int (-u - |v|)dG(u, v \mid x_1, a_1) \). Thus, the difference between the mean optimal and threshold probability optimal regimes can be compared empirically by computing

\[
\arg \min_{a_1} \int (-u - |v|)d\tilde{G}(u, v \mid x_{1i}, a_1), \quad \arg \min_{a_1} \int \tilde{F}_e(\lambda - u - |v|)d\tilde{G}(u, v \mid x_{1i}, a_1),
\]

for \( i = 1, \ldots, n \) and considering the proportion of patients recommended to different treatments.

### 3.2.3 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 the lower quantiles of the induced distribution of \( Y \). We demonstrate this using simulated experiments in Section 3.3. Define the \( \tau^\text{th} \) quantile of the distribution of \( Y \) induced by regime \( \pi \) as \( q^\pi(\tau) = \inf\{y : \text{pr}^\pi(Y \leq y) \geq \tau\} \).

The goal of Quantile Interactive Q-learning (QIQ-learning) is to estimate a pair of decision rules, \( \pi^{\text{QIQ}}_\tau = \{\pi^{\text{QIQ}}_{1,\tau}, \pi^{\text{QIQ}}_{2,\tau}\} \), that maximize \( q^\pi(\tau) \) over \( \pi \) for a fixed, pre-specified \( \tau \). 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 3.2.1, the QIQ-learning second-stage optimal decision is the same as that of TIQ-learning, i.e., \( \pi^{\text{QIQ}}_{2,\tau}(h_2) = \pi^{\tau}_{2}(h_2) = \text{sgn}\{c(h_2)\} \), independent of \( \tau \) and \( \pi^{\text{QIQ}}_{1,\tau} \); see Appendix B for a derivation of this rule. We henceforth denote \( \pi^{\text{QIQ}}_{0,\tau} \) by \( \pi^\tau_2 \).

Next we discuss the form of the optimal first-stage decision rule, \( \pi^{\text{QIQ}}_{1,\tau} \), which in turn motivates a computational algorithm for calculating it. Define

\[
\Gamma(x_1, y) = \text{sgn}\{I\{y, F_c(\cdot), G(\cdot, \cdot \mid x_1, -1) - I\{y, F_c(\cdot), G(\cdot, \cdot \mid x_1, 1)\}\}, \quad (3.8)
\]

\[
y^*_\tau = \inf\{y : \text{pr}^\Gamma(\cdot, y), \pi^\tau_2(Y \leq y) \geq \tau\}, \quad (3.9)
\]

\[
f(y) = q^\Gamma(\cdot, \cdot), \pi^\tau_2(\tau) = \inf\{y : \text{pr}^\Gamma(\cdot, y), \pi^\tau_2(Y \leq y) \geq \tau\}, \quad (3.10)
\]

where \( I(\cdot, \cdot, \cdot) \) is defined in (3.3). Note that \( \Gamma(x_1, y) = \text{sgn}\{d(x_1, y)\} = \pi^{\text{QIQ}}_{1,\tau}(x_1) \mid \lambda = y \), and thus it is a first-stage decision rule. We have introduced the new notation to emphasize the dependence on \( y \).

We show in Lemma 3.2.11 of Section 3.2.7 that \( \lim_{y \to \infty(-\infty)} \text{pr}^\Gamma(\cdot, y), \pi^\tau_2(Y \leq y) = 1(0) \), so that \( y^*_\tau \) is defined for all \( \tau \in (0, 1) \). Furthermore, \( y^*_\tau \) is an upper bound on \( q^\pi_1, \pi^\tau_2(\tau) \) for all \( \pi_1 \) because for
Lemma 3.2.2. below.

The last equality follows because \( \Gamma(\mathbf{X}_1, y) \) minimizes \( E(I[y, F_\epsilon(\cdot), G\{\cdot, \cdot | \mathbf{X}_1, \pi_1(\mathbf{X}_1)\}]) \) with respect to \( a_1 \). Hence,

\[
\left\{ y : \text{pr}^{\pi_1, \pi_2}(Y \leq y) \geq \tau \right\} \subseteq \left\{ y : \text{pr}^{\pi_1, \pi_2}(Y \leq y) \geq \tau \right\},
\]

and taking the infimum on both sides gives the upper bound

\[
y^*_\tau \geq q^{\pi_1, \pi_2}(\tau) \tag{3.11}
\]

for all \( \pi_1 \). 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 optimally according to \( \pi_2^* \).

Note that \( f(y) \) in (3.10) is the quantile obtained under regime \( \pi = \{\Gamma(\cdot, y), \pi_2^*\} \). Thus, because it is a quantile and the bound in (3.11) applies,

\[
\text{pr}^{\Gamma(\cdot, y), \pi_2^*}(Y \leq f(y)) \geq \tau, \quad f(y) = q^{\Gamma(\cdot, y), \pi_2^*}(\tau) \leq y^*_\tau \quad \text{for all } y. \tag{3.12}
\]

Our main results depend on the following lemma, proved below.

**Lemma 3.2.2.**

(A) \( y < y^*_\tau \) implies \( y < f(y) \leq y^*_\tau \); \tag{3.13}

(B) \( f(y^*_\tau^-) = \lim_{\delta \downarrow 0} f(y^*_\tau - \delta) = y^*_\tau \); \tag{3.14}

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

\[
\text{pr}^{\Gamma(\cdot, y^*_\tau), \pi_2^*}(Y \leq y^*_\tau - \delta) \geq \tau; \tag{3.15}
\]

(D) If \( F_\epsilon(\cdot) \) is continuous and strictly increasing, then \( f(y^*_\tau) = y^*_\tau \). \tag{3.16}

It follows that \( f(y^*_\tau) = y^*_\tau \) if and only if \( f(y) \) is left continuous at \( y = y^*_\tau \). In this case, the optimal first-stage rule is \( \pi_{1,\tau}^{\text{opt}}(x_1) = \Gamma(x_1, y^*_\tau) = \text{sgn}\{d(x_1, y^*_\tau)\} \). Part (D) is a sufficient condition guaranteeing left-continuity of \( f(y) \) at \( y^*_\tau \). 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 (3.14) 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 following the proof of Lemma 3.2.2 below.
Proof. We showed in (3.12) 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 (3.12) 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 (3.9), \( y_{\tau}^* \) is the smallest \( y \) satisfying \( \text{pr}^{\Gamma(\cdot, y)}, \pi_2^*(Y \leq y) \geq \tau \).

Using (3.12) 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 (3.12) 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 \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq y_{\tau}^* - \delta) \geq \tau \), 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 \( F_{\tau}(\cdot) \) is continuous and strictly increasing, \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq y) \) is also continuous and monotone increasing because it is an expectation of a continuous, strictly increasing function of \( y \). It can be shown that for any fixed regime \( \pi \), \( \text{pr}^{\pi}(Y \leq y) \) continuous in \( y \) implies \( \text{pr}^{\Gamma(\cdot, y)}, \pi_2^*(Y \leq y) \) is also continuous. Suppose toward a contradiction that \( f(y_{\tau}^*) < y_{\tau}^* \). When \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq y) \) is continuous and strictly increasing, the Mean Value Theorem guarantees existence of at least one point \( \tilde{y} \in \mathbb{R} \) such that \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq \tilde{y}) = \tau \). By definition, \( f(y_{\tau}^*) \) must be one such point, and thus \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq f(y_{\tau}^*)) = \tau \). The assumption \( f(y_{\tau}^*) < y_{\tau}^* \) implies \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq y_{\tau}^*) > \tau \). However, when \( \text{pr}^{\Gamma(\cdot, y)}, \pi_2^*(Y \leq y) \) is continuous, \( \text{pr}^{\Gamma(\cdot, y_{\tau}^*)}, \pi_2^*(Y \leq y_{\tau}^*) = \tau \) by the Mean Value Theorem and by the definition of \( y_{\tau}^* \). Thus, we have a contradiction and conclude that (D) holds.

If the underlying distributions of the histories and \( Y \) were known, the algorithm below produces an optimal regime.

**Population-level algorithm to find** \( \pi_{1,\tau}^{\text{qo}} \)

*Step 1.* Compute \( y_{\tau}^* \) from (3.9) and \( f(y_{\tau}^*) \) from (3.10).

*Step 2.*

a. If \( f(y_{\tau}^*) = y_{\tau}^* \), \( \pi_{1,\tau}^{\text{qo}}(x_1) = \Gamma(x_1, y_{\tau}^*) \) is optimal because it attains the optimal quantile \( y_{\tau}^* \).

b. If \( f(y_{\tau}^*) < y_{\tau}^* \), \( \pi_{1,\tau}^{\text{qo}}(x_1) = \lim_{\delta \to 0} \Gamma(x_1, y_{\tau}^* - \delta) \) is optimal.

In practice the generative model is not known, but the population-level algorithm suggests an estimator of \( \pi_{1,\tau}^{\text{qo}} \). 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 \( F_{\tau}(\cdot) \) and \( G(\cdot, \cdot | x_1, a_1) \); several options are presented in Sections 3.2.4 and 3.2.5; in practice, the choice should be data-driven.
The Quantile Interactive Q-learning algorithm

QIQ1. Follow Steps TIQ1–TIQ3 of the Threshold Interactive Q-learning algorithm in Section 3.2.2.

QIQ2. Let $x_{1i}$ denote patient $i$’s first-stage history. With $I(\cdot, \cdot, \cdot)$ as in (3.3), estimate $q^*_r$ using

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

QIQ3. Estimate $f(y^*_r)$ using

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

QIQ4. a. If $\hat{f}(\hat{y}^*_r) = \hat{y}^*_r$, then $\hat{\pi}^{Q\Omega}_{1,r}(x_1) = \hat{\Gamma}(x_1, \hat{y}^*_r)$ is an estimated optimal first-stage decision rule because it attains the estimated optimal quantile, $\hat{y}^*_r$.

b. If $\hat{f}(\hat{y}^*_r) < \hat{y}^*_r$, define $\hat{y}^{(0)} = \hat{f}(\hat{y}^*_r)$ and iterate $\hat{y}^{(k)} = \hat{f}(\hat{y}^{(k-1)})$ for $k = 1, 2, \ldots, K$ until either $\hat{y}^{(K)} = \hat{y}^*_r$ or $\hat{y}^{(K)} > \hat{y}^*_r - \delta$ for some tolerance $\delta > 0$. Then, $\hat{\pi}^{Q\Omega}_{1,r}(x_1) = \hat{\Gamma}(x_1, \hat{y}^{(K-1)})$ is the estimated optimal first-stage rule because it attains the estimated optimal quantile, $\hat{y}^*_r$, or attains a quantile that is within $\delta$ distance of $\hat{y}^*_r$.

In finite samples, the iterative procedure in QIQ4.b is guaranteed to achieve the estimated optimal quantile exactly in a finite number of iterations; however, the computational time of the iterative scheme may be avoided by specifying a tolerance $\delta > 0$ and using $\hat{\pi}^{Q\Omega}_{1,r}(x_1) = \hat{\Gamma}(x_1, \hat{y}^*_r - \delta)$ as the estimated optimal rule. Treating according to $\hat{\Gamma}(x_1, \hat{y}^*_r - \delta)$ results in the estimated quantile $\hat{f}(\hat{y}^*_r - \delta) > \hat{y}^*_r - \delta$.

By choosing $\delta$ arbitrarily small, this estimated quantile will be arbitrarily close to the estimated optimal quantile $\hat{y}^*_r$.

To complete the $TIQ$- and $QIQ$-learning algorithms, we provide specific estimators for $F_t(\cdot)$ and $G(\cdot, \cdot | x_1, a_1)$ 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.

3.2.4 Working models for $F_t(\cdot)$

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 $F_{\epsilon}(\cdot)$ is given by $\hat{F}_{\epsilon}^{N}(z) = \Phi(z/\hat{\sigma}_{\epsilon})$, where $\hat{\sigma}_{\epsilon}$ is the estimated standard deviation of the second-stage residuals, $\hat{\sigma}_{\epsilon}^{2} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{e_{i}^{\epsilon} \leq z}$, can be used. 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
For more flexibility, the empirical distribution of the residuals, $\hat{F}_{\epsilon}^{E}(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{e_{i}^{\epsilon} \leq z}$, can be used. Standard residual diagnostic techniques may be applied to determine whether a normal assumption seems reasonable for the data.

3.2.5 Working models for $G(\cdot, \cdot | x_{1}, a_{1})$
In addition to modeling $F_{\epsilon}(\cdot)$, TIQ- and QIQ-learning require modeling the bivariate conditional density of $m(H_{2})$ and $c(H_{2})$ given $(X_{1}, A_{1})$. A useful strategy for modeling this bivariate conditional density 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

$$e_{m} = \frac{m(H_{2}) - \mu_{m}(X_{1}, A_{1})}{\sigma_{m}(X_{1}, A_{1})}, \quad e_{c} = \frac{c(H_{2}) - \mu_{c}(X_{1}, A_{1})}{\sigma_{c}(X_{1}, A_{1})},$$

where

$$\mu_{m}(X_{1}, A_{1}) = E\{m(H_{2}) | X_{1}, A_{1}\}, \quad \sigma_{m}^{2}(X_{1}, A_{1}) = E[(m(H_{2}) - \mu_{m}(X_{1}, A_{1}))^{2} | X_{1}, A_{1}].$$

The mean and variance functions of $c(H_{2})$ are defined similarly:

$$\mu_{c}(X_{1}, A_{1}) = E\{c(H_{2}) | X_{1}, A_{1}\}, \quad \sigma_{c}^{2}(X_{1}, A_{1}) = E[(c(H_{2}) - \mu_{c}(X_{1}, A_{1}))^{2} | X_{1}, A_{1}].$$

In simulations, we use parametric mean and variance models for $\mu_{m}(X_{1}, A_{1})$, $\sigma_{m}^{2}(X_{1}, A_{1})$, $\mu_{c}(X_{1}, A_{1})$, and $\sigma_{c}^{2}(X_{1}, A_{1})$, 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 $G(\cdot, \cdot | x_{1}, a_{1})$. Using mean and variance modeling, the following steps would be substituted in Step 3 of the TIQ- and QIQ-learning algorithms. The steps below constitute one of the
simplest estimation algorithms for estimating the mean and variance functions; alternatively, established methods discussed in Carroll and Ruppert (1988) may be substituted for the following algorithm.

**Mean and variance modeling**

Step 3.1 Compute

$$\hat{\theta}_m = \arg \min_{\theta_m} \sum_{i=1}^{n} \left\{ m(H_{2i}; \hat{\beta}_{2,0}) - \mu_m(X_{1i}, A_{1i}; \theta_m) \right\}^2$$

and the resulting estimator $\mu_m(X_1, A_1; \hat{\theta}_m)$ of the mean function $\mu_m(X, A)$.

Step 3.2 Use the estimated mean function from Step 3.1 to obtain

$$\hat{\gamma}_m = \arg \min_{\gamma_m} \sum_{i=1}^{n} \left\{ m(H_{2i}; \hat{\beta}_{2,0}) - \mu_m(X_{1i}, A_{1i}; \hat{\theta}_m) \right\}^2 - \sigma_m^2(X_{1i}, A_{1i}; \gamma_m),$$

so that $\sigma_m^2(X_1, A_1; \hat{\gamma}_m)$ is an estimator of the variance function $\sigma_m^2(X_1, A_1)$. In practice, we may specify a log-linear model for the natural log of the squared residuals from the mean fit in Step 1. Alternatively, a common variance across histories $X_1$ and treatment $A_1$ or within each level of $A_1$ could be estimated.

Step 3.3 Repeat Steps 1 and 2 to obtain estimators of $\mu_c(X_1, A_1; \theta_c)$ and $\sigma_c(X_1, A_1; \gamma_c)$.

Step 3.4 Plug in the estimated parametric mean and variance functions to obtain $\hat{e}_i^m$ and $\hat{e}_i^c$, $i = 1, ..., n$, defined as

$$\hat{e}_i^m = \frac{m(H_{2i}; \hat{\beta}_{2,0}) - \mu_m(X_{1i}, A_{1i}; \hat{\theta}_m)}{\sigma_m(X_{1i}, A_{1i}; \hat{\gamma}_m)}, \quad \hat{e}_i^c = \frac{c(H_{2i}; \hat{\beta}_{2,1}) - \mu_c(X_{1i}, A_{1i}; \hat{\theta}_c)}{\sigma_c(X_{1i}, A_{1i}; \hat{\gamma}_c)}.$$

Then, $\hat{e}_i^m$ and $\hat{e}_i^c$, $i = 1, ..., n$, can be used to estimate the joint distribution of the standardized residuals. Samples drawn from this distribution can be transformed back to samples from $G(\cdot | x_1, a_1)$ to estimate the integral $I\{y, \hat{F}_c(\cdot), \hat{G}(\cdot | x_1, a_1)\}$ with a Monte Carlo average.

**3.2.6 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 pointwise in probability to $m(h_2)$ and $c(h_2)$, respectively, for each $h_2$.

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

39
A3. For each fixed \( x_1, a_1 \), \( \int |d\hat{G}(u, v \mid x_1, a_1) - dG(u, v \mid x_1, a_1)| \) converges to zero in probability.
A4. For each fixed \( a_1 \), \( n^{-1}\sum_{i=1}^{n} |d\hat{G}(u, v \mid X_{1i}, a_1) - dG(u, v \mid X_{1i}, a_1)| \) converges to zero in probability.
A5. \( \Pr\{|d(X_1, y^*_\tau)| = 0\} = 0 \).

In the simulation studies in Section 3.3 and data example in Section 3.4, 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 \( x_1 \) and \( a_1 \), \( dG(\cdot, \cdot \mid x_1, a_1) \) is a density and \( d\hat{G}(\cdot, \cdot \mid x_1, a_1) \) 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 3.2.3.** (Consistency of Threshold Interactive Q-learning) Assume A1–A3 and fix \( \lambda \in \mathbb{R} \). Then, \( \Pr\hat{\lambda}(Y > \lambda) \) converges in probability to \( \Pr\lambda(0)(Y > \lambda) \), where \( \hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2) \).

**Theorem 3.2.4.** (Consistency of Quantile Interactive Q-learning) Assume A1–A5. Then, \( \hat{\tau}(\cdot) \) converges in probability to \( y^*_\tau \) for any fixed \( \tau \), where \( \hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2) \).

### 3.2.7 Proofs of Theoretical Results

Let \( D = \{X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_i\}_{i=1}^{n} \) denote the observed data, which are \( n \) independent and identically distributed realizations of the trajectory \((X_{1}, A_{1}, X_{2}, A_{2}, Y)^\top\). Let \((X_{1}^\tau, A_{1}, X_{2}^\tau, A_{2}, Y)^\top\) be a trajectory that is independent of \( D \) but identically distributed. Let \( H_2 = (X_{1}^\tau, A_{1}, X_{2}^\tau)^\top \) denote the full patient history available prior to second-stage treatment, and \( h_2 \) an observed realization of this vector. Using the set-up and assumptions in Section 3.2.1, the optimal and estimated optimal second-stage rules 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 3.2.2:

\[
\begin{align*}
  d(x_1, y) &= \int F\epsilon(y - u - |v|)dG(u, v \mid x_1, -1) - \int F\epsilon(y - u - |v|)dG(u, v \mid x_1, 1), \\
  \hat{d}(x_1, y) &= \int \hat{F}\epsilon(y - u - |v|)\hat{d}G(u, v \mid x_1, -1) - \int \hat{F}\epsilon(y - u - |v|)\hat{d}G(u, v \mid x_1, 1).
\end{align*}
\]

With this notation, the optimal and estimated optimal first-stage rules for Threshold Interactive Q-learning are \( \pi_{1,\lambda}^{\text{TQ}}(x_1) = \text{sgn}\{d(x_1, \lambda)\} \) and \( \hat{\pi}_{1,\lambda}^{\text{TQ}}(x_1) = \text{sgn}\{\hat{d}(x_1, \lambda)\} \). We define \( \text{sgn}(0) = 1 \).

The following Lemmas are useful for the proofs of Theorems 3.2.3 and 3.2.4. 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 3.2.5. If $X_n$ converges to $\mu$ in probability. Then, $T_n = |\text{sgn}(X_n) - \text{sgn}(\mu)|1_{|\mu|>0}$ 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 $\infty$. □

Lemma 3.2.6. Assume A2 and A3. Then, for fixed $x_1$, $\sup_y |\hat{d}(x_1, y) - d(x_1, y)|$ converges to zero.

Proof. By the triangle inequality,

$$\sup_y |\hat{d}(x_1, y) - d(x_1, y)| \leq \sup_y |\Delta(y; x_1, -1)| + \sup_y |\Delta(y; x_1, 1)|,$$

where

$$\Delta(y; x_1, a_1) = \int \hat{F}_e(y - u - |v|)d\hat{G}(u, v \mid x_1, a_1) - \int F_e(y - u - |v|)dG(u, v \mid x_1, a_1).$$

Thus, we show $\sup_y |\Delta(y; x_1, a_1)|$ converges in probability to zero for an arbitrary $a_1$. Applying the triangle inequality leads to the upper bound

$$\sup_y |\Delta(y; x_1, a_1)| \leq \sup_y \int \hat{F}_e(y - u - |v|)d\hat{G}(u, v \mid x_1, a_1) - dG(u, v \mid x_1, a_1)$$

$$+ \sup_y \int \left|\hat{F}_e(y - u - |v|) - F_e(y - u - |v|)\right|dG(u, v \mid x_1, a_1). \quad (3.17)$$

An upper bound on the right-hand side of (3.17) is

$$\int \left|d\hat{G}(u, v \mid x_1, a_1) - dG(u, v \mid x_1, a_1)\right|$$

$$+ \sup_w \left|\hat{F}_e(w) - F_e(w)\right|\int dG(u, v \mid x_1, a_1)$$

$$= \int \left|d\hat{G}(u, v \mid x_1, a_1) - dG(u, v \mid x_1, a_1)\right| + \sup_w \left|\hat{F}_e(w) - F_e(w)\right|, \quad (3.18)$$

where we have used the fact that $\sup_w \hat{F}_e(w) = 1$ and $\int dG(u, v \mid x_1, a_1) = 1$. The first and second terms in (3.18) are $o_p(1)$ by assumptions A3 and A2. □

Lemma 3.2.7. Assume A1. Then,

$$\sup_{\pi_1, \pi_2} |\text{pr}^{\pi_1, \pi_2}(Y \leq y) - \text{pr}^{\pi_{1}, \pi_{2}}(Y \leq y)|$$

converges to zero in probability.
Proof. Define

\[ \hat{\Delta}_c(y; x_1, a_1, x_2) = F_\varepsilon \{ y - m(x_1, a_1, x_2) - \text{sgn}\{c(x_1, a_1, x_2)\} c(x_1, a_1, x_2) \} - F_\varepsilon \{ y - m(x_1, a_1, x_2) - |c(x_1, a_1, x_2)| \}, \]

\[ \hat{\Delta}_c(x_1, a_1, x_2) = |\text{sgn}\{c(x_1, a_1, x_2)\} - \text{sgn}\{c(x_1, a_1, x_2)\}| \mathbb{1}_{|c(x_1, a_1, x_2)| > 0}. \]

Note that for each fixed \((x_1, a_1, x_2)\),

\[ \left| \hat{\Delta}_c(y; x_1, a_1, x_2) \right| \leq \hat{\Delta}_c(x_1, a_1, x_2). \tag{3.19} \]

Using the definitions given in Section 3.2.2 and inequality (3.19),

\[
\sup_{\pi_1, \pi_2} \left| \text{pr}^{\pi_1, \pi_2}_Y (Y \leq y) - \text{pr}^{\pi_1, \pi_2}_Y (Y \leq y) \right|
\]

\[
= \sup_{\pi_1, \pi_2} \left| \int \int \hat{\Delta}_c(y; x_1, \pi_1(x_1), x_2) dF_{X_2 | X_1, A_1} (x_2 | x_1, \pi_1(x_1)) dF_{X_1} (x_1) \right|
\]

\[
\leq \sup_{\pi_1, \pi_2} \int \int \left| \hat{\Delta}_c(y; x_1, \pi_1(x_1), x_2) \right| dF_{X_2 | X_1, A_1} (x_2 | x_1, \pi_1(x_1)) dF_{X_1} (x_1)
\]

\[
\leq \sup_{\pi_1} \int \int \left| \hat{\Delta}_c(x_1, \pi_1(x_1), x_2) \right| dF_{X_2 | X_1, A_1} (x_2 | x_1, \pi_1(x_1)) dF_{X_1} (x_1),
\]

where we have used the fact that \( \hat{\Delta}_c(x_1, \pi_1(x_1), x_2) \) does not depend on \( y \). Because \( \pi_1(\cdot) \) has range \([-1, 1]\), an upper bound on the right-hand side above is

\[
\int \int \sum_{a_1 \in \{-1, 1\}} \hat{\Delta}_c(x_1, a_1, x_2) dF_{X_2 | X_1, A_1} (x_2 | x_1, a_1) dF_{X_1} (x_1)
\]

\[
= \sum_{a_1 \in \{-1, 1\}} \int \int \hat{\Delta}_c(x_1, a_1, x_2) dF_{X_2 | X_1, A_1} (x_2 | x_1, a_1) dF_{X_1} (x_1)
\]

\[
= \sum_{a_1 \in \{-1, 1\}} E \left\{ \hat{\Delta}_c(X_1, A_1, X_2) | A_1 = a_1, D \right\}, \tag{3.20}
\]

which does not depend on \( \pi_1 \). For each fixed \( a_1 \),

\[
E \left\{ \hat{\Delta}_c(X_1, A_1, X_2) | A_1 = a_1 \right\}
\]

\[
= \int \int E \left\{ \hat{\Delta}_c(x_1, a_1, x_2) \right\} dF_{X_2 | X_1, A_1} (x_2 | x_1, a_1) dF_{X_1} (x_1),
\]

where \( E\{\hat{\Delta}_c(x_1, a_1, x_2)\} \) converges to zero by Lemma 3.2.5 for each \((x_1, a_1, x_2)\). Because \( 0 \leq E\{\hat{\Delta}_c(x_1, a_1, x_2)\} \leq 2 \), applying the Dominated Convergence Theorem gives the result that
\( E \{ \hat{\Delta}_c(X_1, A_1, X_2) \mid A_1 = a_1 \} \) converges to zero, which implies \( E \{ \hat{\Delta}_c(X_1, A_1, X_2) \mid A_1 = a_1, D \} \) is \( o_p(1) \) for each fixed \( a_1 \) by Lemma 3.2.5. Thus, the right-hand side of (3.20) is \( o_p(1) \).

**Lemma 3.2.8.** Assume A2 and A3, and fix \( \lambda \in \mathbb{R} \). Then,

\[
\left| pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) - pr^{\tilde{\pi}_1, \lambda, \pi_2^2}(Y \leq \lambda) \right|
\]

converges to zero in probability.

**Proof.** Define \( \hat{\Delta}_G(x_1; u, v) = dG\{u, v \mid x_1, \pi_1^{\text{Q}}(x_1)\} - dG\{u, v \mid x_1, \pi_1^{\text{Q}}(x_1)\} \), and note that \( \hat{\Delta}_G(x_1; u, v) = \{\pi_1^{\text{Q}}(x_1) - \pi_1^{\text{Q}}(x_1)\}\{dG(u, v \mid x_1, -1) - dG(u, v \mid x_1, 1)\}/2 \). Using the definitions given in Section 3.2.2,

\[
\left| pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) - pr^{\tilde{\pi}_1, \lambda, \pi_2^2}(Y \leq \lambda) \right| = \left| \int \int F_\varepsilon(\lambda - u - y|v)\hat{\Delta}_G(x_1; u, v)dF_x(x_1) \right|
\]

\[
\leq \int \int |d(x_1, \lambda)| \left| \pi_1^{\text{Q}}(x_1) - \pi_1^{\text{Q}}(x_1) \right| dF_x(x_1). \tag{3.21}
\]

Substituting \( \pi_1^{\text{Q}}(x_1) = \text{sgn}\{d(x_1, \lambda)\}, \pi_1^{\text{Q}}(x_1) = \text{sgn}\{\tilde{d}(x_1, \lambda)\}, \) and noting

\[
|d(x_1, \lambda)| \left| \text{sgn}\{\tilde{d}(x_1, \lambda)\} - \text{sgn}\{d(x_1, \lambda)\} \right| \leq \mathbb{I}_{|d(x_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(x_1, \lambda)\} - \text{sgn}\{d(x_1, \lambda)\} \right|,
\]

an upper bound on the right-hand side of (3.21) is

\[
\int \mathbb{I}_{|d(x_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(x_1, \lambda)\} - \text{sgn}\{d(x_1, \lambda)\} \right| dF_x(x_1)
\]

\[
= E \left[ \mathbb{I}_{|d(x_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(X_1, \lambda)\} - \text{sgn}\{d(X_1, \lambda)\} \right| \mid D \right].
\]

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

\[
E \left[ \mathbb{I}_{|d(x_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(X_1, \lambda)\} - \text{sgn}\{d(X_1, \lambda)\} \right| \mid D \right]
\]

\[
= \int E \left[ \mathbb{I}_{|d(x_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(x_1, \lambda)\} - \text{sgn}\{d(x_1, \lambda)\} \right| \mid D \right] dF_x(x_1).
\]

The inside expectation converges to zero by Lemma 3.2.5, and applying the Dominated Convergence Theorem gives the result that the right-hand side above converges to zero. Thus, appealing to Lemma 3.2.5, we have shown \( \left| pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) - pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) \right| \) is bounded above by

\[
E\left| \mathbb{I}_{|d(X_1, \lambda)| > 0} \left| \text{sgn}\{\tilde{d}(X_1, \lambda)\} - \text{sgn}\{d(X_1, \lambda)\} \right| \mid D \right|\]

which is \( o_p(1) \).

**Proof of Theorem 1.** Fix \( \lambda \in \mathbb{R} \). Define \( \Delta(\lambda) = pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) - pr^{\pi_1, \lambda, \pi_2^2}(Y \leq \lambda) \). Then, by the
triangle inequality,

\[ |\Delta(\lambda)| \leq \left| \pr^{\pi_{1, \lambda}}^{\Qi}, \pi_2^*(Y \leq \lambda) - \pr^{\pi_{1, \lambda}}^{\Qi}, \pi_2^*(Y \leq \lambda) \right| + \left| \pr^{\pi_{1, \lambda}}^{\Qi}, \pi_2^*(Y \leq \lambda) - \pr^{\pi_{1, \lambda}}^{\Qi}, \pi_2^*(Y \leq \lambda) \right|. \tag{3.22} \]

The first term on the right-hand side of (3.22) is \( o_p(1) \) by Lemma 3.2.7, and the second term on the right-hand side of (3.22) is \( o_p(1) \) by Lemma 3.2.8.

**Lemma 3.2.9.** Assume A2 and A4. Then, \( \sup_y n^{-1} \sum_{i=1}^n |\hat{d}(\mathbf{X}_{1i}, y) - d(\mathbf{X}_{1i}, y)| \) converges to zero in probability.

**Proof.** An upper bound on \( \sup_y n^{-1} \sum_{i=1}^n |\hat{d}(\mathbf{X}_{1i}, y) - d(\mathbf{X}_{1i}, y)| \) is

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

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

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

\[
+ \sum_{a_1=1,-1} \sup_y \frac{1}{n} \sum_{i=1}^n \left| \int F_e (y - u - |v|) d\hat{G}(u, v \mid \mathbf{X}_{1i}, a_1) - dG(u, v \mid \mathbf{X}_{1i}, a_1) \right|
\]

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

The term \( \sup_w |\hat{F}_e (w) - F_e (w)| \) is \( o_p(1) \) by assumption A2, and for each \( a_1 \),

\[
n^{-1} \sum_{i=1}^n |\hat{d}(\mathbf{X}_{1i}, v) - dG(u, v \mid \mathbf{X}_{1i}, a_1) - dG(u, v \mid \mathbf{X}_{1i}, a_1)| \) is \( o_p(1) \) by assumption A4.

**Lemma 3.2.10.** Assume A2 and A4. Then, \( \sup_y |\Delta(y)| \) converges in probability to zero, where

\[
\Delta(y) = \frac{1}{n} \sum_{i=1}^n \int \hat{F}_e (y - u - |v|) d\hat{G}[u, v \mid \mathbf{X}_{1i}, \text{sgn}\{\hat{d}(\mathbf{X}_{1i}, y)\}] - \frac{1}{n} \sum_{i=1}^n \int F_e (y - u - |v|) dG[u, v \mid \mathbf{X}_{1i}, \text{sgn}\{d(\mathbf{X}_{1i}, y)\}]. \tag{3.23}
\]
Lemma 3.2.11. which is in (3.24) is bounded above by

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

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

\[
\frac{1}{2} \left\{ d\hat{G}(u, v \mid X_{1i}, 1) + d\hat{G}(u, v \mid X_{1i}, -1) \right\} - \frac{\text{sgn}\{\hat{d}(X_{1i}, y)\}}{2} \left\{ d\hat{G}(u, v \mid X_{1i}, -1) - d\hat{G}(u, v \mid X_{1i}, 1) \right\},
\]

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

\[
\sup_y \frac{1}{n} \sum_{i=1}^{n} |\Delta_i(y)| + \sup_y \frac{1}{n} \sum_{i=1}^{n} |\hat{d}(X_{1i}, y)| - |d(X_{1i}, y)|,
\]

(3.24)

where

\[
\Delta_i(y) = \int \hat{F}_i(y - u - |v|) \left\{ d\hat{G}(u, v \mid X_{1i}, 1) + d\hat{G}(u, v \mid X_{1i}, -1) \right\}
- \int F_i(y - u - |v|) \left\{ dG(u, v \mid X_{1i}, 1) + dG(u, v \mid X_{1i}, -1) \right\}.
\]

The term \( \sup_y n^{-1} \sum_{i=1}^{n} |\hat{d}(X_{1i}, y)| - |d(X_{1i}, y)| \) in (3.24) is bounded above by \( \sup_y n^{-1} \sum_{i=1}^{n} |\hat{d}(X_{1i}, y) - d(X_{1i}, y)| \), which is \( o_p(1) \) by Lemma 3.2.9. It can be shown the first term in (3.24) is bounded above by

\[
2 \sup_w |\hat{F}_i(w) - F_i(w)| + \frac{1}{n} \sum_{i=1}^{n} \int \left| d\hat{G}(u, v \mid X_{1i}, 1) - dG(u, v \mid X_{1i}, 1) \right|
+ \frac{1}{n} \sum_{i=1}^{n} \int \left| d\hat{G}(u, v \mid X_{1i}, -1) - dG(u, v \mid X_{1i}, -1) \right|,
\]

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

\[
\text{Lemma 3.2.11. For every fixed } x_1,
\]

\[
\lim_{y \to \infty} \int F_i(y - u - |v|) dG[u, v \mid x_1, a_1 = \text{sgn}\{d(x_1, y)\}] = 1,
\]

\[
\lim_{y \to -\infty} \int F_i(y - u - |v|) dG[u, v \mid x_1, a_1 = \text{sgn}\{d(x_1, y)\}] = 0.
\]
Proof. For each fixed $x_1, a_1$,

$$
\lim_{y \to \infty} \int F_{\epsilon}(y - u - |v|)dG(u, v \mid x_1, a_1) = 1,
$$

$$
\lim_{y \to -\infty} \int F_{\epsilon}(y - u - |v|)dG(u, v \mid x_1, a_1) = 0,
$$

because $\int F_{\epsilon}(y - u - |v|)dG(u, v \mid x_1, a_1)$ is the conditional expectation of a distribution function in $y$. Thus, even if the policy $\text{sgn}\{d(x_1, y)\}$ does not converge as $y \to \infty (-\infty)$, $\lim_{y \to \infty (-\infty)} \int F_{\epsilon}(y - u - |v|)dG[u, v \mid x_1, a_1 = \text{sgn}\{d(h_1, y)\}]$ must converge to 1 (0).

Lemma 3.2.12. For every $x_1$ in the domain of $X_1$,

$$
\int F_{\epsilon}(y - u - |v|)dG[u, v \mid x_1, \text{sgn}\{d(x_1, y)\}]
$$

is non-decreasing in $y$.

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

$$
\int F_{\epsilon}(s - u - |v|)dG[u, v \mid x_1, \text{sgn}\{d(x_1, s)\}]
$$

$$
- \int F_{\epsilon}(t - u - |v|)dG[u, v \mid x_1, \text{sgn}\{d(x_1, t)\}] \quad (3.25)
$$

is non-negative. Because $\int F_{\epsilon}(s - u - |v|)dG[u, v \mid x_1, \text{sgn}\{d(x_1, s)\}]$ can be written as

$$
\frac{1}{2} \left\{ \int F_{\epsilon}(s - u - |v|)dG(u, v \mid x_1, -1) + \int F_{\epsilon}(s - u - |v|)dG(u, v \mid x_1, 1) - |d(x_1, s)| \right\},
$$

(3.25) simplifies to

$$
\frac{1}{2} \left[ \int \{F_{\epsilon}(s - u - |v|) - F_{\epsilon}(t - u - |v|)\} dG(u, v \mid x_1, -1) 
$$

$$
+ \frac{1}{2} \left[ \int \{F_{\epsilon}(s - u - |v|) - F_{\epsilon}(t - u - |v|)\} dG(u, v \mid x_1, 1) 
$$

$$
- \frac{1}{2} \{|d(x_1, s)| - |d(x_1, t)|\}. 
$$

46
The expression above is greater than or equal to zero. To see this, note that

\[
|d(x_1, s) - d(x_1, t)| \leq |d(x_1, s)| - |d(x_1, t)| \\
\leq |d(x_1, s) - d(x_1, t)| \\
\leq \int \{F(x) - u - |v|\} dG(u, v \mid x_1, -1) \\
+ \int \{F(x) - u - |v|\} dG(u, v \mid x_1, 1).
\]

\[\text{Lemma 3.2.13. Assume } F(x) \text{ is continuous. For any fixed } x_1 \text{ in the domain of } X_1,\]
\[
\int F(x) (y - u - |v|) dG[u, v \mid x_1, \text{sgn}\{d(x_1, y)\}]
\]

is continuous in \( y \).

**Proof.** This follows immediately by writing \( \int F(x) (y - u - |v|) dG[u, v \mid x_1, \text{sgn}\{d(x_1, y)\}] \) as

\[
\frac{1}{2} \left\{ \int F(x) (y - u - |v|) dG(u, v \mid x_1, -1) + \int F(x) (y - u - |v|) dG(u, v \mid x_1, 1) - |d(x_1, y)| \right\},
\]

a linear combination of continuous functions.

**Lemma 3.2.14. Assume A2 and A4. Then, }_{y} |L_n(y) - L(y)| \text{ converges in probability to zero, where }\]
\[
L_n(y) = \frac{1}{n} \sum_{i=1}^{n} \int F(x) (y - u - |v|) dG[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}],
\]
\[
L(y) = E \left( \int F(x) (y - u - |v|) dG[u, v \mid X_1, \text{sgn}\{d(X_1, y)\}] \right).
\]

**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 3.2.11, 3.2.12, and 3.2.13, 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 \( y \in \{y_1, \ldots, y_k-1\} \). Thus, \( \lim_{y} |L_n(y) - L(y)| < \delta \) almost surely. Because \( \delta \) is arbitrary, the result holds for each \( \delta \), which implies the limit superior is zero.

47
Lemma 3.2.15. Assume A2 and A4. Then, \( \hat{y}^*_\tau \) converges in probability to \( y^*_\tau \).

Proof. Define

\[
\Delta(y) = \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_c(y - u - |v|)d\tilde{G}[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}] \\
- E \left( \int F_c(y - u - |v|)dG[u, v \mid X_1, \text{sgn}\{d(X_1, y)\}] \right)
\]

By the triangle inequality,

\[
\sup_y |\Delta(y)| \leq \sup_y |\Delta_1(y)| + \sup_y |\Delta_2(y)|,
\]

where

\[
\Delta_1(y) = \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_c(y - u - |v|)d\tilde{G}[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}] \\
- \frac{1}{n} \sum_{i=1}^{n} \int F_c(y - u - |v|)dG[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}],
\]

\[
\Delta_2(y) = \frac{1}{n} \sum_{i=1}^{n} \int F_c(y - u - |v|)dG[u, v \mid X_{1i}, A_1 = \text{sgn}\{d(X_{1i}, y)\}] \\
- E \left( \int F_c(y - u - |v|)dG[u, v \mid X_1, A_1 = \text{sgn}\{d(X_1, y)\}] \right).
\]

The term \( \sup_y |\Delta_1(y)| \) converges to zero in probability by Lemma 3.2.10, and \( \sup_y |\Delta_2(y)| \) converges to zero in probability by Lemma 3.2.14. Thus, we have shown uniform convergence of \( \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_c(y - u - |v|)d\tilde{G}[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}] \) to \( E \left( \int F_c(y - u - |v|)dG[u, v \mid X_1, \text{sgn}\{d(X_1, y)\}] \right) \), which implies the infimums converge. That is, \( \hat{y}^*_\tau = \inf \left( y : \frac{1}{n} \sum_{i=1}^{n} \int \hat{F}_c(y - u - |v|)d\tilde{G}[u, v \mid X_{1i}, \text{sgn}\{d(X_{1i}, y)\}] \geq \tau \right) \)

converges in probability to

\[
y^*_\tau = \inf \left\{ y : E \left( \int F_c(y - u - |v|)dG[u, v \mid X_1, \text{sgn}\{d(X_1, y)\}] \right) \geq \tau \right\}.
\]
Lemma 3.2.16. Assume A2–A4. Let \( x_1 \) be fixed and arbitrary. Then,
\[
\left| \hat{d}(x_1, \hat{y}_r^*) - d(x_1, y_r^*) \right|
\]
converges to zero in probability.

Proof. By the triangle inequality,
\[
\left| \hat{d}(x_1, \hat{y}_r^*) - d(x_1, y_r^*) \right| \leq \left| \hat{d}(x_1, \hat{y}_r^*) - d(x_1, \hat{y}_r^*) \right| + \left| d(x_1, \hat{y}_r^*) - d(x_1, y_r^*) \right|
\]
\[
\leq \sup_y \left| \hat{d}(x_1, y) - d(x_1, y) \right| + \left| d(x_1, \hat{y}_r^*) - d(x_1, y_r^*) \right|.
\]
The right-hand side of the previous expression is \( o_p(1) \) because \( \sup_y \left| \hat{d}(x_1, y) - d(x_1, y) \right| \) is \( o_p(1) \) by Lemma 3.2.6. Note that continuity of \( d(x_1, y) \) is implied by assumption A2, and thus, \( \left| d(x_1, \hat{y}_r^*) - d(x_1, y_r^*) \right| \) is \( o_p(1) \) by Lemma 3.2.15 and the continuous mapping theorem.

Proof of Theorem 2. Choose \( \delta > 0 \) such that \( \Pr\{ \left| d(H_1, y_r^*) \right| \leq \delta \} \leq \eta/2 \), which is possible by assumption A5. We begin by showing \( \sup_y \left| \Delta(y) \right| \) converges to zero in probability, where
\[
\Delta(y) = \Pr\{ \text{sgn}(\hat{d}(\cdot, \hat{y}_r^*)), \hat{\pi}_i^2(Y \leq y) - \Pr\{ \text{sgn}(d(\cdot, y_r^*)), \pi_i^2(Y \leq y) \}.
\]

By the triangle inequality,
\[
\sup_y \left| \Delta(y) \right| \leq \sup_y \left| \Delta_1(y) \right| + \sup_y \left| \Delta_2(y) \right|,
\]
where
\[
\Delta_1(y) = \Pr\{ \text{sgn}(\hat{d}(\cdot, \hat{y}_r^*)), \hat{\pi}_i^2(Y \leq y) - \Pr\{ \text{sgn}(d(\cdot, y_r^*)), \pi_i^2(Y \leq y) \},
\]
\[
\Delta_2(y) = \Pr\{ \text{sgn}(\hat{d}(\cdot, \hat{y}_r^*)), \pi_i^2(Y \leq y) - \Pr\{ \text{sgn}(d(\cdot, y_r^*)), \pi_i^2(Y \leq y) \}.
\]

Note that
\[
\sup_y \left| \Delta_1(y) \right| \leq \sup_{\pi_1, y} \left| \Pr\{ \hat{\pi}_i^2, \hat{\pi}_i^2(Y \leq y) - \Pr\{ \pi_i^2, \pi_i^2(Y \leq y) \},
\]
where the right-hand side is \( o_p(1) \) by Lemma 3.2.7. It can be shown that
\[
\sup_y \left| \Delta_2(y) \right| \leq E \left( \sup_y \left| d(Y_1, y) \left[ \text{sgn}\{ \hat{d}(X_1, \hat{y}_r^*) \} - \text{sgn}\{ d(X_1, y_r^*) \} \right] \right| \right) \]
\[
\leq E \left[ \left| \text{sgn}\{ \hat{d}(X_1, \hat{y}_r^*) \} - \text{sgn}\{ d(X_1, y_r^*) \} \right| \right] .
\]
We write the right-hand side as

\[
E \left[ \mathbb{1}_{|d(x_1, y^*_r)| \leq \delta} \left| sgn \left\{ \tilde{d}(X_1, \hat{y}^*_r) \right\} - sgn \left\{ d(X_1, y^*_r) \right\} \right| \right] \\
+ E \left[ \mathbb{1}_{|d(x_1, y^*_r)| > \delta} \left| sgn \left\{ \tilde{d}(X_1, \hat{y}^*_r) \right\} - sgn \left\{ d(X_1, y^*_r) \right\} \right| \right] \\
\leq 2pr\{ |d(X_1, y^*_r)| \leq \delta \}
+ E \left[ \mathbb{1}_{|d(x_1, y^*_r)| > \delta} \left| sgn \left\{ \tilde{d}(X_1, \hat{y}^*_r) \right\} - sgn \left\{ d(X_1, y^*_r) \right\} \right| \right].
\]  

(3.27)

By our choice of \( \delta \), \( 2pr\{ |d(X_1, y^*_r)| \leq \delta \} \leq \eta \). The second term on the right-hand side of (3.27) is bounded above by \( E[\mathbb{1}_{|d(x_1, y^*_r)| > 0} | sgn \{ \tilde{d}(X_1, \hat{y}^*_r) \} - sgn \{ d(X_1, y^*_r) \} | \right] \). Taking the expectation of this upper bound with respect to \( D \),

\[
E \left[ \mathbb{1}_{|d(x_1, y^*_r)| > \delta} \left| sgn \left\{ \tilde{d}(X_1, \hat{y}^*_r) \right\} - sgn \left\{ d(X_1, y^*_r) \right\} \right| \right] \\
= \int E \left[ \mathbb{1}_{|d(x_1, y^*_r)| > \delta} \left| sgn \left\{ \tilde{d}(x_1, \hat{y}^*_r) \right\} - sgn \left\{ d(x_1, y^*_r) \right\} \right| \right] dF_{X_1}(x_1),
\]

where the inside expectation converges to zero by Lemmas 3.2.5 and 3.2.16. The Dominated Convergence Theorem applies, giving the result that \( E[\mathbb{1}_{|d(x_1, y^*_r)| > 0} | sgn \{ \tilde{d}(X_1, \hat{y}^*_r) \} - sgn \{ d(X_1, y^*_r) \} | \right] \) converges to zero. Thus, by Lemma 3.2.5, \( E[\mathbb{1}_{|d(x_1, y^*_r)| > 0} | sgn \{ \tilde{d}(X_1, \hat{y}^*_r) \} - sgn \{ d(X_1, y^*_r) \} | \right] \) is \( o_p(1) \).

We have shown that (3.26) is \( o_p(1) \), which implies

\[
q^{\pi_1, \tau, \pi_2}(\tau) = \inf \left\{ y : pr^{\pi_1, \tau, \pi_2}(Y \leq y) \geq \tau \right\}
\]

converges in probability to

\[
q^{\pi_1, \tau, \pi_2}(\tau) = \inf \left\{ y : pr^{\pi_1, \tau, \pi_2}(Y \leq y) \geq \tau \right\}.
\]

\[ \blacksquare \]
\begin{align*}
X_1 &\sim \text{Normal}_{p=2}(\mathbf{1}_2, \Sigma), \quad A_t \in \{-1, 1\}, \ 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_1^\top)^\top, \\
Y &= H_2^\top \beta_{2,0} + A_2 H_2^\top \beta_{2,1} + \epsilon, \quad \epsilon \sim \text{Normal}(0, 1),
\end{align*}

where \( \mathbf{1}_p \) is a \( p \times 1 \) vector of 1s. The matrix \( \Sigma \) is a covariance 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

\[
\gamma_0 = (1, 0.5, 0)^\top, \quad \gamma_1 = (-1, -0.5, 0)^\top, \quad \beta_{2,0} = (0.25, -1, 0.5)^\top, \quad \beta_{2,1} = (1, -0.5, -0.25)^\top,
\]

which were chosen to ensure that the mean-optimal treatment produced a more variable response for some patients. Results are based on \( M = 1,000 \) generated data sets. For each data set, 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 $F_{\ell}(\cdot)$, which is correctly specified for the generative model above. The mean-variance modeling techniques along with a Gaussian copula model for the standardized residuals discussed in Section 2.4 is also correctly specified and is used as the estimator for $G(\cdot, \cdot | x_1, a_1)$.

To study the performance of the TIQ-learning algorithm, we compare values of the cumulative distribution function of the final response that arise when treatment is assigned according to the estimated TIQ-learning, IQ-learning, and $Q$-learning regimes. Define $\Pr_{\pi_j}(Y > \lambda)$ to be the true probability that $Y$ exceeds $\lambda$ given treatments are assigned according to $\pi_j = (\tilde{\pi}_{1,j}, \tilde{\pi}_{2,j})$, the regime estimated from the $j^{th}$ generated data set. For threshold values $\lambda = -2, 2, 4$, we estimate $\Pr(\tilde{\pi}_j(Y > \lambda))$ using $\sum_{j=1}^M \tilde{\Pr}_{\pi_j}(Y > \lambda)/M$, where $\tilde{\Pr}_{\pi_j}(Y > \lambda)$ is an estimate of $\Pr_{\pi_j}(Y > \lambda)$ obtained by calculating the proportion of test patients consistent with regime $\pi_j$ whose observed $Y$ values are greater than $\lambda$. Thus, our estimate is an average over training data sets and test set observations. Results in Fig. 3.1 when $\lambda = -2$ and 4 show a clear advantage of TIQ-learning in terms of the proportion of distribution mass above $\lambda$ as $C$, the degree of heteroskedasticity in the second-stage covariates $X_2$, increases. All three learning methods perform similarly when $\lambda = 2$, as this threshold is close to the center of the distribution.

Figure 3.2 illustrates how the optimal treatment regime for a test set of 1,000 individuals changes as $\lambda$ varies. Results are shown for our generative model with $C = 0.5$. 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 in the middle plot are averaged over 100 Monte Carlo iterations and closely resemble the true policies on the left.
Figure 3.3: **Left to Right:** $\tau = 0.1, 0.5, 0.75$. Solid black, true optimal quantiles; dotted black, quantiles under treatment randomization; dashed turquoise/purple/orange lines with circles/squares/triangles, quantiles under $QIQ{-}$, $Q{-}$, and $IQ{-}$-learning.

Looking at the true optimal treatments, it appears the $\lambda = 2$ threshold is close to a shift in the regime for many patients. Because the generative distribution of $Y$ is continuous for a fixed first-stage treatment and optimal second-stage treatment, thresholds at which the regime shifts for a given patient indicate there is no unique best first-stage treatment for that patient. In finite samples, it may be difficult to detect the unique best treatment in a neighborhood of a regime change. This appears to be the case when $\lambda = 2$, as $TIQ$-learning has the most difficulty identifying the true optimal treatments at this threshold. Although the estimated $Q$-learning regime does not depend on $\lambda$, it is plotted for each $\lambda$ value to aid visual comparison. The treatments recommended by $Q$-learning differ the most from the optimal treatments when $\lambda$ is far from the center of the distribution.

### 3.3.2 Quantile Interactive $Q$-learning simulations

To study the performance of the $QIQ$-learning algorithm, we compare quantiles of $Y$ that arise 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 $\hat{q}^{\pi_j} (\tau)$ 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 $\sum_{j=1}^{M} \hat{q}^{\pi_j}(\tau)/M$, where $\hat{q}^{\pi_j}(\tau)$ is an estimate of $\hat{q}^{\pi_j}(\tau)$ 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 3.2.2 is satisfied, so the true optimal regime is attained asymptotically. The results in Fig. 3.3 indicate that the lowest quantile, $\tau = 0.1$, suffers under the $Q$-learning regime as heterogeneity in the second-stage histories increases, represented by 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 gives similar results to maximizing the median, and methods that directly optimize the mean are expected to perform best in this setting.

3.4 Application to STAR*D

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 (SMART; Lavori and Dawson, 2004; Murphy, 2005a) studying personalized treatment strategies for patients with major depressive disorder. In STAR*D, depression is measured by the Quick Inventory of Depressive Symptomatology (QIDS) score, a one-number summary obtained from a survey 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., 2013b) 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, thus 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 (Fava et al., 2003; Rush et al., 2004), a drug in the class of SSRIs (Edwards, 1992). 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 were allowed to exit 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; this results in a reduced data set of 132 patients. 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−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 3.1. We describe all models used in the
Table 3.1: Variables used in the STAR*D analysis.

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

analysis below.

**Step 1: Second-stage regression**

We assume the linear working model $Y = H_2^\top \beta_2 + A_2 H_2^\top \tilde{\beta}_2 + \epsilon$ at stage two, where $H_2 = (1, qids1, slope1, A1)^\top$, $E(\epsilon) = 0$, var($\epsilon$) = $\sigma^2$, and $\epsilon$ is independent of $H_2$ and $A_2$. We fit this model using least squares.

**Step 2: Estimation of $F_\cdot(\cdot)$**

A normal quantile-quantile plot of the residuals from the previous regression step indicates slight deviation from normality, so we use the empirical estimator of $F_\cdot(\cdot)$ given in Section 3.2.4.

**Step 3: Estimation of $G(\cdot, \cdot \mid x_1, a_1)$**

We estimate the conditional mean and variance functions of $m(H_2) = H_2^\top \beta_{2,0}$ and $c(H_2) = H_2^\top \tilde{\beta}_{2,1}$ following steps described in Section 3.2.5. For the mean functions, we take $H_{1,0} = H_{1,1} = (1, X_{1,0})^\top$ with $X_1 = (qids0, slope0)^\top$ 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 $\{m(H_2; \tilde{\beta}_{2,0}), c(H_2; \tilde{\beta}_{2,1})\}$ given $X_1$ and $A_1$, resulting in our estimate of $G(\cdot, \cdot \mid x_1, a_1)$ which we denote by $\hat{G}(\cdot, \cdot \mid x_1, a_1)$.

**Step 4: The estimated optimal decision rules**

The estimated first-stage optimal rule is

$\hat{x}_{1,\lambda}^{\text{no1}}(x_1) = \arg \min_{a_1} \int \hat{F}_{\epsilon}(21 - u - |v|)d\hat{G}(u, v \mid x_1, a_1).$
The estimated optimal treatment at stage two obtained in Step 1 is

$$\hat{\pi}^*_2(h_2) = \text{sgn}(-1.66 + 0.15 * \text{qids1} - 4.03 * \text{slope1} - 0.68 * A1).$$

Based on our discussion in Section 3.2.2, we compare the estimated first-stage treatment recommendations to those recommended by the mean-optimal rule, \(\arg \min_{a_1} \int (-u - |v|)d\bar{G}(u, v \mid x_1, a_1),\) for each observed \(x_1\) in the data. Only one patient out of 132 is recommended differently. In addition, the difference in raw values of \(\int F_e(21 - u - |v|)d\bar{G}(u, v \mid x_1, a_1)\) for \(a_1 = 1, -1\) as well as \(\int (-u - |v|)d\bar{G}(u, v \mid x_1, a_1)\) 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 3.2: Estimated value of dynamic and non-dynamic regimes using the Augmented Inverse Probability Weighted Estimator.

<table>
<thead>
<tr>
<th>Estimated Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIQ-learning</td>
</tr>
<tr>
<td>Q-learning</td>
</tr>
<tr>
<td>(1, 1)</td>
</tr>
<tr>
<td>(-1, 1)</td>
</tr>
<tr>
<td>(1, -1)</td>
</tr>
<tr>
<td>(-1, -1)</td>
</tr>
</tbody>
</table>

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 \((H_2, A_2)\) 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 \(\tilde{Y} = H_{2,0}^\top \beta_{2,0} + |H_{2,1}^\top \beta_{2,1}|\) given \((H_1, A_1)\) using the estimated version of \(\tilde{Y}, H_{2,0}^\top \tilde{\beta}_{2,0} + |H_{2,1}^\top \tilde{\beta}_{2,1}|\), obtained from the second-stage regression. Recall that \(\tilde{Y}\) is the predicted future optimal outcome at stage one. We specify the working model

$$E(\tilde{Y} \mid X_1, A_1) = H_{1,0}^\top \beta_{1,0}^Q + A_1H_{1,1}^\top \beta_{1,1}^Q,$$

where \(H_{1,0}^\top = H_{1,1}^\top = (1, X_1^\top)\) and \(X_1 = (\text{qids0, slope0})^\top\). We fit the model using least squares. Then, the Q-learning estimated optimal first-stage rule is \(\hat{\pi}^{Q-opt}(x_1) = \text{sgn}(-0.95 + 0.13 * \text{qids1} + 2.17 * \text{slope1}).Q\)-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 3.2. Also included in Table 3.2 are value estimates for four non-dynamic regimes that treat everyone according to the decision rules \( \pi_1(x_1) = a_1 \) and \( \pi_2(h_2) = a_2 \) 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 \textit{TIQ}-learning and \textit{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 is appropriate and equivalent to maximizing remission probabilities for each patient with \textit{TIQ}-learning.
Chapter 4

Constrained $IQ$-learning

4.1 Introduction

Intolerable side-effects, high cost, and other undesirable factors may reduce the overall quality of life of patients being treated for chronic illnesses. Unfavorable quality of life outcomes may lead to non-compliance and, as a result, reduced efficacy of treatment strategies. In some cases, the risk of death may be increased by severe adverse events or comorbid conditions (Charlson et al., 1994). Thus, when treating patients with chronic illnesses, clinicians usually aim to assign the treatment-dose combination that optimally balances efficacy, side-effect burden, cost, and overall quality of life for each patient.

The classic tradeoff between treatment efficacy and tolerability has been studied within many disease-specific subfields including depression and other mental illnesses (Bostwick, 2010; Correll, 2010; Papakostas, 2010), HIV/AIDS (Burgoyne and Tan, 2008), occupational HIV exposure (Bassett et al., 2004), urinary conditions (Corcos et al., 2006), chronic pain management (Lange et al., 2010), and cancer (Delforge et al., 2010; Burstein et al., 2010; Lorenzo et al., 2011). These are only a few examples from the large volume of medical literature dedicated to the topic of balancing efficacy and competing outcomes. As demonstrated by the previous examples, two common approaches to this problem in the medical literature are to consider efficacy as the primary outcome in a clinical trial and report side effect profiles as a secondary analysis or to suggest guidelines for clinical practice based on meta-analyses.

Dynamic treatment regimes (DTRs) offer an alternative framework for developing evidence-based clinical practice for long-term treatment of chronic diseases. DTRs recommend treatments based on individual patient information, such as past treatments and outcomes, through a sequence of decision rules, one for each decision point. A large collection of DTR methods exists for optimizing the average response of a single outcome (e.g., efficacy), including: $Q$-learning (Watkins, 1989; Watkins and Dayan, 1992; Murphy, 2005a); Interactive $Q$-learning (Laber et al., 2013b); regularized $Q$-learning (Moodie and Richardson, 2010; Chakraborty et al., 2010; Song et al., 2011; Goldberg et al., 2012); $g$-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). Optimization of mean survival times is considered by Goldberg and Kosorok (2012). In general, these methods allow decision-makers to incorporate side effect burden, cost, and other competing outcomes only by forming a composite response that is a function of two or more of these competing outcomes.

Examples of methods for composite outcomes include Lizotte et al. (2010) and Wang et al. (2012b). However, Laber et al. (2014) point out several challenges of constructing a composite outcome: a population-level composite outcome ignores individual preferences, patients may not know or be able to effectively communicate their preferences, and preferences may evolve over time. The authors also show that misspecification of a composite outcome can unfavorably affect the quality of the estimated DTR. In Laber et al. (2014), preference elicitation is avoided and multiple outcomes are incorporated directly by presenting the decision-maker with a set of acceptable treatment strategies at each decision time point. For example, one treatment is recommended for a given patient if it is clearly superior in terms of efficacy and side effects. Otherwise, a set of non-inferior treatments is recommended when there is no clear winner, and the decision is left to clinician and patient judgement.

In this chapter, we pose a constrained optimization framework for estimating a DTR that maximizes a general distributional feature of a univariate or bivariate outcome (e.g., average or median efficacy, average efficacy and average quality of life score) subject to general distributional constraints on one or more competing outcomes (e.g., a fixed percentile of a side-effect score, efficacy variability, average cost). Sequential Multiple Assignment Randomized Trials (SMARTs; Lavori and Dawson, 2004; Murphy, 2005b) are a type of clinical trial design useful for estimating optimal DTRs. At each stage, patients are randomized to a feasible treatment that may depend on previous treatments and outcomes. We focus attention on estimating optimal constrained DTRs from SMART data with binary treatments available at each of two stages corresponding to two decision time points. Two-stage, binary treatment SMARTs are common in practice; for example, see the website of The Methodology Center at Pennsylvania State University; Director: Dr. Linda Collins (2012, http://methodology.psu.edu/ra/adap-inter/projects) and Laber (2013). In addition, this set-up facilitates concise notation and methodological development. We also restrict our focus to a two competing outcome setting.

4.2 Optimal Constrained Treatment Regimes

4.2.1 Notation and Set-up

Let \( \{(X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_i, Z_i)\}_{i=1}^n \) denote independent and identically distributed realizations of the trajectory \((X_1, A_1, X_2, A_2, Y, Z)\), one for each of \( n \) subjects. Baseline subject information is contained in \( X_1 \in \mathbb{R}^{p_1} \); \( A_1 \in \{-1, 1\} \) is the first-line treatment; \( X_2 \in \mathbb{R}^{p_2} \) contains interim information collected between the first- and second-line treatment assignments; \( A_2 \in \{-1, 1\} \) is the second-line treatment; \( Y \in \mathbb{R} \) is the primary outcome of interest; and \( Z \in \mathbb{R} \) is a second outcome of interest. For example, \( Y \)
might be efficacy and $Z$ a competing outcome such as side effect burden. Let $\mathbf{H}_2 = (X_1^T, A_1, X_1^T)^T$, so that $\mathbf{H}_2$ includes all information available to a decision-maker prior to assigning the second-stage treatment. First-stage decision rule $\pi_1$ maps the support of $X_1$ to the space of possible first-stage treatments, and second-stage decision rule $\pi_2$ maps the support of $\mathbf{H}_2$ to the space of possible second-stage treatments. The pair of decision rules $\pi = (\pi_1, \pi_2)$ is a DTR, where a patient presenting with $X_1 = x_1$ at stage one would be assigned treatment $\pi_1(x_1)$, and a patient presenting with $\mathbf{H}_2 = h_2$ would be assigned $\pi_2(h_2)$.

Let $pr_{Y,Z}^\pi(y, z)$ denote the joint distribution of $(Y, Z)$ when treatments are assigned according to the regime $\pi$, evaluated at $(y, z)$. Similarly, let $pr_{Y}^\pi(y)$ and $pr_{Z}^\pi(z)$ denote the marginal distributions of $Y$ and $Z$ under regime $\pi$, evaluated at $y$ and $z$, respectively. We assume the following:

$$Y = m_Y(\mathbf{H}_2) + A_2 c_Y(\mathbf{H}_2) + \epsilon_Y,$$

$$Z = m_Z(\mathbf{H}_2) + A_2 c_Z(\mathbf{H}_2) + \epsilon_Z,$$

for functions $\{m_Y(\cdot), c_Y(\cdot), m_Z(\cdot), c_Z(\cdot)\}$ and mean-zero error terms $(\epsilon_Y, \epsilon_Z)$ that are independent of $(\mathbf{H}_2, A_2)$ and have joint distribution $F_{\epsilon_Y, \epsilon_Z}(\cdot, \cdot)$.

Let $M(pr_{Y,Z}^\pi)$ denote a feature of the joint distribution function of $(Y, Z)$ under $\pi$ that we would like to maximize, and let $S(pr_{Y,Z}^\pi)$ be a $r$-dimensional feature of this distribution that we would like to bound above. In addition, let $\Pi$ denote a class of regimes of interest and for $b, c \in \mathbb{R}^r$ write $b \preceq c$ to denote $b_j \leq c_j$ for $j = 1, \ldots, r$. If one exists, we define the optimal $S$-constrained regime, $\pi_{S}^{\text{opt}}$, as a solution to

$$\max_{\pi \in \Pi} M(pr_{Y,Z}^\pi)$$

$$\text{s.t. } S(pr_{Y,Z}^\pi) \preceq \kappa,$$  \hspace{1cm} (4.1)

where $\kappa \in \mathbb{R}^r$ is a known vector of constants.

In practice, one might choose to simplify the constrained optimization problem in (4.1) by specifying an objective $M(pr_{Y,Z}^\pi)$ that depends only on the marginal distribution of $Y$. For example, if $Y$ is efficacy, one might specify $M(pr_{Y,Z}^\pi) = M(pr_Y^\pi)$ to be average efficacy, $\int y \, dpr_Y^\pi(y)$, the $\tau$-th quantile of the efficacy distribution, $M(pr_Y^\pi) = \inf\{y : pr_Y^\pi(y) \geq \tau\}$, or the probability that $Y$ exceeds a lower bound, say $l^*$, for which $M(pr_Y^\pi) = 1 - pr_Y^\pi(l^*)$. If $Z$ is a measure of side-effect burden, then one may specify the average side-effect burden, $S(pr_{Y,Z}^\pi) = S(pr_Z^\pi) = \int z \, dpr_Z^\pi(z)$, as a constraint. Alternatively, $S(pr_{Y,Z}^\pi)$ could be a two-dimensional constraint that considers both the variance of $Y$ and the probability that $Z$ exceeds some upper bound, say $u^*$, so that $S(pr_{Y,Z}^\pi) = \{\text{Var}^\pi Y, 1 - pr_Z^\pi(u^*)\}^T$, where $\text{Var}^\pi Y$ denotes the variance of $Y$ under regime $\pi$.

Remark 1. Mean-optimal, single-outcome methods such as $Q$-learning (Watkins, 1989; Watkins and

---

60
Dayan, 1992; Murphy, 2005a), Interactive Q-learning (Laber et al., 2013b), \( g \)-estimation (Robins, 2004), \( A \)-learning (Murphy, 2003), and policy search (Orellana et al., 2010; Zhao et al., 2012; Zhang et al., 2012b,a, 2013) solve (4.1) without the constraint and with \( M(\Pr_{Y,Z}^\pi) = M(\Pr_Y^\pi) \) specified as the average outcome under \( \pi \), i.e., \( M(\Pr_Y^\pi) = \int y \, d\Pr_Y^\pi(y) \).

**4.2.2 Estimating distributional summaries of \( Y \) and \( Z \)**

If the underlying joint distribution of the data \((X_1, A_1, X_2, A_2, Y, Z)\) is known, constrained optimization methods could be used to solve (4.1). In most settings, however, the data generating distribution is unknown, and both the objective function \( M(\Pr_{Y,Z}^\pi) \) and constraints \( S(\Pr_{Y,Z}^\pi) \) must be estimated. Let \( \hat{\Pr}_{Y,Z}^\pi(y, z) \) denote an estimator of the joint cumulative distribution function \( \Pr_{Y,Z}^\pi(y, z) \). If one exists, we define the estimated optimal \( S \)-constrained treatment regime, \( \hat{\pi}_S^{\text{opt}} \), as a solution to

\[
\max_{\pi \in \Pi} M(\hat{\Pr}_{Y,Z}^\pi) \\
\text{s.t. } S(\hat{\Pr}_{Y,Z}^\pi) \preceq \kappa,
\]

(4.2)

for known constant vector \( \kappa \in \mathbb{R}^r \). Broadly, solving (4.2) proceeds by estimating \( \Pr_{Y,Z}^\pi(y, z) \), sampling from the estimated distribution function to estimate the objective and constraints, and finding the solution using a grid search or stochastic programming algorithm. For some functionals \( M \) and \( S \), availability of an exact analytical form may negate the need for Monte Carlo approximation using samples from the estimated distribution. However, the methods developed in this chapter are computationally difficult in many settings, and thus, it may be more convenient in practice to approximate the functionals using samples from the estimated distribution.

**Remark 4.2.1.** We say (4.1) is **separable** if the objective and constraints depend only on the marginal distributions of \( Y \) and \( Z \) under regime \( \pi \), i.e., \( \Pr_{Y,Z}^\pi(y, z) \) and \( \Pr_{Y,Z}^\pi(y, z) \), and not on the joint distribution \( \Pr_{Y,Z}^\pi(y, z) \). Consider the problem of maximizing the mean of \( Y \), \( M(\Pr_{Y,Z}^\pi) = M(\Pr_Y^\pi) = \int y \, d\Pr_Y^\pi(y) \) subject to bounds on the variance of \( Y \) and probability that \( Z \) exceeds an upper bound \( u^* \), \( S(\Pr_{Y,Z}^\pi) = S(\Pr_Y^\pi, \Pr_Z^\pi) = \{\text{Var} \Pr_Y, 1 - \Pr_Z(u^*)\} \). Then, this example satisfies our definition of a separable constrained optimization problem. In this case, we need only estimate \( \Pr_Y^\pi(y) \) and \( \Pr_Z^\pi(z) \) instead of the joint distribution, \( \Pr_{Y,Z}^\pi(y, z) \).

Using the assumptions stated in Section 4.2.1, we now give the form of the marginal distributions of \( Y \) and \( Z \) under a given policy \( \pi = (\pi_1, \pi_2) \). Let \( \pi_2(h_2) = \text{sgn}\{f_2(h_2)\} \) for some function \( f_2(h_2) \). Similar to Section 3.2.2, define the following: \( F_{X_1}(\cdot) \) is the cumulative distribution function of vector \( X_1, F_{X_2 \mid x_1, A_1}(\cdot \mid x_1, a_1) \) is the conditional cumulative distribution function of vector \( X_2 \) given \( X_1 = x_1 \) and \( A_1 = a_1 \), and \( F_{e_Y}(\cdot) \) is the cumulative distribution function of \( e_Y \). In addition, define

\[
J_Y^\pi(x_1, x_2, y) = F_{e_Y}(y - m_Y(x_1, \pi_1(x_1), x_2)) - \text{sgn}\{f_2(x_1, \pi_1(x_1), x_2)\} c_Y(x_1, \pi_1(x_1), x_2).
\]
Then,

\[ \text{pr}^X_Y(y) = \int \int J^X_Y(x_1, x_2, y) dF_{X_2 | X_1, A_1 \{x_2 \mid x_1, \pi_1(x_1) \}} dF_{X_1}(x_1). \]

Let \( G_Y(\cdot, \cdot, \cdot \mid x_1, a_1) \) denote the joint conditional distribution of
\( \{m_Y(X_1, A_1, X_2), c_Y(X_1, A_1, X_2), f_2(X_1, A_1, X_2)\} \) given \( X_1 = x_1 \) and \( A_1 = a_1 \), so that

\[ G_Y(s, t, w \mid x_1, a_1) = \int \mathbb{1}_{m_Y(x_1, a_1, x_2) \leq s} \mathbb{1}_{c_Y(x_1, a_1, x_2) \leq t} \mathbb{1}_{f_2(x_1, a_1, x_2) \leq w} dF_{X_2 \mid X_1, A_1}(x_2 \mid x_1, a_1). \]

Then, one can show by interchanging the order of integration that

\[ \int \int F_{c_Y} \{y - s - \text{sgn}(w)t\} dG_Y \{s, t, w \mid x_1, \pi_1(x_1)\} dF_{X_1}(x_1) \]

\[ = \int \int J^Y_Y(x_1, x_2, y) dF_{X_2 \mid X_1, A_1 \{x_2 \mid x_1, \pi_1(x_1)\}} dF_{X_1}(x_1), \]

where the right side is equivalent to \( \text{pr}^X_Y(y) \). Thus,

\[ \text{pr}^X_Y(y) = E \left[ \int F_{c_Y} \{y - s - \text{sgn}(w)t\} dG_Y \{s, t, w \mid X_1, \pi_1(X_1)\} \right], \quad (4.3) \]

where the expectation is over the marginal distribution of \( X_1 \). Similarly, it can be shown that

\[ \text{pr}^Z_Y(z) = E \left[ \int F_{c_Z} \{z - u - \text{sgn}(w)v\} dG_Z \{u, v, w \mid X_1, \pi_1(X_1)\} \right], \quad (4.4) \]

where \( F_{c_Z} (\cdot) \) is the marginal distribution function of \( c_Z \), and \( G_Z(\cdot, \cdot, \cdot \mid x_1, a_1) \) is the joint conditional distribution of \( \{m_Z(X_1, A_1, X_2), c_Z(X_1, A_1, X_2), f_2(X_1, A_1, X_2)\} \) given \( X_1 = x_1 \) and \( A_1 = a_1 \).

Define

\[ \mathbb{1}_{x_1, a_1, x_2}(s, t, u, v, w) \]

\[ = \mathbb{1}_{m_Y(x_1, a_1, x_2) \leq s} \mathbb{1}_{c_Y(x_1, a_1, x_2) \leq t} \mathbb{1}_{m_Z(x_1, a_1, x_2) \leq u} \mathbb{1}_{c_Z(x_1, a_1, x_2) \leq v} \mathbb{1}_{f_2(x_1, a_1, x_2) \leq w} \]

and

\[ G_{Y,Z}(s, t, u, v, w \mid x_1, a_1) = \int \mathbb{1}_{x_1, a_1, x_2}(s, t, u, v, w) dF_{X_2 \mid X_1, A_1}(x_2 \mid x_1, a_1). \]
Thus, $G_{Y,Z}(s, t, u, v, w \mid x_1, a_1)$ is the joint conditional distribution of

$$\{m_Y(X_1, A_1, X_2), c_Y(X_1, A_1, X_2), m_Z(X_1, A_1, X_2), c_Z(X_1, A_1, X_2), f_2(X_1, A_1, X_2)\}$$

given $X_1 = x_1$ and $A_1 = a_1$. The joint cumulative distribution of $(Y, Z)$ under regime $\pi$ and evaluated at $(y, z)$ is

$$pr^\pi_{Y,Z}(y, z) = E \left[ \int F_{\epsilon_Y, \epsilon_Z}(y, z, s, t, u, v, w) dG_{Y,Z}(s, t, u, v, w \mid X_1, \pi_1(X_1)) \right],$$  \hspace{0.5cm} (4.5)

where we define $F_{\epsilon_Y, \epsilon_Z}(y, z, s, t, u, v, w) = F_{\epsilon_Y, \epsilon_Z}(y - s - \text{sgn}(w)t, z - u - \text{sgn}(w)v)$ in the previous expression, and the outer expectation is with respect to $X_1$.

Expressions (4.3), (4.4), and (4.5) are similar in form to (3.4) in the development of TIQ-learning in Section 3.2.2. This suggests that estimating $pr_\pi^Y(y)$, $pr_\pi^Z(z)$, and $pr^\pi_{Y,Z}(y, z)$ can be accomplished using steps similar to those described in the TIQ-learning algorithm. Let $\tilde{F}_{\epsilon_Y}(\cdot), \tilde{F}_{\epsilon_Z}(\cdot), \tilde{F}_{\epsilon_Y, \epsilon_Z}(\cdot), \tilde{G}_Y(\cdot, \cdot \mid x_1, a_1), \tilde{G}_Z(\cdot, \cdot \mid x_1, a_1)$, and $\tilde{G}_Y(\cdot, \cdot, \cdot \mid x_1, a_1)$ denote estimators of the corresponding true distribution functions. Let $E_n$ denote empirical expectation with respect to $X_1$. Then, for a given $\pi_1$, the following expressions are estimators of $pr^\pi_Y(y)$, $pr^\pi_Z(z)$, and $pr^\pi_{Y,Z}(y, z)$:

$$\hat{pr}^\pi_Y(y) = E_n \int \tilde{F}_{\epsilon_Y}(y - s - \text{sgn}(w)t) d\tilde{G}_Y(s, t, w \mid X_1, \pi_1(X_1)),$$  \hspace{0.5cm} (4.6)

$$\hat{pr}^\pi_Z(z) = E_n \int \tilde{F}_{\epsilon_Z}(z - u - \text{sgn}(w)v) d\tilde{G}_Z(u, v, w \mid X_1, \pi_1(X_1)),$$  \hspace{0.5cm} (4.7)

$$\hat{pr}^\pi_{Y,Z}(y, z) = E_n \int \tilde{F}_{\epsilon_Y, \epsilon_Z}(y, z, s, t, u, v, w) d\tilde{G}_Z(s, t, u, v, w \mid X_1, \pi_1(X_1)).$$  \hspace{0.5cm} (4.8)

**Approaches to estimating $G_Y$, $G_Z$, and $G_{Y,Z}$**

We assume $S$-constrained decision rules at stage two can be expressed as $\pi_{2,S}(h_2) = \text{sgn}(f_2(h_2))$ for some smooth function $f_2(\cdot)$, where we define $\text{sgn}(x) = 1_{x \geq 0} - 1_{x < 0}$. Thus, we arbitrarily assign treatment 1 at the second stage when $f_2(h_2) = 0$.

Recall from Section 4.2.1 that $H_2$ denotes the full history prior to treatment assignment at stage two, i.e., $H_2 = (X_1^T, A_1, X_2^T)^T$. Our strategy for modeling $G_{Y,Z}(\cdot, \cdot, \cdot \mid x_1, a_1)$ is to model the joint distribution of the standardized residuals,

$$e^m_Y = \frac{m_Y(H_2) - \mu_Y^m(X_1, A_1)}{\sigma_Y^m(X_1, A_1)},$$

$$e^m_Z = \frac{m_Z(H_2) - \mu_Z^m(X_1, A_1)}{\sigma_Z^m(X_1, A_1)},$$

$$e^{f_2} = \frac{f_2(H_2) - \mu_{f_2}(X_1, A_1)}{\sigma_{f_2}(X_1, A_1)},$$

$$e^c_Y = \frac{c_Y(H_2) - \mu_Y^c(X_1, A_1)}{\sigma_Y^c(X_1, A_1)},$$

$$e^c_Z = \frac{c_Z(H_2) - \mu_Z^c(X_1, A_1)}{\sigma_Z^c(X_1, A_1)},$$

$$e^{f_2} = \frac{f_2(H_2) - \mu_{f_2}(X_1, A_1)}{\sigma_{f_2}(X_1, A_1)},$$

$$e^{f_2} = \frac{f_2(H_2) - \mu_{f_2}(X_1, A_1)}{\sigma_{f_2}(X_1, A_1)},$$

(4.9)
that are obtained from mean and variance modeling of $m_Y(H_2)$, $c_Y(H_2)$, $m_Z(H_2)$, $c_Z(H_2)$, and $f_2(H_2)$ given $X_1$ and $A_1$ (Carroll and Ruppert, 1988). In (4.9), the mean functions are defined as

$$
\mu^m_Y(X_1, A_1) = E\{m_Y(H_2) \mid X_1, A_1\}, \quad \mu^c_Y(X_1, A_1) = E\{c_Y(H_2) \mid X_1, A_1\},
$$

$$
\mu^m_Z(X_1, A_1) = E\{m_Z(H_2) \mid X_1, A_1\}, \quad \mu^c_Z(X_1, A_1) = E\{c_Z(H_2) \mid X_1, A_1\},
$$

$$
\mu_{f_2}(X_1, A_1) = E\{f_2(H_2) \mid X_1, A_1\},
$$

and the standard deviation functions are defined as

$$
\sigma^m_Y(X_1, A_1) = E\{[m_Y(H_2) - \mu^m_Y(X_1, A_1)]^2 \mid X_1, A_1\}^{1/2},
$$

$$
\sigma^c_Y(X_1, A_1) = E\{[c_Y(H_2) - \mu^c_Y(X_1, A_1)]^2 \mid X_1, A_1\}^{1/2},
$$

$$
\sigma^m_Z(X_1, A_1) = E\{[m_Z(H_2) - \mu^m_Z(X_1, A_1)]^2 \mid X_1, A_1\}^{1/2},
$$

$$
\sigma^c_Z(X_1, A_1) = E\{[c_Z(H_2) - \mu^c_Z(X_1, A_1)]^2 \mid X_1, A_1\}^{1/2},
$$

$$
\sigma_{f_2}(X_1, A_1) = E\{[f_2(H_2) - \mu_{f_2}(X_1, A_1)]^2 \mid X_1, A_1\}^{1/2}.
$$

Our strategy for modeling $G_Y(\cdot, \cdot, \cdot \mid x_1, a_1)$ is to model the joint conditional distribution of a subset of the standardized residuals, $(e^m_Y, e^c_Y, e_{f_2})$. If the constraints of the optimization problem are separable, then we would forego modeling the joint distribution of (4.9) for modeling the joint distribution of $(e^m_Y, e^c_Y, e_{f_2})$ to obtain an estimator of $G_Z(\cdot, \cdot, \cdot \mid x_1, a_1)$ instead of $G_Y(\cdot, \cdot, \cdot \mid x_1, a_1)$.

In practice, clinical data are often expensive to obtain, and thus sample sizes are moderate or small. It may be useful to consider parametric models for $m_Y(H_2)$, $c_Y(H_2)$, $m_Z(H_2)$, $c_Z(H_2)$, and $f_2(H_2)$, as well as their mean and variance functions. For example, the following steps may be used to estimate $m_Y(H_2) = m_Y(H_2; \tilde{\beta}_Y^m), \mu^m_Y(X_1, A_1) = \mu^m_Y(X_1, A_1; \theta_Y^m) \text{ and } \sigma_Y^m(X_1, A_1) = \sigma_Y^m(X_1, A_1; \gamma_Y^m)$, thus obtaining estimated versions of the standardized residuals.

1. Regress $Y$ on $H_2$ and $A_2$ to obtain $\tilde{\beta}^m_Y$ and resulting estimator $m_Y(H_2; \tilde{\beta}_Y^m)$ of $m_Y(H_2)$.

2. Compute $\tilde{\theta}_Y = \arg\min \sum_{i=1}^n \{m_Y(H_2; \tilde{\beta}_Y^m) - \mu^m_Y(X_1, A_1; \theta_Y^m)\}^2$ and resulting estimator $\mu_Y^m(X_1, A_1; \tilde{\theta}_Y^m)$ of $\mu_Y^m(X_1, A_1)$.

3. Using the estimated mean function, $\mu_Y^m(X_1, A_1; \tilde{\theta}_Y^m)$ from Step 2, compute

$$
\tilde{\gamma}_Y^m = \arg\min_{\gamma^m_Y} \sum_{i=1}^n \left( \log \left( \left\{ m(H_2; \tilde{\beta}_Y^m) - \mu_Y^m(X_1, A_1; \tilde{\theta}_Y^m) \right\}^2 \right) - \log \left( \sigma_Y^m(X_1, A_1; \gamma_Y^m)^2 \right) \right)^2.
$$

Then, $\mu_Y^m(X_1, A_1; \tilde{\gamma}_Y^m)$ is the resulting estimator of $\mu_Y^m(X_1, A_1)$.
4. Define the estimated standardized residuals as

\[ \hat{\epsilon}_Y, i = \frac{m_Y(H_2; \hat{\theta}_Y) - \mu_Y(X_{1i}, A_{1i}; \hat{\theta}_Y)}{\sigma_Y(X_{1i}, A_{1i}; \hat{\gamma}_Y)}, \]

for \( i = 1, \ldots, n. \)

Steps 1 – 4 can be repeated with minor adjustments to obtain the remaining estimated standardized residuals \( \{\hat{\epsilon}_Y, i, \hat{\epsilon}_Z, i, \hat{\epsilon}_1, i\}. \) One convenient strategy for modeling the joint distribution of the standardized residuals is to use a copula; for example, Chapter 3 we use a Gaussian copula. Random samples from the estimated copula can be transformed back to samples from \( \hat{G}_Y(\cdot, \cdot, \cdot | x_1, a_1), \hat{G}_Z(\cdot, \cdot, \cdot | x_1, a_1), \) or \( \hat{G}_{Y,Z}(\cdot, \cdot, \cdot | x_1, a_1) \) using the estimated mean and variance functions.

**Approaches to estimating** \( F_{e_Y}, F_{e_Z}, \) and \( F_{e_Y,e_Z} \)

Estimating the marginal distribution functions of \( e_Y \) and \( e_Z \) follows from our discussion in 3.2.4. One way to extend this approach to model \( F_{e_Y,e_Z}(\cdot, \cdot) \) is by using a mean-zero multivariate normal distribution with covariance \( \Sigma_{Y,Z}. \) Then \( \Sigma_{Y,Z} \) can be estimated using \( \hat{\Sigma}_{Y,Z}, \) the sample covariance matrix of observed residuals from the regression of \( Y \) on \( (H_2, A_2) \) and the regression of \( Z \) on \( (H_2, A_2). \)

**Sampling from estimated marginal distribution functions** \( \hat{p}_{r_Y} \) and \( \hat{p}_{r_Z} \)

Below, we provide an algorithm for sampling from \( \hat{p}_{r_Y}(y), \) assuming \( \hat{F}_{e_Y}(\cdot) \) and \( \hat{G}_{Y}(\cdot, \cdot, \cdot | x_1, a_1) \) have already been estimated. Sampling from \( \hat{p}_{r_Z}(z) \) follows with minor adjustments.

**S1.** Use the observed responses \( Y_i, i = 1, \ldots, n, \) to estimate the percentiles, 1%, 2%, \ldots, 99%, of the marginal distribution of \( Y. \) Denote estimated percentile \( j \) by \( \hat{y}_j. \)

**S2.** For each percentile, evaluate \( \hat{p}_{r_Y}(\hat{y}_j) \) by computing

\[ \hat{p}_{r_Y}(\hat{y}_j) = E_n \int \hat{F}_{e_Y}(\hat{y}_j - s - \text{sgn}(w)t) d\hat{G}_Z(s, t, w | X_1, \pi_1(X_1)). \]

**S3.** Randomly sample from a multinomial distribution with bin probabilities equal to the estimated probabilities that \( Y \) is between percentile \( j \) and \( j + 1, \) for each \( j. \) Take each sample observation to be the midpoint of the two adjacent percentiles that were used to obtain the corresponding bin probability.

**Sampling from estimated joint distribution function** \( \hat{p}_{r_{Y,Z}} \)

Our method for sampling from estimated joint distribution function \( \hat{p}_{r_{Y,Z}}(y, z) \) is an extension of the previous algorithm. The algorithm is as follows.
S1. Use the observed responses $Y_i$ and $Z_i$, $i = 1, \ldots, n$, to estimate the percentiles, $1\%, 2\%, \ldots, 99\%$, of the marginal distributions of both $Y$ and $Z$. Denote the estimated $j^{th}$ percentiles by $\hat{y}_j$ and $\hat{z}_j$.

S2. For each pair of percentiles, $(\hat{y}_j, \hat{z}_j)$, evaluate $\hat{p}_{Y,Z}(\hat{y}_j, \hat{z}_j)$ by computing

$$
\hat{p}_{Y,Z}(\hat{y}_j, \hat{z}_j) = E_n \int \hat{F}_{Y,Z}(s, u, v, w \mid X_1, \pi_1(X_1)) \, ds \, du \, dv \, dw.
$$

S3. Compute the estimated probability that $(Y, Z)$ under regime $\pi$ lies in the rectangle between marginal percentiles $j$ and $j + 1$ of $Y$ and marginal percentiles $j'$ and $j' + 1$ of $Z$ for each $(j, j')$ pair. Note that this probability can be calculated as

$$
\hat{p}_{Y,Z}(\hat{y}_{j+1}, \hat{z}_{j'+1}) - \hat{p}_{Y,Z}(\hat{y}_j, \hat{z}_{j'+1}) - \hat{p}_{Y,Z}(\hat{y}_{j+1}, \hat{z}_{j'}) + \hat{p}_{Y,Z}(\hat{y}_j, \hat{z}_j).
$$

S4. Randomly sample from a multinomial distribution with bin probabilities equal to the estimated probabilities that $(Y, Z)$ lies in each $(j, j + 1)$ by $(j', j' + 1)$ rectangle, and take each sample observation to be the center of the corresponding rectangle.

After obtaining samples from an appropriate sampling algorithm, distributional summaries are estimated using the corresponding sample statistic. For example, the sample mean of samples from $\hat{p}_{Y}(y)$ would be used to estimate $M(\hat{p}_{Y}(y)) = \int y \, d\hat{p}_{Y}(y)$.

### 4.3 Future work

Thus far we have discussed techniques for estimating the functionals $M$ and $S$ in the constrained optimization problem displayed in (4.1). It remains to develop a computationally efficient algorithm for optimizing the estimated version of the problem shown in (4.2). When $X_1$ and $X_2$ are low dimensional, say one or two patient characteristics contained in each, and attention is restricted to linear decision rules of the form $\pi_1(x_1) = \text{sgn}(x_1^T \beta_1)$ and $\pi_2(x_2) = \text{sgn}(x_2^T \beta_2)$, searching over a grid of fixed parameters $\beta^* = (\beta_1^*, \beta_2^*)$ is one way to estimate $\pi_{S_{\text{opt}}}(y)$. Each fixed $\beta^*$ completely determines a regime $\pi$. For each $\beta^*$ in the grid, the algorithm would first check feasibility of the regime by estimating the constraints $S(\hat{p}_{Y,Z}(y))$. If the constraints are not satisfied, the algorithm would move on to the next $\beta^*$ in the grid search. If the constraints are satisfied, the algorithm would proceed by estimating the objective function, $M(\hat{p}_{Y}(y))$, and storing the result. If one exists, the estimated solution to (4.2) is the $\pi \in \Pi$ corresponding to the $\beta^*$ that gives the maximal estimated objective function.

As the dimension of the patient histories increases, using a grid search to find a solution to (4.2) becomes computationally infeasible. Note that our objective function and constraints in the formulation...
of the problem are estimated from data and therefore potentially noisy versions of the true functions. Thus, our problem is a member of a class known as stochastic optimization problems (for a review of stochastic optimization, see Fouskakis and Draper, 2002). We conjecture that existing literature on stochastic programming algorithms will offer insight on how best to proceed when $X_1$ and $X_2$ are of moderate or high dimension.
REFERENCES


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


Appendix A

Supplementary Material for
IQ-learning

A.1 Inconsistency of Q-Learning

From Chapter 2, the closed-form expression in (2.7) facilitates study of nonlinearity introduced by the nonsmooth maximization operator and resulting inconsistency of Q-learning. To this end, suppose

\[ Q_2(h_2, a_2) = h_{2,0}^T \beta_{2,0}^* + a_2 h_{2,1}^T \beta_{2,1}^* \]

so that the second-stage Q-function is correctly specified, and thus

\[ \max_{a_2 \in \{-1, 1\}} Q_2(H_2, a_2) = H_{2,0}^T \beta_{2,0}^* + H_{2,1}^T \beta_{2,1}^* \].

Consider the coefficient indexing the best-fitting linear model to the first-stage Q-function,

\[ \beta_1^* = \arg \min_{\beta_{1,0}, \beta_{1,1}} E \left\{ \max_{a_2 \in \{-1, 1\}} Q_2(H_2, a_2) - H_{1,0}^T \beta_{1,0} - A_1 H_{1,1}^T \beta_{1,1} \right\}^2, \]

so that

\[ \beta_1^* = \Sigma_1^{-1} E B_1 \left( H_{2,0}^T \beta_{2,0}^* + H_{2,1}^T \beta_{2,1}^* \right), \]

where

\[ B_1 = (H_{1,0}^T, A_1 H_{1,1}^T)^T \]

and

\[ \Sigma_1 = E B_1 B_1^T. \]

If

\[ m(H_2) = B_1^T \gamma + \rho \]

where \( \rho \) is a mean-zero random variable which is independent of patient histories and outcomes, then

\[ \beta_1^* = \gamma + \Sigma_1^{-1} E B_1 \left( H_{2,0}^T \beta_{2,0}^* + H_{2,1}^T \beta_{2,1}^* \right), \]

and for

\[ b_1 = (h_{1,0}^T, a_1 h_{1,1}^T)^T \]

it follows that

\[ b_1^T \beta_1^* = E \{ m(H_2) | h_1, a_1 \} + b_1^T \Sigma_1^{-1} E B_1 | H_{2,1}^T \beta_{2,1}^* \].

In addition, suppose that

\[ H_{2,1}^T \beta_{2,1}^* = B_1^T \eta + \nu, \]

where

\[ \nu \]

is a mean-zero normal random variable with variance \( \sigma^2 \) that is independent of patient histories and outcomes. Then

\[ b_1^T \beta_1^* - E \{ m(H_2) | h_1, a_1 \} \]

is equal to

\[ b_1^T \Sigma^{-1} E B_1 \left[ B_1^T \eta \left\{ 1 - 2 \Phi \left( \frac{-B_1^T \eta}{\sigma} \right) \right\} + \sqrt{\frac{2\sigma^2}{\pi}} \exp \left\{ \frac{-(B_1^T \eta)^2}{2\sigma^2} \right\} \right], \]
which can be re-expressed as $E \left( \left| H_{I2,1}^T \beta_{I2,1}^* \right| \left| H_1 = h_1, A_1 = a_1 \right. \right) + r(h_1, a_1)$, where

$$r(h_1, a_1) = 2b_1^I \Sigma^{-1} E B_1 B_1^I \eta \left\{ \phi \left( \frac{-b_1^I \eta}{\sigma} \right) - \phi \left( \frac{-B^I \eta}{\sigma} \right) \right\} + \sqrt{\frac{2\sigma^2}{\pi}} \left[ b_1^I \Sigma^{-1} E B_1 \exp \left\{ \frac{-(B^I \eta)^2}{2\sigma^2} \right\} - \exp \left\{ \frac{-(b_1^I \eta)^2}{2\sigma^2} \right\} \right]$$

is a remainder term that shows how far the optimal linear approximation is from the truth, $E \left( \left| H_{I2,1}^T \beta_{I2,1}^* \right| \left| H_1 = h_1, A_1 = a_1 \right. \right) = E \left\{ \left| \Delta(H_2) \right| \left| H_1 = h_1, A_1 = a_1 \right. \right\}$. The remainder is identically zero if $\eta = 0$ and $\sigma = 0$, the case of no second-stage treatment effect with probability one, i.e., $\mathrm{pr}(H_{I2,1}^T \beta_{I2,1}^* = 0) = 1$. The remainder is close to zero when the distribution of $B^I \eta / \sigma$ is concentrated sufficiently far from zero and $b_1^I \eta / \sigma$ is also far from zero. The remainder term is largest for small to moderate values of $B^I \eta / \sigma$. This is relevant, as in many applications we do not expect large signal-to-noise ratios. Thus, even under simple generative models like the one described above, the $Q$-learning algorithm with its linear approximations need not be even approximately consistent.

### A.2 Proofs of Asymptotic Results

Let $L^\infty(\mathcal{F})$ denote the space of uniformly bounded real-valued functions on $\mathcal{F}$ equipped with the supremum norm. Write $Z_n = n^{1/2} (\Delta_L, \Delta_\mu, \Delta_\sigma)$. Then by (A1N), $Z_n$ converges in distribution to Normal $\{0, \Sigma_N(h_1, a_1)\}$. Similarly, define $W_n = n^{1/2} (\Delta_L, \Delta_\theta, \Delta_\gamma, \Delta_\beta, \Delta_\xi)$. Then by (A1E) $W_n$ converges in distribution to Normal $\{0, \Sigma_E(h_1, a_1)\}$. For convenience we abbreviate $\mu(h_1, a_1; \theta)$, $\sigma(h_1, a_1; \gamma)$, $L(h_1, a_1; \alpha)$, and $\xi(H_2, H_1, A_1; \theta, \gamma, \beta_2)$ as $\mu, \sigma, L$, and $\xi$, respectively, and write $\hat{\mu}, \hat{\sigma}, \hat{L},$ and $\hat{\xi}$ as shorthand for $\mu(h_1, a_1; \hat{\theta}), \sigma(h_1, a_1; \hat{\gamma}), L(h_1, a_1; \hat{\alpha}),$ and $\xi(H_2, H_1, A_1; \hat{\theta}, \hat{\gamma}, \hat{\beta}_2)$. Similarly, we write $m^*, \sigma^*, L^*$, and $\xi^*$ as shorthand for $m(h_1, a_1; \theta^*), \sigma(h_1, a_1; \gamma^*), L(h_1, a_1; \alpha^*),$ and $\xi(H_2, H_1, A_1; \theta^*, \gamma^*, \beta_2^*)$.

**Proof of Theorem 1, Part 1.** Notice that

$$n^{1/2} \left\{ \tilde{Q}^T \left( Q \right)^N(h_1, a_1) - L(h_1, a_1; \alpha^*) - \frac{1}{\sigma^*} \int |z| \phi \left( \frac{z - \mu^*}{\sigma^*} \right) dz \right\} = n^{1/2} \left\{ I(\hat{L}, \hat{\mu}, \hat{\sigma}) - I(L^*, \mu^*, \sigma^*) \right\},$$

where $I(\cdot)$ is as defined immediately preceding Theorem 1 in the main paper. Inspection reveals that $\nabla I(L, \mu, \sigma)$ exists and is continuous in a neighborhood of $(L^*, \mu^*, \sigma^*)$. Hence, by a first-order Taylor series approximation, the right hand side above is equal to

$$n^{1/2} \left\{ I(\hat{L}, \hat{\mu}, \hat{\sigma}) - I(L^*, \mu^*, \sigma^*) \right\} = \nabla I(L^*, \mu^*, \sigma^*)^T Z_n + o_P(1).$$
The result follows from Slutsky’s lemma.

**Proof of Theorem 1, Part 2.** The proof proceeds by showing that

\[
n^{1/2} \left\{ \tilde{Q}_1^{IQ,E}(h_1, a_1) - L(h_1, a_1; \alpha^*) - \frac{1}{\sigma^*} \int |z| \kappa \left( \frac{z - \mu^*}{\sigma^*} \right) dz \right\} = \{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^T, 1\} W_n + o_P(1). \quad \text{(A.1)}
\]

The term on the left hand side of the above display equals

\[
n^{1/2} E_n |\tilde{\mu} + \tilde{\xi}| - n^{1/2} E|\mu^* + \sigma^* \xi^*| + W_{n,1}.
\]

The first two terms in the above display are equal to

\[
n^{1/2}(E_n - E)|\tilde{\mu} + \tilde{\xi}| + n^{1/2} E \left(|\tilde{\mu} + \tilde{\xi}| - |\mu^* + \sigma^* \xi^*|\right).
\]

From (A2), it follows that \(n^{1/2}(E_n - E)\) converges weakly to \(G_\infty\) in \(l^\infty(F)\), where \(G_\infty\) is a mean-zero Gaussian process with covariance function \(\text{Cov}\{G_\infty(f), G_\infty(g)\} = E(f - Ef)(g - Eg)\) (see, for example, Kosorok, 2008). Note that by the second part of (A2), the foregoing covariance function is continuous in a neighborhood of \((\theta^*, \gamma^*, \beta_2^*)\). Thus, using the equicontinuity of \(n^{1/2}(E_n - E)\), it follows that \(n^{1/2}(E_n - E)|\tilde{\mu} + \tilde{\xi}| = W_{n,5} + o_{P^*}(1)\), where \(P^*\) denotes outer probability. So far, we have shown that the right hand side of (A.1) is equal to \(n^{1/2} \left\{ E(|\tilde{\mu} + \tilde{\xi}| - |\mu^* + \sigma^* \xi^*|) \right\} + W_{n,1} + W_{n,5} + o_{P^*}(1)\).

From (A2), \(J\) is continuously differentiable in a neighborhood of \((\theta^*, \gamma^*, \beta_2^*)\). Using a first-order Taylor series approximation, we have

\[
n^{1/2} E(|\tilde{\mu} + \tilde{\xi}| - |\mu^* + \sigma^* \xi^*|) = \nabla J(\theta^*, \gamma^*, \beta_2^*)^T(W_{n,2}, W_{n,3}, W_{n,4}) + o_P(1).
\]

Thus, we have shown that the right hand side of (A.1) equals \(\{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^T, 1\} W_n + o_P(1)\). The result follows from Slutsky’s Lemma (Kosorok, 2008).

### A.3 Additional simulation results

Figures A.1 - A.3 display additional results regarding the value, \(V^\pi = E^\pi Y\), of the estimated regimes from Section 2.3. Histograms of the raw value estimates from each Monte Carlo iteration from the NormHomo IQ-learning regime, Q-learning with linear models, and Q-learning with SVR are displayed in Figures A.1 - A.3 for three values of the scaling constant, \(C\). In general, the estimated value distribution for NormHomo is shifted slightly higher than both of the Q-learning estimated value distributions. Results shown are for \(C = 0.05, 1.0, 2.0\); results were similar across other values of \(C\).
Figure A.1: Histograms of the 2,000 estimated values from each regime for $C=0.05$.

Figure A.2: Histograms of the 2,000 estimated values from each regime for $C=1.0$.

Figure A.3: Histograms of the 2,000 estimated values from each regime for $C=2.0$. 
Here we provide additional simulation results to demonstrate the robust performance of IQ-learning across a broad range of model settings. As in the main portion of the paper, the generative model is

\[
X_1 \sim \text{Normal}_p\{0.1, \Omega_{AR_1}(0.5)\}, \quad A_t \sim \text{Uniform}\{-1, 1\}, \quad t = 1, 2,
\]

\[
X_2 = (1.5 - 0.5A_1)X_1 + \zeta_A, \quad \xi,
\]

\[
Y = H_2^T\beta_{2,0} + A_2H_2^T\beta_{2,1} + \phi,
\]

where \(\Omega_{AR_1}(0.5)\) is a diagonal matrix, \(H_2 = (1, X_2^T, A_1, X_2^T, A_1X_2^T)^T\), and \(\zeta_A = \sqrt{1.5 + 0.5A_1}\). Thus, the class is indexed by the dimension \(p\), the distributions of \(\xi\) and \(\phi\), and the coefficient vectors \(\beta_{2,0}\) and \(\beta_{2,1}\). We fix the main effect parameter \(\beta_{2,0}\) and vary the second-stage treatment effect size by scaling \(\beta_{2,1}\) as follows:

\[
\beta_{2,0} = \frac{1_{2p+2}}{||1_{2p+2}||}, \quad \beta_{2,1} = C \frac{(-0.251_{p+1}^T, 1_{p+1}^T)^T}{||(-0.251_{p+1}^T, 1_{p+1}^T)||},
\]

where \(C\) ranges over a grid from 0 to 2, and \(1_d\) denotes a \(d\)-dimensional vector of 1s. In addition, we fix the theoretical \(R^2\) of the second-stage regression model by generating \(\phi \sim \text{Normal}\{0, \sigma^2_{\phi}(C)\}\), where the variance \(\sigma^2_{\phi}(C)\) depends on the scaling constant \(C\). We consider training sets of size \(n = \{250, 500\}\) and \(n = 500\) and vary the second-stage \(R^2 \in \{0.4, 0.6, 0.8\}\). Results for \(n = 250\) with \(R^2 = 0.6\) are included in Section 2.3 of the paper. Here we provide results for the remaining combinations of \(n\) and \(R^2\). We include simulations with \(\xi\) generated from a \(\text{Normal}_p\{0, I_p\}\) as well as where elements of \(\xi\) generated independently from a \(t\)-distribution with five degrees of freedom. We include results for dimension \(p = 4\), followed by results for \(p = 8\) when \(R^2 = 0.6\). In each simulation, results are based on \(M = 2,000\) Monte Carlo data sets.

Define \(H_1 = (1, X_1^T)^T\). As in Section 2.3, we consider linear working models for the mean and variance functions of the form

\[
Q_2(h_2, a_2; \beta_2) = h_2^T\beta_{2,0} + a_2h_2^T\beta_{2,1}, \quad Q_1(h_1, a_1; \beta_1) = h_1^T\beta_{1,0} + a_1h_1^T\beta_{1,1},
\]

\[
L(h_1, a_1; \alpha) = h_1^T\alpha_0 + a_1h_1^T\alpha_1, \quad \mu(h_1, a_1; \theta) = h_1^T\theta_0 + a_1h_1^T\theta_1,
\]

\[
\log\{\sigma(h_1, a_1; \gamma)\} = h_1^T\gamma_0 + a_1h_1^T\gamma_1.
\]

In Section 2.3, we considered two IQ-learning estimators: (i) normal estimator \(\hat{g}^N(\cdot; h_1, a_1)\) of the residual distribution and a restricted variance model, \(\log\{\sigma(h_1, a_1; \gamma)\} = \gamma_0 + a_1\gamma_1\), that depends only on treatment (NormHomo); and (ii) nonparametric estimator \(\hat{g}^E(\cdot; h_1, a_1)\) of the residual distribution with a log-linear variance model that depends on \(h_1\) and \(a_1\) (NonparHetero). When \(\xi \sim \text{Normal}_p\{0, I_p\}\), both these estimators are correctly specified. When the elements of \(\xi\) are generated independently from \(t_5\), only NonparHetero is correctly specified.

Figures A.4 - A.17 display the results. For all settings, the IMSE ratio of \(Q\)-learning to IQ-learning is greater than one, indicating that the IQ-learning estimators more accurately estimate the first-stage \(Q\)-function. Increasing the sample size to \(n = 500\) and specifying higher \(R^2\) values leads to the greatest
Figure A.4: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.4$; $p = 4$; $n = 250$. **Left to Right:** Ratio of average value, coded so values greater than one are favorable to $IQ$-learning; integrated mean squared error ratio of $Q_1$-function estimates, coded so values greater than one are favorable to $IQ$-learning; coverage of 95% confidence intervals for $Q_1(h_1, a_1)$; width of 95% confidence intervals for $Q_1(h_1, a_1)$. 
Figure A.5: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.4$; $p = 4$; $n = 250$.

Figure A.6: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.8$; $p = 4$; $n = 250$. 
Figure A.7: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.8$; $p = 4$; $n = 250$.

Figure A.8: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.4$; $p = 4$; $n = 500$. 82
Figure A.9: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.4$; $p = 4$; $n = 500$.

Figure A.10: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, \mathbf{I}_p)$; $R^2 = 0.6$; $p = 4$; $n = 500$. 

83
Figure A.11: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.6$; $p = 4$; $n = 500$.

Figure A.12: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.8$; $p = 4$; $n = 500$. 

84
Figure A.13: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.8$; $p = 4$; $n = 500$. 

Figure A.14: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 8$; $n = 250$. 

85
Figure A.15: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from $t_5$; $R^2 = 0.6$; $p = 8$; $n = 250$.

Figure A.16: Measures of performance of $Q$-learning vs. $IQ$-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 8$; $n = 500$. 
Figure A.17: Measures of performance of Q-learning vs. IQ-learning; components of $\xi$ independently
drawn from $t_5$; $R^2 = 0.6$; $p = 8$; $n = 500$. 

Gains in IMSE of IQ-learning compared to Q-learning. In general, coverage of 95% confidence intervals
for $Q_1$ and average value ratios seem consistent across all settings of the parameters. In particular, the
IQ-learning estimators obtain close to the 95% nominal coverage level in all settings across values of
$C$, while Q-learning suffers from poor coverage, especially for high $R^2$ values and large effect sizes.

Results in Figures A.18 and A.19 arise from generative models where only the NonparHetero
IQ-learning estimator is correctly specified. In these settings, we vary the distribution $\xi$ and substitute
$\zeta_{X_1,A_1}$ for $\zeta_{A_1}$. That is, we specify a model for $\zeta$ that depends on both the first-stage treatment
and first-stage covariates according to the relationship

$$
\zeta_{X_1,A_1} = \exp\left[\log(2)/4 + .25^T_{p-1}X_1 + A_1\{\log(2)/4 + 1^T_{p-1}X_1\}\right] .
$$

Results are based on $n = 250$ training samples, $M = 1,000$ Monte Carlo data sets, and dimension
$p = 4$. The second-stage $R^2$ is not fixed. Figure A.18 presents results from the case where the components of $\xi$ are generated independently from a Lognormal(0,1) distribution. Figure A.19 results arise
when elements of $\xi$ are drawn independently from a mixture of the Normal(4,1) and Normal(0,1) distribu-
tions. The NonparHetero IQ-learning estimator clearly outperforms NormHomo in Figures A.18
and A.19, whereas their performance is nearly indistinguishable when both are correctly specified.

In Figure A.18, we see that both IQ-learning estimators improve IMSE over Q-learning, with
Figure A.18: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from Lognormal(0, 1);

Figure A.19: Measures of performance of $Q$-learning vs. $IQ$-learning; components of $\xi$ independently drawn from a mixture of $\text{Normal}(4, 1)$ and $\text{Normal}(0, 1)$, each with probability 0.5.
NonparHetero achieving greater gains in performance. In addition, the coverage plot in Figure A.18 shows that the $IQ$-learning estimators fall short of the nominal 95% level, even though the widths of these confidence intervals are much larger than those observed in the correctly specified simulations. Coverage is still improved when compared to $Q$-learning. The NonparHetero $IQ$-learning estimator achieves the highest coverage — nearly 90% for most values of $C$. The ratio of average value is near one for both $IQ$-learning estimators, indicating little difference in the mean of the final response when treating according to $IQ$-learning versus $Q$-learning estimated regimes. The results in Figure A.19 are similar to those in Figure A.18. The NonparHetero $IQ$-learning estimator produced the lowest IMSE, however, this did not translate into any improvement over the average value of $Q$-learning. The NormHomo estimator displayed the poorest coverage in this case, but NonparHetero came close to achieving the nominal level for all values of $C$. 
Appendix B

Supplementary Material for $TIQ$- and $QIQ$-learning

B.1 $TIQ$-learning with second-stage heteroskedasticity

Here we assume

$$Y = m(H_2) + A_2 c(H_2) + \eta(H_2, A_2) \epsilon,$$  \hspace{1cm} (B.1)

where $\eta(H_2, A_2) = \exp\{r(H_2) + A_2 s(H_2)\}/2$ for functions $r(\cdot)$ and $s(\cdot)$. In addition, $E(\epsilon) = 0$, $\text{var}(\epsilon) = 1$, and $\epsilon$ is independent of $(H_2, A_2)$. Under model (B.1), the $\lambda$-optimal second-stage decision rule is

$$\pi_{2,\lambda}^{\text{TIQ}}(h_2) = \text{sgn} \left[ \frac{\lambda - m(h_2) + c(h_2)}{\exp \left( \frac{r(h_2) - s(h_2)}{2} \right)} - \frac{\lambda - m(h_2) - c(h_2)}{\exp \left( \frac{r(h_2) + s(h_2)}{2} \right)} \right].$$ \hspace{1cm} (B.2)

To see this, define

$$\text{pr}^{\pi_1, \pi_2}(Y > \lambda) = E\{E\{\text{pr}^{\pi_1, \pi_2}(Y > \lambda \mid H_2, a_2) \mid_{a_2=\pi_2(H_2)} X_1, a_1} \mid_{a_1=\pi_1(X_1)}\}. $$

$$= E \left[ E \left( \text{pr} \left[ \epsilon > \frac{\lambda - m(H_2) - \pi_2(H_2) c(H_2)}{\exp \left( \frac{r(H_2) + \pi_2(H_2) s(H_2)}{2} \right)} \right] \mid X_1, \pi_1(X_1) \right) \right].$$

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

$$\frac{\lambda - m(h_2) - \pi_2(h_2) c(h_2)}{\exp \left( \frac{r(h_2) + \pi_2(h_2) s(h_2)}{2} \right)},$$

leading to $\pi_{2,\lambda}^{\text{TIQ}}(h_2)$ in (B.2). Define $G(\cdot, \cdot, \cdot, \cdot \mid x_1, a_1)$ to be the joint conditional distribution of $\{m(h_2), c(h_2), r(h_2), s(h_2)\}$ given $(X_1 = x_1, A_1 = a_1)$. Let $F_\epsilon(\cdot)$ denote the cumulative distribution
function of $\epsilon$. The first-stage $\lambda$-optimal decision rule is

$$\pi_{1,\lambda}^{\text{TQ}}(\mathbf{x}_1) = \arg\min_{a_1} \int F_\epsilon \left( \frac{\lambda - t - \text{sgn}\{K(t, u, v, w)\} u}{\exp\{(v + \text{sgn}\{K(t, u, v, w)\} w)/2\}} \right) \times G(t, u, v, w \mid \mathbf{x}_1, a_1) \, dt \, du \, dv \, dw,$$

where

$$K(t, u, v, w) = \frac{\lambda - t + u}{\exp\{(v - w)/2\}} - \frac{\lambda - t - u}{\exp\{(v + w)/2\}}.$$

Thus, estimation of $\pi_{1,\lambda}^{\text{TQ}}$ involves specifying estimators for $F_\epsilon(\cdot)$ and the four-dimensional conditional density $G(\cdot, \cdot, \cdot, \cdot \mid \mathbf{x}_1, a_1)$. Alternatively, a suitable transformation of the response may be employed to obtain constant variance at the second stage, and then the methods described in Chapter 3 may be applied.

### B.2 TIQ-learning with patient-specific thresholds

In this section, let $H_1 = X_1$ to simplify notation. Denote the patient-specific threshold optimal second-stage rule by $\pi_{2,\lambda(h_1)}^{\text{TQ}}(h_2)$, where $t = 1$ or $t = 2$. Then, $\pi_{2,\lambda(h_1)}^{\text{TQ}}(h_2) = \pi_2^{\text{TQ}}(h_2) = \text{sgn}\{c(h_2)\}$ whether $t = 1$ or $t = 2$. To see this, note for fixed $\pi_1$,

$$\Pr[\pi_{1,2}\{Y > \lambda(H_t)\}] = E(E[\Pr[\pi_{1,2}\{Y > \lambda(H_t)\} \mid H_2, a_2] \mid_{a_2 = \pi_2(h_2)} \mid H_1, a_1] \mid_{a_1 = \pi_1(H_1)}).$$

Because $H_1 \subset H_2$, conditioning on $H_2$ reduces $\lambda(H_2)$ to a constant for each patient. Thus, the derivation of the optimal second-stage rule in the main portion of the paper applies, giving the result that $\pi_{2,\lambda(h_1)}^{\text{TQ}}(h_2) = \pi_2^{\text{TQ}}(h_2) = \text{sgn}\{c(h_2)\}$.

When the threshold depends on first-stage history $H_1$, $\lambda(h_1)$ replaces $\lambda$ in Step TIQ4 of the TIQ-learning algorithm in Section 3.2.2 and no additional modeling is needed. When the threshold depends on $H_2$, the joint conditional distribution of $\{\lambda(H_2), m(H_2), c(H_2)\}$ given $(H_1 = h_1, A_1 = a_1)$ must be estimated. Let $G(\cdot, \cdot, \cdot \mid h_1, a_1)$ denote this trivariate distribution and $\hat{G}(\cdot, \cdot, \cdot \mid h_1, a_1)$ an estimator. The estimated optimal first-stage decision rule is

$$\hat{\pi}_{1,\lambda(h_1)}^{\text{TQ}}(h_1) = \arg\min_{a_1} \int \hat{F}_\epsilon(t - u - |v|) \hat{G}(t, u, v \mid h_1, a_1) \, dt \, du \, dv.$$

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 $h_1$. 

91
### B.3 $QIQ$-learning optimal second-stage decision rule

We show the $\tau$-optimal $QIQ$-learning second-stage rule is $\pi_{QIQ}^{\tau, 2}(h_2) = \text{sgn}\{c(h_2)\}$ under the assumption of constant variance at the second-stage. Define the set

$$S_{\pi_1, \pi_2} = \{y : \Pr^{\pi_1, \pi_2}(Y \leq y) \geq \tau\},$$

so that $q^{\pi_1, \pi_2}(\tau) = \inf S_{\pi_1, \pi_2}$. In the $TIQ$-learning portion of Chapter 3, we showed

$$\Pr^{\pi_1, \pi_2}(Y > y) \leq E\left(E\left[\Pr\{\epsilon > y - m(H_2) - |c(H_2)|\} \mid X_1, a_1\right] \mid a_1 = \pi_1(x_1)\right);$$

or equivalently,

$$\Pr^{\pi_1, \pi_2}(Y \leq y) \geq E\left(E\left[\Pr\{\epsilon \leq y - m(H_2) - |c(H_2)|\} \mid X_1, a_1\right] \mid a_1 = \pi_1(x_1)\right).$$

The right side of the previous expression is $\Pr^{\pi_1, \pi_2}(Y \leq y)$ because $|c(h_2)| = \pi_2^*(h_2)c(h_2)$. Since $\Pr^{\pi_1, \pi_2}(Y \leq y) \geq \Pr^{\pi_1, \pi_2}(Y \leq y)$ for all $y$, 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, $\pi_{QIQ}^{\tau, 2}(h_2) = \pi_2^*(h_2) = \text{sgn}\{c(h_2)\}$ is optimal because this inequality holds for arbitrary $\pi_1$ and $\pi_2$.

### B.4 $QIQ$-learning toy example: $f(y^*_\tau) \neq y^*_\tau$

Suppose all subjects have the same first-stage covariates, i.e., $X_1 = x_1$ with probability one. Fix $\tau = 0.5$ and let $p(y \mid x_1, a_1)$ denote the conditional density of $Y$ given $(X_1 = x_1, A_1 = a_1)$. Suppose

$$p(y \mid x_1, 1) = \begin{cases} -2.5 \text{ with probability } 0.1 \\ -1.5 \text{ with probability } 0.2 \\ -0.5 \text{ with probability } 0.2 \\ 0.5 \text{ with probability } 0.2 \\ 1.5 \text{ with probability } 0.2 \\ 2.5 \text{ with probability } 0.1 \end{cases}$$

and

$$p(y \mid x_1, -1) = \begin{cases} \text{Uniform}(-2, 0) \text{ with probability } 0.5 \\ 0 \text{ with probability } 0.5. \end{cases}$$
Figure B.1: Cumulative distribution functions of $Y$ given $x_1$ 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 $\text{pr}(Y \leq y_\tau^* \mid x_1, a_1)$, namely $A_1 = 1$, the resulting quantile, $f(y_\tau^*) = -0.5$, is suboptimal.
Then, $f(y^*_r) < y^*_r$ because $y^*_r = 0$ and $f(y^*_r) = -1$. Recall $y^*_r = \inf \{y : \Pr^*(\cdot, y), \pi^*(Y \leq y) \geq \tau \}$ by definition. Figure B.1 provides plots of the cumulative distribution functions of $Y$ when $A_1 = -1, 1$. In this example, $f(y^*_r^-) = y^*_r$, where $y^*_r^-$ denotes the left limit of $y^*_r$. 