ABSTRACT

ZHANG, YICHI. List-based Interpretable Dynamic Treatment Regimes. (Under the direction of Dr. Eric B. Laber and Dr. Marie Davidian.)

Precision medicine is currently a topic of great interest in clinical and intervention science. It is well recognized as a path to delivering better healthcare while reducing cost, resource consumption, and treatment burden. A key component of precision medicine is that it is evidence-based, i.e., data-driven, and consequently there has been tremendous interest in estimation of precision medicine strategies using observational or randomized study data. One way to formalize precision medicine is to use a sequence of decision rules, one per stage of clinical intervention, that map up-to-date patient information to a recommended treatment. An optimal regime is defined as optimizing the mean of some cumulative clinical outcome if applied to a population of interest. It is well-known that even under simple generative models the optimal treatment regime can be a highly nonlinear function of patient information. Consequently, a focal point of recent methodological research has been the development of flexible models for estimating optimal treatment regimes. However, in many settings, estimation of an optimal treatment regime is an exploratory analysis intended to generate new hypotheses for subsequent research and not to directly dictate treatment to new patients. Also, the development of treatment regimes for application in clinical practice requires the long-term, joint effort of statisticians and clinical scientists. In such settings, an estimated treatment regime that is interpretable in a domain context may be of greater value than an unintelligible treatment regime built using "black-box" estimation methods.

We propose a simple, yet flexible class of treatment regimes whose members are representable as a short list of if-then statements. Regimes in this class can be presented either

as a paragraph or as a simple flowchart which are immediately interpretable to domain experts. In Chapter 1, we consider single-stage problems. We derive a robust estimator of the optimal regime within this class and demonstrate its finite sample performance using simulation experiments. In Chapter 2, we consider multiple-stage problems and make further methodological development. Though the discreteness of lists precludes smooth, i.e., gradient-based, methods of estimation and leads to non-standard asymptotics, we provide a computationally efficient estimation algorithm, prove consistency of the proposed estimator, and derive rates of convergence. In Chapter 3, we present the R package DTRList, which implements the proposed method. We illustrate its usage on a simulated dataset.

List-based Interpretable Dynamic Treatment Regimes

by
Yichi Zhang

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

2016

APPROVED BY:

Dr. Anastasios Tsiatis	Dr. Leonard A. Stefanski
Dr. Eric B. Laber Co-chair of Advisory Committee	Dr. Marie Davidian Co-chair of Advisory Committee

BIOGRAPHY

Yichi Zhang was born in Guangzhou, China. He obtained his Bachelor's degree in Probability and Statistics from Peking University in 2011. In the meantime, he also got a Bachelor's degree in Economics. He continued his study at North Carolina State University. He obtained his Master's degree in Statistics in 2012 on his path of pursuing a Doctoral degree. His research interests include dynamic treatment regimes and non-standard asymptotics.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Dr. Eric B. Laber, Dr. Marie Davidian and Dr. Anastasios Tsiatis for their valuable guidance during my PhD journey, which will also have an ongoing impact on my future career.

I would like to thank my committee member Dr. Leonard A. Stefanski for his insightful comments. I would also like to thank the professors in the department for their interesting and thought-provoking lectures.

I would like to thank my parents Quanxin Zhang and Yinian Liang for their constant love and support through my life.

I would like to express my love and appreciation to my wife Runchao Jiang for all the wonderful moments that we share.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes

1.1 Introduction

Treatment regimes formalize clinical decision making as a function from patient information to a recommended treatment. Proponents of personalized medicine envisage the widespread clinical use of evidence-based, i.e., data-driven, treatment regimes. The potential benefits of applying treatment regimes are now widely recognized. By individualizing treatment, a treatment regime may improve patient outcomes while reducing cost and the consumption of resources. This is important in an era of growing medical costs and an aging global population. However, the widespread integration of data-driven treatment

regimes into clinical practice is, and should be, an incremental process wherein: (i) data are used to generate hypotheses about optimal treatment regimes; (ii) the generated hypotheses are scrutinized by clinical collaborators for scientific validity; (iii) new data are collected for validation and new hypothesis generation, and so on. Within this process, it is crucial that estimated treatment regimes be interpretable to clinicians. Nevertheless, optimality, not interpretability, has been the focal point in the statistical literature on treatment regimes.

A treatment regime said to be optimal if it maximizes the expectation of a prespecified clinical outcome when used to assign treatment to a population of interest. There is a large literature on estimating optimal treatment regimes using data from observational or randomized studies. Broadly, these estimators can be categorized as regression-based or classification-based estimators. Regression-based estimators model features of the conditional distribution of the outcome given treatment and patient covariates. Examples include estimators of the regression of an outcome on covariates, treatment, and their interactions (e.g., Su et al., 2009; Qian and Murphy, 2011; Tian et al., 2014), and estimators of point treatment effects given covariates (e.g., Robins, 1994; Vansteelandt and Joffe, 2014). Regression-based methods rely on correct specification of some or all of the modeled portions of the conditional distribution of the outcome. Thus, a goal of many regression-based estimators is to ensure correct model specification under a large class of generative models (Zhao et al., 2009; Qian and Murphy, 2011; Moodie et al., 2013; Laber et al., 2014; Taylor et al., 2015). However, as flexibility is introduced into the model, interpretability tends to diminish, and in the extreme case the estimated treatment regime becomes an unintelligible black box.

Classification-based estimators, also known as policy-search or value-search estimators, estimate the marginal mean of the outcome for every treatment regime within a pre-specified class and then take the maximizer as the estimated optimal regime. Examples include marginal structural mean models (Robins et al., 2008; Orellana et al., 2010); robust marginal mean models (Zhang et al., 2012c); and outcome weighted learning (Zhang et al., 2012a; Zhao et al., 2012, 2015b). Classification-based estimators often rely on fewer assumptions about the conditional distribution of the outcome given treatment and patient information and thus may be more robust to model misspecification than regression-based estimators (Zhang et al., 2012c,a). Furthermore, because classification-based methods estimate an optimal regime within a pre-specified class, it is straightforward to impose structure on the estimated regime, e.g., interpretability, by restricting this class. We use robust marginal mean models with a highly interpretable yet flexible class of regimes to estimate a high-quality regime that can be immediately understood by clinical and intervention scientists.

To obtain an interpretable and parsimonious treatment regime, we use the concept of decision list, which was developed in the computer science literature for representing flexible but interpretable classifiers (Rivest, 1987; Clark and Niblett, 1989; Marchand and Sokolova, 2005; Letham et al., 2012; Wang and Rudin, 2015); see Freitas (2014) for a recent position paper on the importance of interpretability in predictive modeling and additional references on interpretable classifiers. As a treatment regime, a decision list comprises a sequence of "if-then" clauses that map patient covariates to a recommended treatment. Figure 1.1 shows a decision list for patients with chronic depression (see Section 4.2). This decision list recommends treatments as follows: if a patient

presents with somatic anxiety score above 1 and retardation score above 2, the list recommends nefazodone; otherwise, if the patient has Hamilton anxiety score above 23 and sleep disturbance score above 2, the list recommends psychotherapy; and otherwise the list recommends nefazodone + psychotherapy (combination). Thus, a treatment regime represented as a decision list can be conveyed as either a diagram or text and is easily understood, in either form, by domain experts. Indeed, decision lists have frequently been used to display estimated treatment regimes (Shortreed et al., 2011; Moodie et al., 2012; Shortreed et al., 2014; Laber and Zhao, 2015) or to describe theory-based, i.e., not data-driven, treatment regimes (Shiffman, 1997; Marlowe et al., 2012).

Another important attribute of a decision list is that it "short circuits" measurement of patient covariates; e.g., in Figure 1.1, the Hamilton anxiety score and sleep disturbance score do not need to be collected for patients with somatic anxiety score above 1 and retardation score above 2. This is important in settings where patient covariates are expensive or burdensome to collect (e.g., Gail et al., 1999; Gail, 2009; Baker et al., 2009; Huang et al., 2015). We provide an estimator of the treatment regime that minimizes an expected cost among all regimes that optimize the marginal mean outcome.

1.2 Methodology

1.2.1 Framework

We assume that the observed data are $\{(X_i, A_i, Y_i)\}_{i=1}^n$, which comprise n independent identically distributed observations, one for each subject in an experimental or observational study. Let (X, A, Y) denote a generic observation. Then $X \in \mathbb{R}^p$ are baseline

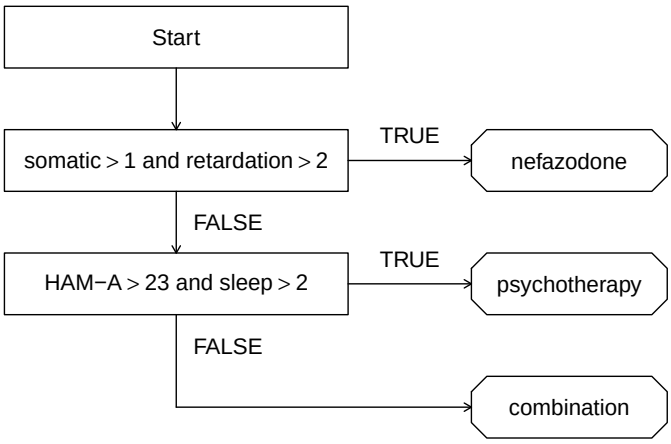


Figure 1.1 Estimated decision list for treating patients with chronic depression.

patient covariates; $A \in \mathcal{A} = \{1, \ldots, m\}$ is the treatment assigned; and $Y \in \mathbb{R}$ is the outcome, coded so that higher values are better. A treatment regime, π, is a function from $\mathbb{R}^p$ into $\mathcal{A}$, so that under π a patient presenting with $X = x$ is recommended treatment $\pi(x)$.

The value of a regime π is the expected outcome if all patients in the population of interest are assigned treatment according to π. To define the value, we use the set of potential outcomes $\{Y^*(a)\}_{a \in \mathcal{A}}$, where $Y^*(a)$ is the outcome that would be observed if a subject were assigned treatment a. Define $Y^*(\pi) = \sum_{a \in \mathcal{A}} Y^*(a) I\{\pi(X) = a\}$ to be the potential outcome under regime π, and $R(\pi) = \mathrm{E}\{Y^*(\pi)\}$ to be the value of regime π. An optimal regime, say π^{opt}, satisfies $R(\pi^{\mathrm{opt}}) \geq R(\pi)$ for all π. Let Π denote a class of regimes of interest. Classification-based estimation methods form an estimator of $R(\pi)$, say $\widehat{R}(\pi)$, and then estimate π^{opt} using $\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{R}(\pi)$. The success of this approach requires: (i) a high-quality estimator of $R(\pi)$; (ii) a sufficiently rich class Π; and (iii) an efficient

algorithm for maximizing $\widehat{R}(\pi)$ over Π. We discuss these topics in the next three sections.

1.2.2 Estimation of $R(\pi)$

We make several standard assumptions: (A1) consistency: $Y = Y^*(A)$; (A2) no unmeasured confounders: $\{Y^*(a)\}_{a\in\mathcal{A}}$ are conditionally independent of A given X; and (A3) positivity: there exists $\delta > 0$ so that $\Pr(A = a|X) \geq \delta$ for all $a \in \mathcal{A}$. Assumption (A2) is automatically satisfied in a randomized study but is untestable in observational studies (Robins et al., 2000). Under (A1)–(A3), it can be shown (Tsiatis, 2006) that

$$R(\pi) = \mathrm{E}\left(\sum_{a=1}^{m}\left[\frac{I(A=a)}{\omega(X,a)}\left\{Y-\mu(X,a)\right\}+\mu(X,a)\right]I\left\{\pi(X)=a\right\}\right), \qquad (1.1)$$

where $\omega(x,a) = \Pr(A = a|X = x)$ and $\mu(x,a) = \mathrm{E}(Y|X = x, A = a)$. Alternate expressions for $R(\pi)$ exist (Zhang et al., 2012a); however, estimators based on (1.1) possess a number of desirable properties (see below).

To construct an estimator of $R(\pi)$ from (1.1) we replace $\omega(x,a)$ and $\mu(x,a)$ with estimated working models and replace the expectation with its sample analog. If treatment is randomly assigned independently of subject covariates, then $\omega(x,a)$ can be estimated by $n^{-1}\sum_{i=1}^{n} I(A_i = a)$. Otherwise, we posit a multinomial logistic regression model of the form $\omega(x,a) = \exp\left(u^{\mathrm{T}}\gamma_a\right)\big/\left\{1+\sum_{j=1}^{m-1}\exp\left(u^{\mathrm{T}}\gamma_j\right)\right\}$, $a = 1,\ldots,m-1$, where $u = u(x)$ is a known feature vector, and $\gamma_1,\ldots,\gamma_{m-1}$ are unknown parameters. Let $\widehat{\omega}(x,a)$ denote the maximum likelihood estimator of $\omega(x,a)$, where $\gamma_1,\ldots,\gamma_{m-1}$ are replaced by maximum likelihood estimators $\widehat{\gamma}_1,\ldots,\widehat{\gamma}_{m-1}$. We posit a generalized linear model for $\mu(x,a)$, $g\{\mu(x,a)\} = z^{\mathrm{T}}\beta_a$, where $g(\cdot)$ is a known link function, $z = z(x)$ is a known feature vector

constructed from x, and $\beta_1, \ldots, \beta_m$ are unknown parameters. We use $\widehat{\mu}(x,a) = g^{-1}(z^{\mathrm{T}}\widehat{\beta}_a)$ as our estimator of $\mu(x,a)$, where $\widehat{\beta}_1, \ldots, \widehat{\beta}_m$ are the maximum likelihood estimators of $\beta_1, \ldots, \beta_m$.

Given estimators $\widehat{\omega}(x,a)$ and $\widehat{\mu}(x,a)$, an estimator of $R(\pi)$ based on (1.1) is

$$\widehat{R}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m}\left[\frac{I(A_i = a)}{\widehat{\omega}(X_i,a)}\left\{Y_i - \widehat{\mu}(X_i,a)\right\} + \widehat{\mu}(X_i,a)\right] I\{\pi(X_i) = a\}. \tag{1.2}$$

For any fixed π, $\widehat{R}(\pi)$ is doubly robust in the sense that it is a consistent estimator of $R(\pi)$ if either the model for $\omega(x,a)$ or $\mu(x,a)$ is correctly specified (Tsiatis, 2006; Zhang et al., 2012c). As a direct consequence, $\widehat{R}(\pi)$ is guaranteed to be consistent in a randomized study, as $\omega(x,a)$ is known by design. Furthermore, if both models are correctly specified, then $\widehat{R}(\pi)$ is semiparametric efficient; i.e., it has the smallest asymptotic variance among the class of regular, asymptotically linear estimators (Tsiatis, 2006).

1.2.3 Regimes Representable as Decision Lists

Gail and Simon (1985) present an early example of a treatment regime using data from the NSABP clinical trial. The treatment regime they propose is

> **If** age $\leq$ 50 and PR $\leq$ 10 **then** chemotherapy alone;
>
> **else** chemotherapy with tamoxifen,

where age (in years) denotes the age of the patient and PR denotes the progesterone receptor level (in fmol). The simple if-then structure of the foregoing treatment regime makes it immediately interpretable.

Formally, a treatment regime, π, that is representable as a decision list of length L is described by $\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}$, where c_j is a logical condition that is true or false for each $x \in \mathbb{R}^p$, and $a_j \in \mathcal{A}$ is a recommended treatment, $j = 0, \ldots, L$. As a special case, $L = 0$ is allowed. The corresponding treatment regime $\{a_0\}$ gives the same treatment a_0 to every patient. Hereafter, let Π denote the set of regimes that are representable as a decision list. Clearly, the regime proposed by Gail and Simon (1985) is a member of Π.

Define $\mathcal{T}(c_j) = \{x \in \mathbb{R}^p : c_j \text{ is true for } x\}$, $j = 1, \ldots, L$; $\mathcal{R}_1 = \mathcal{T}(c_1)$, $\mathcal{R}_j = \{\cap_{\ell<j}\mathcal{T}(c_\ell)^c\} \bigcap \mathcal{T}(c_j)$, $j = 2, \ldots, L$; and $\mathcal{R}_0 = \bigcap_{\ell=1}^{L} \mathcal{T}(c_\ell)^c$, where S^c is the complement of the set S. Then a regime $\pi \in \Pi$ can be written as $\pi(x) = \sum_{\ell=0}^{L} a_\ell I\,(x \in \mathcal{R}_\ell)$, which has structure

$$
\begin{aligned}
&\textbf{If } c_1 \textbf{ then } a_1;\\
&\textbf{else if } c_2 \textbf{ then } a_2;\\
&\textbf{...}\\
&\textbf{else if } c_L \textbf{ then } a_L;\\
&\textbf{else } a_0.
\end{aligned}
\tag{1.3}
$$

In principle, the conditions c_j, and hence the sets $\mathcal{T}(c_j)$, can be arbitrary. To ensure

parsimony and interpretability, we restrict c_j so that $\mathcal{T}(c_j)$ is one of the following sets:

$$
\begin{aligned}
&[1]\colon \{x \in \mathbb{R}^p : x_{j_1} \leq \tau_1\}, && [6]\colon \{x \in \mathbb{R}^p : x_{j_1} \leq \tau_1 \text{ or } x_{j_2} \leq \tau_2\},\\
&[2]\colon \{x \in \mathbb{R}^p : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} \leq \tau_2\}, && [7]\colon \{x \in \mathbb{R}^p : x_{j_1} \leq \tau_1 \text{ or } x_{j_2} > \tau_2\},\\
&[3]\colon \{x \in \mathbb{R}^p : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} > \tau_2\}, && [8]\colon \{x \in \mathbb{R}^p : x_{j_1} > \tau_1 \text{ or } x_{j_2} \leq \tau_2\},\\
&[4]\colon \{x \in \mathbb{R}^p : x_{j_1} > \tau_1 \text{ and } x_{j_2} \leq \tau_2\}, && [9]\colon \{x \in \mathbb{R}^p : x_{j_1} > \tau_1 \text{ or } x_{j_2} > \tau_2\},\\
&[5]\colon \{x \in \mathbb{R}^p : x_{j_1} > \tau_1 \text{ and } x_{j_2} > \tau_2\}, && [10]\colon \{x \in \mathbb{R}^p : x_{j_1} > \tau_1\},
\end{aligned}
\tag{1.4}
$$

where $j_1 < j_2 \in \{1, \ldots, p\}$ are indices and $\tau_1, \tau_2 \in \mathbb{R}$ are thresholds. We believe that the conditions that dictate the sets in (1.4), e.g., $x_{j_1} \leq \tau_1$ and $x_{j_2} \leq \tau_2$, are more easily interpreted than those dictated by linear thresholds, e.g., $\alpha_1 x_{j_1} + \alpha_2 x_{j_2} \leq \alpha_3$, as the former are more commonly seen in clinical practice.

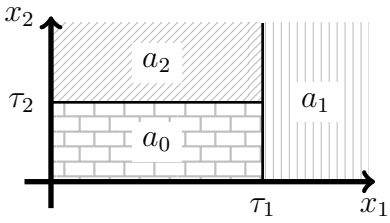
In the proposed setup, at most two variables are involved in any single condition. Having a small number of variables in each clause yields two important properties. First, the resulting treatment regime is parsimonious, and the most important variables for treatment selection are automatically identified. Second, application of the treatment regime allows for patient measurements to be taken in sequence so that the treatment recommendation can be short-circuited. For example, consider a decision list described by $\{(c_1, a_1), (c_2, a_2), a_0\}$. It is necessary to measure the variables that compose c_1 on all subjects, but the variables composing c_2 need only be measured for those who do not satisfy c_1.

Uniqueness and Minimal Cost of a Decision List

For a decision list π described by $\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}$, let $\mathcal{N}_\ell$ denote the cost of measuring the covariates required to check logical conditions $c_1, \ldots, c_\ell$. Hereafter, for simplicity, we assume that this cost is equal to the number of covariates needed to check $c_1, \ldots, c_\ell$, but it can be extended easily to a more complex cost function reflecting risk, burden, and availability. The expected cost of applying treatment rule $\pi(x) = \sum_{\ell=0}^{L} a_\ell I\left(x \in \mathcal{R}_\ell\right)$ is $N(\pi) = \sum_{\ell=1}^{L} \mathcal{N}_\ell \Pr\left(X \in \mathcal{R}_\ell\right) + \mathcal{N}_L \Pr\left(X \in \mathcal{R}_0\right)$, which is smaller than $\mathcal{N}_L = \mathcal{N}_L \sum_{\ell=0}^{L} \Pr(X \in \mathcal{R}_\ell)$, the cost of measuring all covariates in the treatment regime. This observation reflects the benefit of the short-circuit property.

A decision list π described by $\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}$ need not be unique in that there may exist an alternative decision list π' described by $\{(c'_1, a'_1), \ldots, (c'_{L'}, a'_{L'}), a'_0\}$ such that $\pi(x) = \pi'(x)$ for all x but $L \neq L'$, or $L = L'$ but $c_j \neq c'_j$ or $a_j \neq a'_j$ for some $j \in \{1, \ldots, L\}$. This is potentially important because the expected costs $N(\pi)$ and $N(\pi')$ might differ substantially. Figure 1.2 shows two representations, π and π', of the same decision list both with $L = L' = 2$ but with different clauses. The cost of the decision list in the middle panel, π, is $N(\pi) = \mathcal{N}_1 \Pr(X_1 > \tau_1) + \mathcal{N}_2 \Pr(X_1 \leq \tau_1)$, whereas the cost of the decision list in the right panel, π', is $N(\pi') = \mathcal{N}_2 \geq N(\pi)$ with strict inequality if $\mathcal{N}_2 > \mathcal{N}_1$ and $\Pr(X_1 > \tau_1) > 0$. Thus, π is preferred to π' in settings where X_2 is a biomarker that is expensive, burdenome, or potentially harmful to collect (e.g., Huang et al., 2015, and references therein).

Therefore, among all decision lists achieving the value $R(\pi^{\text{opt}})$, where π^{opt} is an optimal regime as defined previously, we seek to estimate the one that minimizes the cost. Defining $\mathcal{L}_r$ to be the level set $\left\{\pi \in \Pi : R(\pi) = r\right\}$, then the goal is to estimate a



Decision list π:

If $x_1 > \tau_1$ **then** a_1;
else if $x_2 > \tau_2$ **then** a_2;
else a_0.

Decision list π':

If $x_1 \leq \tau_1$ **and** $x_2 > \tau_2$ **then** a_2;
else if $x_1 > \tau_1$ **then** a_1;
else a_0.

Figure 1.2 Left: diagram of a decision list dictated by regions $\mathcal{R}_1 = \{x \in \mathbb{R}^2 : x_1 > \tau_1\}$, $\mathcal{R}_2 = \{x \in \mathbb{R}^2 : x_1 \leq \tau_1, x_2 > \tau_2\}$, and $\mathcal{R}_0 = \{x \in \mathbb{R}^2 : x_1 \leq \tau_1, x_2 \leq \tau_2\}$, and treatment recommendations a_1, a_2, and a_0. Middle: representation of the decision list that requires only x_1 in the first clause. Right: alternative representation of the same decision list that requires both x_1 and x_2 in the first clause.

regime in the set $\arg\min_{\pi \in \mathcal{L}\{R(\pi^{\text{opt}})\}} N(\pi)$. Define $\widehat{\mathcal{L}}(r) = \{\pi \in \Pi : \widehat{R}(\pi) = r\}$. Let $\widetilde{\pi}$ be an estimator of an element in the set $\arg\max_{\pi \in \Pi} \widehat{R}(\pi)$. In the following we provide an algorithm that ensures our estimator, $\widehat{\pi}$, belongs to the set $\arg\min_{\pi \in \widehat{\mathcal{L}}\{\widehat{R}(\widetilde{\pi})\}} \widehat{N}(\pi)$, where $\widehat{N}(\pi)$ is defined by replacing the probabilities in $N(\pi)$ with sample proportions.

1.2.4 Computation

Estimation proceeds in two steps: (i) approximate an element $\widetilde{\pi} \in \arg\max_{\pi \in \Pi} \widehat{R}(\pi)$, where $\widehat{R}(\pi)$ is constructed using (1.2); and (ii) find an element $\widehat{\pi} \in \arg\min_{\pi \in \widehat{\mathcal{L}}\{\widehat{R}(\widetilde{\pi})\}} \widehat{N}(\pi)$.

Approximation of $\arg\max_{\pi \in \Pi} \widehat{R}(\pi)$

Maximizing $\widehat{R}(\pi)$ over $\pi \in \Pi$ is computationally burdensome in problems with more than a handful of covariates because of the indicator functions in (1.2) and the discreteness of the decision list. However, the tree structure of decision lists suggests a greedy algorithm in the spirit of classification and regression trees (CART, Breiman et al., 1984). Assume that for the jth covariate, there is a candidate set of finitely many possible cutoff val-

ues $\mathcal{X}_j$. These cutoffs might be dictated by clinical guidelines, e.g., if the covariate is a comorbid condition then the thresholds might reflect low, moderate, and high levels of impairment; alternatively, these cutoffs could be chosen to equal empirical or theoretical percentiles of that covariate. There is no restriction imposed on these cutoffs. Let $\mathcal{C}$ denote the set of all conditions that induce regions of the form in (1.4) with the cutoffs $\tau_{j_k} \in \mathcal{X}_{j_k}$ for $k = 1, 2$, $j_k \in \{1, \ldots, p\}$.

Before giving the details of the algorithm, we provide a conceptual overview. The algorithm first uses exhaustive search to find a decision list with exactly one clause, of the form $\pi = \{(c_1, a_1), a_1'\}$, which maximizes $\widehat{R}(\pi)$. Let $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}_1'\}$ denote this decision list. The algorithm then uses exhaustive search to find the decision list that maximizes $\widehat{R}(\pi)$ over decision lists with exactly two clauses, the first of which must be either $(\widetilde{c}_1, \widetilde{a}_1)$ or $(\widetilde{c}_1', \widetilde{a}_1')$, where $\widetilde{c}_1'$ is the negation of $\widetilde{c}_1$ such that $\mathcal{T}(\widetilde{c}_1') = \mathcal{T}(\widetilde{c}_1)^c$; e.g., if $\widetilde{c}_1$ has the form $x_{j_1} \leq \tau_1$ and $x_{j_2} \leq \tau_2$, then $\widetilde{c}_1'$ would be $x_{j_1} > \tau_1$ or $x_{j_2} > \tau_2$. Although the decision lists $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}_1'\}$ and $\{(\widetilde{c}_1', \widetilde{a}_1'), \widetilde{a}_1\}$ yield identical treatment recommendations and have the same value, their first clauses are distinct, and may lead to substantially different final decision lists. Hence it is necessary to consider both possibilities for the first clause. The algorithm proceeds recursively by adding one clause at a time until some stopping criterion (described below) is met.

Hereafter, for a decision list π described by $\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}$ for some $L \geq 0$, write

$\widehat{R}[\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}]$ to denote $\widehat{R}(\pi)$; e.g., for $L = 0$, $\widehat{R}[\{a_0\}]$ is the estimated value of the regime that assigns treatment a_0 to all patients. For any decision list with a vacuous condition, e.g., $\{\cap_{\ell<j}\mathcal{T}(c_\ell)^c\} \bigcap \mathcal{T}(c_j) = \emptyset$ for some j, define $\widehat{R}[\{(c_1, a_1), \ldots, (c_L, a_L), a_0\}] =$

$-\infty$. Let z_ρ be the 100ρ percentile of the standard normal distribution. Let Π_{temp} denote the set of regimes to which additional clauses can be added, and let Π_{final} denote the set of regimes that have met one of the stopping criteria. The algorithm is as follows, and an illustrative example with a step-by-step run of the algorithm is given in the Appendix.

Step 1. Choose a maximum list length $L_{\max}$ and a critical level $\alpha \in (0,1)$. Compute $\widetilde{a}_0 = \arg\max_{a_0 \in \mathcal{A}} \widehat{R}\left[\{a_0\}\right]$. Set $\Pi_{\text{temp}} = \emptyset$ and $\Pi_{\text{final}} = \emptyset$.

Step 2. Compute $(\widetilde{c}_1, \widetilde{a}_1, \widetilde{a}'_1) = \arg\max_{(c_1,a_1,a'_1) \in \mathcal{C}\times\mathcal{A}\times\mathcal{A}} \widehat{R}\left[\{(c_1,a_1),a'_1\}\right]$ and $\widehat{\Delta}_1 = \widehat{R}\left[\{(\widetilde{c}_1,\widetilde{a}_1),\widetilde{a}'_1\}\right] - \widehat{R}\left[\{\widetilde{a}_0\}\right]$. If $\widehat{\Delta}_1 < z_{1-\alpha}\big\{\widehat{\text{Var}}\big(\widehat{\Delta}_1\big)\big\}^{1/2}$ then let $\pi = \{\widetilde{a}_0\}$, set $\Pi_{\text{final}} = \{\pi\}$, and go to Step 5; otherwise let $\pi = \{(\widetilde{c}_1,\widetilde{a}_1),\widetilde{a}'_1\}$, $\pi' = \{(\widetilde{c}'_1,\widetilde{a}'_1),\widetilde{a}_1\}$, set $\Pi_{\text{temp}} = \{\pi,\pi'\}$, and proceed to Step 3, where $\widetilde{c}'_1$ is the negation of $\widetilde{c}_1$.

Step 3. Pick an element $\overline{\pi} \in \Pi_{\text{temp}}$, say $\overline{\pi} = \big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),\overline{a}'_{j-1}\big\}$, where $j-1$ is the length of $\overline{\pi}$. Remove $\overline{\pi}$ from Π_{temp}. With the clauses $(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1})$ held fixed, compute $(\widetilde{c}_j,\widetilde{a}_j,\widetilde{a}'_j) = \arg\max_{(c_j,a_j,a'_j)\in\mathcal{C}\times\mathcal{A}\times\mathcal{A}} \widehat{R}\left[\big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),(c_j,a_j),a'_j\big\}\right]$ and $\widehat{\Delta}_j = \widehat{R}\left[\big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),(\widetilde{c}_j,\widetilde{a}_j),\widetilde{a}'_j\big\}\right] - \widehat{R}\left[\big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),\overline{a}'_{j-1}\big\}\right]$. If $\widehat{\Delta}_j < z_{1-\alpha}\big\{\widehat{\text{Var}}\big(\widehat{\Delta}_j\big)\big\}^{1/2}$, then let $\pi = \big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),\overline{a}'_{j-1}\big\}$, and set $\Pi_{\text{final}} = \Pi_{\text{final}} \bigcup \{\pi\}$; otherwise if $j = L_{\max}$, let $\pi = \big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),(\widetilde{c}_j,\widetilde{a}_j),\widetilde{a}'_j\big\}$, and set $\Pi_{\text{final}} = \Pi_{\text{final}} \bigcup \{\pi\}$; otherwise set $\Pi_{\text{temp}} = \Pi_{\text{temp}} \bigcup \{\pi,\pi'\}$, where $\widetilde{c}'_j$ is the negation of $\widetilde{c}_j$, $\pi = \big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),(\widetilde{c}_j,\widetilde{a}_j),\widetilde{a}'_j\big\}$, and $\pi' = \big\{(\overline{c}_1,\overline{a}_1),\ldots,(\overline{c}_{j-1},\overline{a}_{j-1}),(\widetilde{c}'_j,\widetilde{a}'_j),\widetilde{a}_j\big\}$.

Step 4. Repeat Step 3 until Π_{temp} becomes empty.

Step 5. Compute $\widetilde{\pi} = \arg\max_{\pi\in\Pi_{\text{final}}} \widehat{R}(\pi)$. Then $\widetilde{\pi}$ is the estimated optimal decision list.

The above description is simplified to illustrate the main ideas. The actual implementation of this algorithm avoids exhaustive searches by pruning the search space $\mathcal{C} \times \mathcal{A} \times \mathcal{A}$. It also avoids explicit construction of Π_{temp} and Π_{final}. Complete implementation details are provided in the Appendix. In the algorithm, the decision list stops growing if either the estimated increment in the value, $\widehat{\Delta}_j$, is not sufficiently large compared to an estimate of its variation, $\left\{\widehat{\text{Var}}\left(\widehat{\Delta}_j\right)\right\}^{1/2}$, or if it reaches the pre-specified maximal length $L_{\max}$. We estimate $\text{Var}(\widehat{\Delta}_j)$ using large sample theory; the expression is given in the Appendix. This variance estimator is a crude approximation, as it ignores uncertainty introduced by the estimation of the decision lists; however, it can be computed quickly, and in simulated experiments it appears sufficient for use in a stopping criterion. The significance level α is a user-chosen tuning parameter. In our simulation experiments, we chose $\alpha = 0.05$; results were not sensitive to this choice (see Appendix). To avoid lengthy lists, we set $L_{\max} = 10$. Nevertheless, in our simulations and applications the estimated lists never reach this limit. Finally, it may be desirable in practice to restrict the set of candidate clauses so that, for each j, the number of subjects in $\widehat{\mathcal{R}}_j = \{\cap_{\ell<j}\mathcal{T}(\widehat{c}_\ell)^c\} \bigcap \mathcal{T}(\widehat{c}_j)$ exceeds some minimal threshold. This can be readily incorporated into the above algorithm by simply discarding candidate clauses that induce partitions that contain an insufficient number of observations.

The time complexity of the proposed algorithm is $O\left[2^{L_{\max}} m p^2 \left\{n + (\max_j \#\mathcal{X}_j)^2\right\}\right]$ (see Appendix), where $\#\mathcal{X}_j$ is the number of cutoff values in $\mathcal{X}_j$. Because $2^{L_{\max}}$ and m are constants that are typically small relative to $p^2\left\{n + (\max_j \#\mathcal{X}_j)^2\right\}$, the time complexity is essentially $O(np^2)$ provided that $\max_j \#\mathcal{X}_j$ is either fixed or diverges more slowly than $n^{1/2}$. Hence, the time complexity is the same as a single least squares fit, indicating that

the proposed algorithm runs very fast and scales well in both dimension p and sample size n.

Finding an Element of $\arg\min_{\pi \in \widehat{\mathcal{L}}\left\{\widehat{R}(\widetilde{\pi})\right\}} \widehat{N}(\pi)$

To find an element within the set $\arg\min_{\pi \in \widehat{\mathcal{L}}\left\{\widehat{R}(\widetilde{\pi})\right\}} \widehat{N}(\pi)$, we enumerate all regimes in $\mathcal{L}\left\{\widehat{R}(\widetilde{\pi})\right\}$ with length no larger than $L_{\max}$ and select among them the list with the minimal cost. The enumeration algorithm is recursive and requires a substantial amount of bookkeeping; therefore, we describe the basic idea here and defer implementation details to the Appendix. Suppose $\widetilde{\pi}$ is described by $\{(\widetilde{c}_1, \widetilde{a}_1), \ldots, (\widetilde{c}_L, \widetilde{a}_L), \widetilde{a}_0\}$. Call a condition of the form $x_j \leq \tau_j$ an atom. There exist $K \leq 2L$ atoms, say $d_1, \ldots, d_K$, such that each clause $\widetilde{c}_\ell$, $\ell = 1, \ldots, L$, can be expressed using the union, intersection, and/or negation of at most two of these atoms. The algorithm proceeds by generating all lists with clauses representable using the foregoing combinations of at most two atoms. To reduce computation time, we use a branch-and-bound scheme (Brusco and Stahl, 2006) that avoids constructing lists with vacuous conditions or those that are provably worse than an upper bound on $\min_{\pi \in \widehat{\mathcal{L}}\left\{\widehat{R}(\widetilde{\pi})\right\}} N(\pi)$. In the simulation experiments in the next section, the average runtime for the enumeration algorithm was less than one second running on a single core of a 2.3GHz AMD Opteron™ processor and 1GB of DDR3 RAM.

1.3 Simulation Experiments

We use a series of simulated experiments to examine the finite sample performance of the proposed method. The average value $\mathbb{E}\{R(\widehat{\pi})\}$ and the average cost $\mathbb{E}\{N(\widehat{\pi})\}$ are the primary performance measures. We consider generative models with (i) binary

and continuous outcomes; (ii) binary and trinary treatments; (iii) correctly and incorrectly specified models; and (iv) low- and high-dimensional covariates. The class of data-generating models that we consider is as follows. Covariates are drawn from a p-dimensional Gaussian distribution with mean zero and autoregressive covariance matrix such that $\text{cov}(X_k, X_\ell) = 4(1/5)^{|k-\ell|}$, and the treatments are sampled uniformly so that $P(A = a|X = x) = 1/m$ for all $x \in \mathbb{R}^p$ and $a \in \mathcal{A}$. Let $\phi(x, a)$ be a real-valued function of x and a; given $X = x$ and $A = a$, continuous outcomes are normally distributed with mean $2 + x_1 + x_3 + x_5 + x_7 + \phi(x, a)$ and variance 1, whereas binary outcomes follows a Bernoulli distribution with success probability $\text{expit}\{2 + x_1 + x_3 + x_5 + x_7 + \phi(x, a)\}$, where $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$. Table 1.1 lists the expressions of ϕ used in our generative models and the number of treatments, m, in $\mathcal{A}$. Under these outcome models, the optimal regime is $\pi^{\text{opt}}(x) = \arg\max_a \phi(x, a)$.

For comparison, we estimate π^{opt} by parametric Q-learning, nonparametric Q-learning, outcome weighted learning (OWL, Zhao et al., 2012) and modified covariate approach (MCA, Tian et al., 2014). For parametric Q-learning, we use linear regression when Y is continuous and logistic regression when Y is binary. The linear component in the regression model has the form $\sum_{a=1}^{m} I(A = a)(1, X^{\text{T}})\beta_a$, where $\beta_1, \ldots, \beta_m$ are unknown coefficient vectors. A LASSO penalty (Tibshirani, 1996) is used to reduce overfitting; the amount of penalization is chosen by minimizing 10-fold cross-validated prediction error. For nonparametric Q-learning, we use support vector regression when Y is continuous and support vector machines when Y is binary (Zhao et al., 2011), both are implemented using a Gaussian kernel. Tuning parameters for non-parametric Q-learning are selected by minimizing 10-fold cross-validated prediction error. For OWL, both linear and Gaus-

sian kernels are used and we follow the same tuning strategy as in Zhao et al. (2012). For MCA, we incorporate the efficiency augmentation term described in Tian et al. (2014). Both OWL and MCA are limited to two treatment options.

To implement our method, the mean model, $\mu(x, a)$, in (1.1), is estimated as in parametric Q-learning. The propensity score $\omega(x, a)$ is estimated by $n^{-1}\sum_{i=1}^{n} I(A_i = a)$. All the comparison methods result in treatment regimes that are more difficult to interpret than a decision list; thus, our intent is to show that decision lists are competitive in terms of the achieved value of the estimated regime, $\mathbb{E}\{R(\widehat{\pi})\}$, while being significantly more interpretable and less costly.

Results in Table 1.2 are based on the average over 1000 Monte Carlo replications with data sets of size $n = 500$ if $m = 2$ and Y is continuous; $n = 750$ if $m = 3$ and Y is continuous; $n = 1000$ if $m = 2$ and Y is binary; and $n = 1500$ if $m = 3$ and Y is binary. The value $R(\widehat{\pi})$ and cost $N(\widehat{\pi})$ were computed using an independent test set of size 10^6.

Table 1.2 shows that the decision list is competitive in terms of the value obtained across the entire suite of simulation experiments. If π^{opt} can be represented as a decision list, the proposed method produces the best value. However, even in settings in which the optimal regime is not a decision list, the estimated decision list appears to perform well. Recall that the proposed algorithm attempts to find the best approximation of the optimal regime within the class of regimes that are representable as a decision list. Figure 1.3 shows the average estimated decision list in misspecified settings II and III with continuous outcome and $p = 10$. In these settings, the estimated decision list provides a reasonable approximation of the true optimal regime. In addition, the cost of the decision list is notably smaller than the cost of the parametric Q-learning estimator or the MCA

Table 1.1 The second column gives the number of treatment options m. The third column gives the set of ϕ functions used in the outcome models. The fourth column specifies the form of the optimal regime $\pi^{\text{opt}}(x) = \arg\max_a \phi(x,a)$ where: "linear" indicates that $\pi^{\text{opt}}(x) = \arg\max_a\{(1,x^{\text{T}})\beta_a\}$ for some coefficient vectors $\beta_a \in \mathbb{R}^{p+1}$, $a \in \mathcal{A}$; "decision list" indicates that π^{opt} is representable as a decision list; and "nonlinear" indicates that $\pi^{\text{opt}}(x)$ is neither linear nor representable as a decision list.

Setting	m	Expression of ϕ	Form of π^{opt}
I	2	$\phi_1(x,a) = I(a=2)\{3I(x_1 \le 1, x_2 > -0.6) - 1\}$	decision list
II	2	$\phi_2(x,a) = I(a=2)(x_1 + x_2 - 1)$	linear
III	2	$\phi_3(x,a) = I(a=2)\arctan(\exp(1+x_1) - 3x_2 - 5)$	nonlinear
IV	2	$\phi_4(x,a) = I(a=2)(x_1 - x_2 + x_3 - x_4)$	linear
V	3	$\phi_5(x,a) = I(a=2)\{4I(x_1 > 1) - 2\}$ $+ I(a=3)I(x_1 \le 1)\{2I(x_2 \le -0.3) - 1\}$	decision list
VI	3	$\phi_6(x,a) = I(a=2)(2x_1) + I(a=3)(-x_1x_2)$	nonlinear
VII	3	$\phi_7(x,a) = I(a=2)(x_1 - x_2) + I(a=3)(x_3 - x_4)$	linear

estimator. Nonparametric Q-learning OWL always use all covariates, so their costs are always equal to p.

In the Appendix, we derive point estimates and prediction intervals for $R(\widehat{\pi})$. We also present simulation results to illustrate the accuracy of variable selection for the decision list.

1.4 Applications

1.4.1 Breast Cancer Data

Gail and Simon (1985) compared the treatment effects of chemotherapy alone and chemotherapy with tamoxifen using data collected from the NSABP trial. Their regime recommended chemotherapy alone to patients with age ≤ 50 and PR ≤ 10 and chemotherapy

Table 1.2 The average value and the average cost of estimated regimes in simulated experiments. In the header, p is the dimension of patient covariates; DL refers to the proposed method using decision list; Q_1 refers to parametric Q-learning; Q_2 refers to nonparametric Q-learning; OWL_1 and OWL_2 refer to outcome weighted learning with linear kernel and Gaussian kernel, respectively; MCA refers to modified covariate approach with efficiency augmentation. OWL and MCA are not applicable under Setting V, VI and VII.

p	Setting	Value						Cost		
		DL	Q_1	Q_2	OWL_1	OWL_2	MCA	DL	Q_1	MCA
Continuous response										
10	I	2.78	2.53	2.53	2.33	2.29	2.54	1.64	9.0	5.1
	II	2.70	2.80	2.79	2.61	2.54	2.80	1.64	9.0	5.1
	III	2.59	2.54	2.53	2.29	2.24	2.55	1.68	9.1	4.9
	IV	2.89	3.37	3.35	3.16	3.09	3.37	2.50	9.5	7.4
	V	2.90	2.67	2.59	–	–	–	1.90	9.5	–
	VI	3.98	3.46	3.95	–	–	–	1.61	9.2	–
	VII	3.22	3.75	3.73	–	–	–	2.56	9.7	–
50	I	2.76	2.51	2.36	2.21	2.19	2.53	1.80	21.3	9.2
	II	2.70	2.79	2.73	2.26	2.27	2.79	1.64	21.4	9.3
	III	2.59	2.52	2.35	2.16	2.12	2.54	1.71	23.1	9.0
	IV	2.89	3.36	3.27	2.76	2.70	3.36	2.53	25.4	14.9
	V	2.87	2.63	2.33	–	–	–	2.14	28.5	–
	VI	3.95	3.43	3.47	–	–	–	1.69	26.6	–
	VII	3.21	3.74	3.61	–	–	–	2.55	30.8	–
Binary response										
10	I	0.77	0.74	0.69	0.73	0.73	0.74	1.94	8.9	4.1
	II	0.71	0.72	0.60	0.71	0.71	0.72	1.69	9.2	5.3
	III	0.73	0.73	0.68	0.72	0.72	0.73	2.10	9.2	4.7
	IV	0.71	0.76	0.66	0.75	0.74	0.75	2.40	9.6	8.4
	V	0.75	0.73	0.62	–	–	–	2.52	9.6	–
	VI	0.79	0.75	0.64	–	–	–	2.09	9.5	–
	VII	0.77	0.81	0.69	–	–	–	2.83	9.9	–
50	I	0.76	0.73	0.69	0.71	0.70	0.73	2.64	21.9	8.3
	II	0.71	0.72	0.60	0.70	0.69	0.71	1.87	26.2	6.4
	III	0.73	0.72	0.67	0.70	0.69	0.72	2.53	25.0	7.3
	IV	0.71	0.76	0.66	0.73	0.72	0.74	2.55	31.0	13.8
	V	0.74	0.72	0.61	–	–	–	3.15	30.4	–
	VI	0.78	0.75	0.63	–	–	–	2.41	29.6	–
	VII	0.76	0.81	0.68	–	–	–	2.97	35.7	–

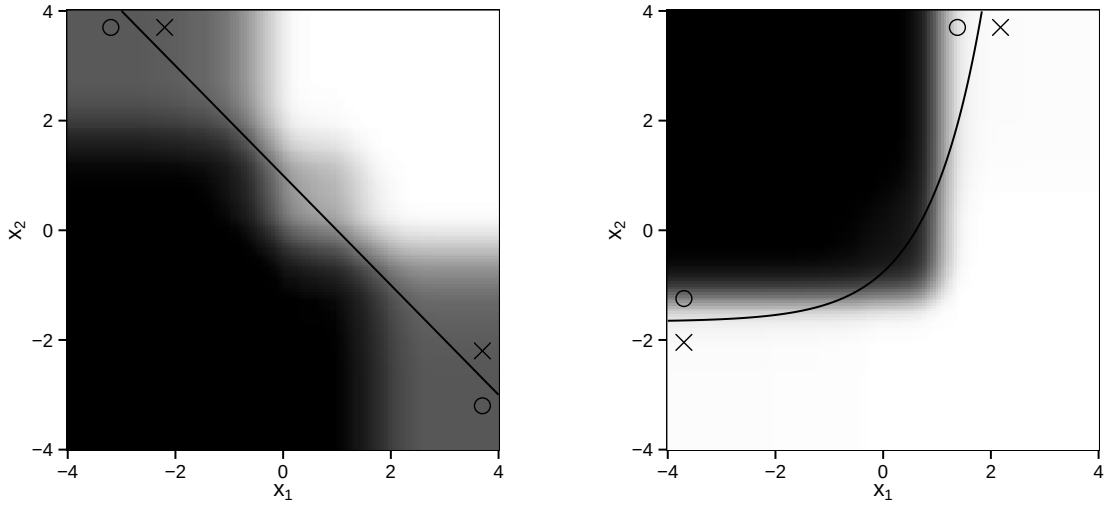


Figure 1.3 Left: average estimated regimes under setting II. Right: average estimated regimes under setting III. In both settings π^{opt} cannot be represented as decision list. The solid line is the treatment decision boundary under π^{opt}. The region where treatment 1 is better than treatment 2 is marked by circles, while the region where treatment 2 is better than treatment 1 is marked by crosses. For every point $(x_1, x_2)^{\text{T}}$, we compute the proportion of 1000 replications that the estimated regime recommends treatment 1 to a patient with covariate $(x_1, x_2, 0, \ldots, 0) \in \mathbb{R}^{10}$. The larger the proportion, the darker the shade.

plus tamoxifen to all others. Because the variables involved in the treatment regime constructed by Gail and Simon were chosen using clinical judgment, it is of interest to see what regime emerges from a more data-driven procedure. Thus, we use the proposed method to estimate an optimal treatment regime in the form of a decision list using data from the NSABP trial.

As in Gail and Simon (1985), we take three-year disease-free survival as the outcome, so that $Y = 1$ if the subject survived disease-free for three years after treatment, and $Y = 0$ otherwise. Patient covariates are age (years), PR (fmol), estrogen receptor level (ER, fmol), tumor size (centimeters), and number of histologically positive nodes (number of nodes, integer). We estimated the optimal treatment regime representable as a decision list using data from the 1164 subjects with complete observations on these variables. Because treatment assignment was randomized in NSABP, we estimated $\omega(x, a)$ by the sample proportion of subjects receiving treatment a. Based on exploratory analyses, we estimated $\mu(x, a)$ using a logistic regression model with transformed predictors $z = z(x) = \{\text{age}, \log(1 + \text{PR}), \log(1 + \text{ER}), \text{tumor-size}, \log(1 + \text{number-of-nodes})\}^{\text{T}}$.

The estimated optimal treatment regime representable as a decision list is given in the top panel of Figure 1.4; the regime estimated by Gail and Simon is given in the bottom panel of this figure. The structure of the two treatment regimes is markedly similar. The treatment recommendations from the two regimes agree for 92% of the patients in the NSABP data. In this data set, 33% of the patients have a PR value less than 3; 13% of the patients have a PR values between 3 and 10; and 54% of the patients have a PR value greater than 10.

In a previous analysis of the NSABP data, Zhang et al. (2012c) recommended that

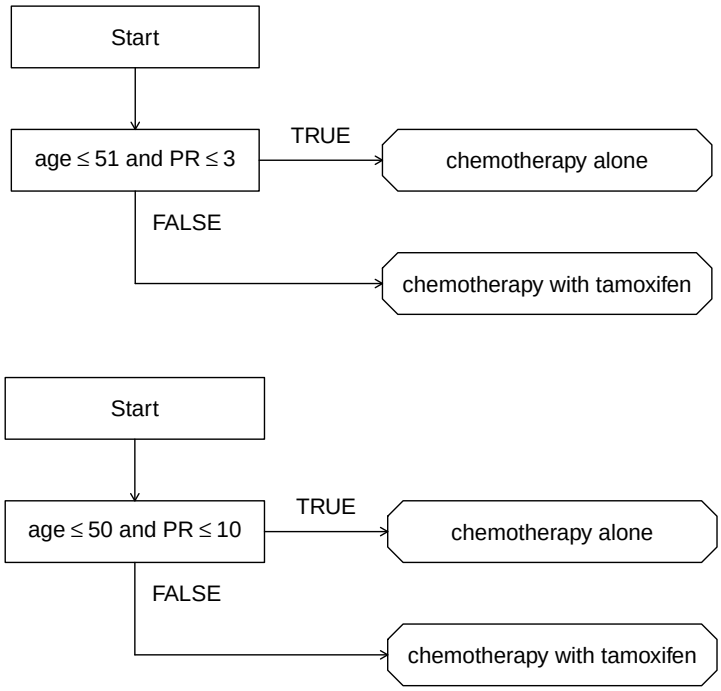


Figure 1.4 Top: estimated optimal treatment regime representable as a decision list. Bottom: treatment regime proposed by Gail and Simon (1985).

patients with age+7.98 log(1+PR) $\leq$ 60 receive chemotherapy alone and all others receive chemotherapy plus tamoxifen. However, this regime was built using only age and PR as potential predictors with no data-driven variable selection. In contrast, the proposed method selects age and PR from the list of potential predictors. For completeness, we also implemented parametric Q-learning using a logistic regression model with covariate vector z. The estimated regime recommends chemotherapy alone if $1.674 - 0.021$ age $-$ $0.076\log(1+\text{PR}) - 0.116\log(1+\text{ER}) - 0.024$ tumor-size $- 0.274\log(1+\text{number-of-nodes}) \geq$ 0 and chemotherapy with tamoxifen otherwise. The treatment recommendation dictated by parametric Q-learning agrees with that dictated by decision list for 86% of the subjects in the data set.

To estimate the survival probability under each estimated regime, we use cross-validation. The data set was randomly divided into a training set containing 80% of the subjects and a test set containing 20% of the subjects. The optimal regime was estimated using both approaches on the training set, and its value was computed using (1.2) (with $\widehat{\mu} \equiv 0$) on the test set. To reduce variability, this process was repeated 100 times. The estimated survival probability is 0.65 for the regime representable as decision list and 0.66 for the regime obtained from parametric Q-learning. Thus, the proposed method greatly improves interpretability while preserving quality.

1.4.2 Chronic Depression Data

Keller et al. (2000) compared nefazodone, psychotherapy, and combination of nefazodone and psychotherapy for treating patients with chronic depression in a three-arm randomized clinical trial. Among the three treatments considered, combination therapy was

shown to be the most beneficial in terms of efficacy as measured by the Hamilton Rating Scale for Depression score (HRSD). However, the combination treatment is significantly more expensive and burdensome than monotherapy. Therefore, it is of interest to construct a treatment regime that recommends combination therapy only to subjects for whom there is a significant benefit over monotherapy.

Because lower HRSD indicates less severe symptoms, we define outcome $Y = -\text{HRSD}$ to be consistent with our paradigm of maximizing the mean outcome. Patient covariates comprise 50 pretreatment variables, including personal habits and difficulties, medication history and various scores from several psychological questionnaires; a list of these variables is given in the Appendix. We estimate an optimal regime using data from the $n = 647$ (of 680 enrolled) subjects in the clinical trial with complete data. Because treatments were randomly assigned, we estimated $\omega(x, a)$ by the sample proportion of subjects receiving treatment a. We estimated $\mu(x, a)$ using a penalized linear regression model with all patient covariates and treatment by covariate interactions. Penalization was implemented with a LASSO penalty tuned using 10-fold cross-validated prediction error.

The estimated optimal treatment regime representable as a decision list is displayed in Figure 1.1. One explanation for this rule is as follows. Those with strong physical anxiety symptoms (somatic) and significant cognitive impairment (retardation) may be unlikely to benefit from psychotherapy alone or in combination with nefazodone and are therefore recommended to nefazodone alone. Otherwise, because psychotherapy is a primary tool for treating anxiety (HAM-A) and nefazodone is associated with sleep disturbance (sleep), it may be best to assign subjects with moderate to severe anxiety and severe sleep

disturbance to psychotherapy alone. All others are assigned to the combination therapy.

The estimated regime contains only four covariates. In contrast, the regime estimated by parametric Q-learning using linear regression and LASSO penalty involves a linear combination of twenty-four covariates, making it difficult to explain and expensive to implement. To compare the quality of these two regimes, we use random-split cross-validation as in Section 4.1. The estimated HRSD score under the regime representable as decision list is 12.9, while that under the regime estimated by parametric Q-learning is 11.8. Therefore, by using decision lists we are able to obtain a remarkably more parsimonious regime with high quality, which facilitates easier interpretation.

1.5 Discussion

Data-driven treatment regimes have the potential to improve patient outcomes and generate new clinical hypotheses. Estimation of an optimal treatment regime is typically conducted as a secondary, exploratory analysis aimed at building knowledge and informing future clinical research. Thus, it is important that methodological developments are designed to fit this exploratory role. Decision lists are a simple yet powerful tool for estimation of interpretable treatment regimes from observational or experimental data. Because decision lists can be immediately interpreted, clinical scientists can focus on the scientific validity of the estimated treatment regime. This allows the communications between the statistician and clinical collaborators to focus on the science rather than the technical details of a statistical model.

Due to the "if-then" format and the conditions given in (1.4), the estimated regime, as a function of the data, is discrete. Thus, a theoretical proof of the consistency of the

treatment recommendations using decision lists is heavily technical and will be presented elsewhere. We provide some empirical evidence in the Appendix that the estimated regime gives consistent treatment decisions.

Chapter 2

Interpretable Dynamic Treatment Regimes

2.1 Introduction

Precision medicine is now almost universally recognized as a path to delivering the best possible healthcare (Collins and Varmus, 2015; Ashley, 2015; Jameson and Longo, 2015). Furthermore, technological advancements and investment in big-data infrastructure have made it possible to collect, store, and curate large amounts of patient-level data to inform the practice of precision medicine (Krumholz, 2014). Quantitative researchers have responded with a surge of methodological developments aimed at 'mathematizing' precision medicine in the form of treatment regimes, a sequence of decision rules, one per stage of clinical intervention, that map up-to-date patient information to a treatment recommendation; an optimal treatment regime is defined as maximizing the mean of some desirable clinical outcome if applied to a population of interest. It can be shown that

even under the simplest generative models the optimal regime is a nonlinear function of patient information (Robins, 2004; Schulte et al., 2014; Laber et al., 2014); consequently, to avoid model misspecification, a recent trend is to apply flexible supervised learning methods to estimate optimal treatment regimes. These flexible methods include direct-search using large-margin classifiers (Zhao et al., 2012, 2015b; Kang et al., 2014; Zhao et al., 2015a; Xu et al., 2015b); Q-learning with non-parametric regression models (Qian and Murphy, 2011; Zhao et al., 2011; Moodie et al., 2013; Zhou and Kosorok, 2016); and tree-based methods (Zhang et al., 2012b; Laber and Zhao, 2015; Zhang et al., 2015; Doove et al., 2015). Further testament to the popularity of these methods is that the Journal of the American Statistical Association's Theory and Methods Invited Paper and the Case Studies and Applications Invited Paper at the 2016 Joint Statistical Meetings will feature non-parametric methods for estimating treatment regimes (Zhou et al., 2015; Xu et al., 2015a).

Flexible estimation methods mitagate the risk of model misspecification but potentially at the price of rendering the estimated regime unintelligible. This price is may be too high in settings where the primary role of an estimated optimal regime is to generate new scientific hypotheses or inform future research. For example, in the context of sequential multiple assignment clinical trials (SMARTs, Murphy, 2005; Lei et al., 2012) estimation of an optimal treatment regime is typically included as a secondary, exploratory analysis as sizing the trial to ensure high-quality estimation of an optimal regime is complex (Laber et al., 2016). Tree-based regimes, like regression or classification trees, offer flexibility while retaining interpretability. Here, we propose a method for estimation of an optimal treatment regime that comprises a sequence of decision rules

each of which is represented as a sequence if-then statements mapping logical clauses to treatment recommendations. Decision rules of this form are a special case of tree-based rules, known as decision lists (Rivest, 1987; Marchand and Sokolova, 2005; Letham et al., 2012; Wang and Rudin, 2015; Zhang et al., 2015), that are immediately interpretable in a domain context as they can be expressed in either flow-chart or paragraph form. Thus, regimes of this form are amenable to critique and examination by clinicians and facilitate collaborative, iterative development of data-driven precision medicine. Furthermore, we shall show that despite the structure imposed by the decision lists, they are sufficiently expressive so as to provide high-quality regimes even under non-linear generative models previously used in the literature to illustrate the value of non-parametric estimation methods.

In addition to the clinical and scientific value of interpretable, list-based regimes, the proposed work provides a number of important methodological contributions. Unlike existing tree-based methods for estimating optimal treatment regimes, the proposed methodology applies to problems with an arbitrary number of treatment stages and treatments per stage. In principle, robust policy-search (Zhang et al., 2013) could be used with CART (Breiman et al., 1984) to estimate a multi-stage, tree-based treatment regime; however, this method relies on inverse probability of treatment weighting which rapidly becomes unstable as the number of treatment stages increases. A second contribution is that we prove that the proposed estimator is consistent for the optimal regime within the class of list-based regimes and derive rates of convergence for the proposed estimator. These theoretical results are non-trivial because the discreteness of the list precludes the use of standard asymptotic approaches; to our knowledge these are first

results on convergence rates for decision lists and are therefore of independent interest. A third contribution is the proposed estimation algorithm used to construct the decision lists at each stage. This algorithm reduces computation time of naive recursive-splitting algorithm from $O(n^3)$ to $O(n \log n)$ where n is the number of subjects in the sample, furthermore we modify the splitting criteria proposed by Zhang et al. (2015) to avoid (asymptotically) becoming stuck in a local mode.

In Section 2.2, we describe list-based treatment regimes and describe our estimation algorithm. In Section 2.3, we prove consistency of the proposed estimator and derive rates of convergence. In Section 2.4, we demonstrate the finite sample performance of the proposed method using simulation experiments. We illustrate the proposed method using data from a clinical trail in Section 2.5 and make concluding remarks in Section 2.6.

2.2 Methodology

2.2.1 Framework

Consider n *i.i.d.* observations collected from a sequential clinical trial with T stages; the proposed methodology also applies to observational data provided that standard causal assumptions required for Q-learning are satisfied (see Schulte et al., 2014, for a statement of these assumptions). In the assumed setup the observed data are $\{(\boldsymbol{S}_{it}, A_{it}, Y_{it}) : t = 1, \ldots, T\}_{i=1}^{n}$, which comprise *i.i.d.* trajectories of the form $\{(\boldsymbol{S}_t, A_t, Y_t) : t = 1, \ldots, T\}$ where: $\boldsymbol{S}_t \in \mathbb{R}^{p_t}$ is a vector of covariates measured at the beginning of the t-th stage; $A_t \in \mathcal{A}_t$ is the treatment actually received during the t-th stage; and $Y_t \in \mathbb{R}$ is a scalar outcome measured at the end of the t-th stage. Let $m_t = |\mathcal{A}_t|$ denote the number of

available treatment options at the t-th stage. The final outcome of interest is the sum of immediate outcomes, $Y = \sum_{t=1}^{T} Y_t$. We assume that larger values of Y are better. Let $\boldsymbol{X}_t$ denote the information available to the decision maker at stage t so that $\boldsymbol{X}_1 = \boldsymbol{S}_1$ and $\boldsymbol{X}_t = (\boldsymbol{X}_{t-1}^{\mathrm{T}}, A_{t-1}, Y_{t-1}, \boldsymbol{S}_t^{\mathrm{T}})^{\mathrm{T}}$ for $t > 1$. Let $\mathcal{X}_t \subset \mathbb{R}^{d_t}$ be the support of $\boldsymbol{X}_t$, where $d_t = \sum_{s=1}^{t} p_s + 2(t-1)$ is the dimension of $\boldsymbol{X}_t$.

A treatment regime $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ is a sequence of functions $\pi_t : \mathcal{X}_t \to \mathcal{A}_t$ so that under $\boldsymbol{\pi}$ a patient presenting with $\boldsymbol{X}_t = \boldsymbol{x}_t$ at stage t is recommended treatment $\pi_t(\boldsymbol{x}_t)$. For any regime $\boldsymbol{\pi}$, let $\mathbb{E}^{\boldsymbol{\pi}}$ denote expectation with respect to distribution induced by assigning treatments according to $\boldsymbol{\pi}$. Given a class of regimes Π, an optimal regime satisfies, $\boldsymbol{\pi}^{\mathrm{opt}} \in \Pi$ and $\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{opt}}} Y \geq \mathbb{E}^{\boldsymbol{\pi}} Y$ for all $\boldsymbol{\pi} \in \Pi$. Our goal is to construct an estimator of $\boldsymbol{\pi}^{\mathrm{opt}}$ when Π is the class of list-based regimes. Each decision rule π_t in a list-based regime has the form:

$$
\begin{aligned}
&\texttt{If } \boldsymbol{x}_t \in R_{t1} \texttt{ then } a_{t1};\\
&\texttt{else if } \boldsymbol{x}_t \in R_{t2} \texttt{ then } a_{t2};\\
&\ldots\\
&\texttt{else if } \boldsymbol{x}_t \in R_{tL_t} \texttt{ then } a_{tL_t},
\end{aligned}
\tag{2.1}
$$

where: each $R_{t\ell}$ is a subset of $\mathcal{X}_t$ with the restriction that $R_{tL_t} = \mathcal{X}_t$; $a_{t\ell} \in \mathcal{A}_t$; $\ell = 1, \ldots, L_t$; and L_t is the length of π_t. Thus, a compact representation of π_t is $\{(R_{t\ell}, a_{t\ell})\}_{\ell=1}^{L_t}$. To increase interpretability, we restrict $R_{t\ell}$ to clauses involving thresholding with at most

two covariates, hence $R_{t\ell}$ is an element of

$$\begin{aligned}\mathcal{R}_t = \{&\mathcal{X}_t,\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1\},\\ &\{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} \leq \tau_2\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} > \tau_2\},\\ &\{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1 \text{ and } x_{j_2} \leq \tau_2\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1 \text{ and } x_{j_2} > \tau_2\} :\\ &1 \leq j_1 < j_2 \leq d_t, \tau_1, \tau_2 \in \mathbb{R}\}, \end{aligned} \tag{2.2}$$

where j_1, j_2 are indices and τ_1, τ_2 are thresholds. We also impose an upper bound, $L_{\max}$, on list length L_t for all t. Hence, the class of regimes of interest is $\Pi = \otimes_{t=1}^{T} \Pi_t$, where $\Pi_t = \{\{R_{t\ell}, a_{t\ell}\}_{\ell=1}^{L_t} : R_{t\ell} \in \mathcal{R}_t, a_{t\ell} \in \mathcal{A}_t, L_t \leq L_{\max}\}$.

Remark 1. We omit sets of the form $\{\boldsymbol{x}_t \in \mathcal{X}_t : x_{j_1} \leq \tau_1 \text{ or } x_{j_2} \leq \tau_2\}$ in the definition of $\mathcal{R}_t$ because such sets are expressible in terms of the sets already in $\mathcal{R}_t$. For example, the clause "if $\boldsymbol{x}_t \in R_{t1}$ then a_{t1}" with $R_{t1} = \{\boldsymbol{x}_t \in \mathcal{X}_t : x_{tj_1} \leq \tau_1 \text{ or } x_{tj_2} \leq \tau_2\}$ can be written as "if $\boldsymbol{x}_t \in R'_{t1}$ then a_{t1}; else if $\boldsymbol{x}_t \in R'_{t2}$ then a_{t1}" with $R'_{t1} = \{\boldsymbol{x}_t \in \mathcal{X}_t : x_{tj_1} \leq \tau_1\}$ and $R'_{t2} = \{\boldsymbol{x}_t \in \mathcal{X}_t : x_{tj_2} \leq \tau_2\}$. Moreover, the latter form has the benefit of avoiding the measurement of x_{j_2} for subjects satisfying $x_{j_1} \leq \tau_1$, which may be an important consideration if x_{j_2} refers to some biomarker that is expensive to measure (see Zhang et al., 2015, for discussion of decision lists and measurement cost).

Remark 2. Under certain generative models, distinct sets in $\mathcal{R}_t$ may correspond to the same group of subjects with probability one. For example, if X_{t1} takes values in $\{0, 1\}$, the set $\{\boldsymbol{x} \in \mathcal{X}_t : x_1 \leq 0\}$ and the set $\{\boldsymbol{x} \in \mathcal{X}_t : x_1 \leq 0.5\}$ correspond to the same group of subjects. To address this issue, it is tempting to require the threshold for x_1 to take values in the support of X_{t1}. Nevertheless, such requirement is not sufficient to

ensure that different sets in $\mathcal{R}_t$ correspond to different groups of subjects. To see this, suppose $(X_{t1}, X_{t2})^{\mathrm{T}}$ can take three possible values: $(0,0)^{\mathrm{T}}$, $(1,0)^{\mathrm{T}}$ and $(1,1)^{\mathrm{T}}$, e.g., if X_{t1} and X_{t2} are indicators of two symptoms where the second symptom can be present only when the first symptom is present. In this case, the set $\{\boldsymbol{x} \in \mathcal{X}_t : x_1 \leq 0\}$ and the set $\{\boldsymbol{x} \in \mathcal{X}_t : x_1 \leq 0 \text{ and } x_2 \leq 0\}$ correspond to the same group of subjects. Therefore, we allow the thresholds to take arbitrary values. In our theoretical analysis, we quantify dissimilarity of sets in $\mathcal{R}_t$ using a distance that accounts for the distribution of $\boldsymbol{X}_t$.

To estimate $\boldsymbol{\pi}^{\text{opt}}$ we combine non-parametric Q-learning with policy-search (see Taylor et al., 2015, for a discussion of this idea in the context of single decision point). To develop our ideas, we first provide a high-level schematic for our algorithm, then we describe implementation and modeling details, and finally we discuss a computational insight that improves computation time.

Define $Q_T(\boldsymbol{x}_T, a_T) = \mathbb{E}(Y_T | \boldsymbol{X}_T = \boldsymbol{x}_T, A_T = a_T)$ then it can be shown that $\pi_T^{\text{opt}} = \arg\max_{\pi \in \Pi_T} \mathbb{E} Q_T \{\boldsymbol{X}_T, \pi(\boldsymbol{X}_T)\}$. Recursively, for $t = T-1, \ldots, 1$ define $Q_t(\boldsymbol{x}_t, a_t) = \mathbb{E}\left[Y_t + Q_{t+1}\left\{\boldsymbol{X}_{t+1}, \pi_{t+1}^{\text{opt}}(\boldsymbol{X}_{t+1})\right\} \middle| \boldsymbol{X}_t = \boldsymbol{x}_t, A_t = a_t\right]$ and subsequently it can be shown that $\pi_t^{\text{opt}} = \arg\max_{\pi_t \in \Pi_t} \mathbb{E} Q_t \{\boldsymbol{X}_t, \pi_t(\boldsymbol{X}_t)\}$ (Schulte et al., 2014). For each $t = 1, \ldots, T$ let $\mathcal{Q}_t$ denote a postulated class of models for Q_t. Q-learning with policy-search follows directly from the foregoing definitions; a schematic is as follows.

(S1) Construct an estimator of Q_T in $\mathcal{Q}_T$, e.g., one could use penalized least squares $\widehat{Q}_T = \arg\min_{Q_T \in \mathcal{Q}_T} \sum_{i=1}^n \{Y_{iT} - Q_T(\boldsymbol{X}_{iT}, A_{iT})\}^2 + \mathcal{P}_T(Q_T)$, where $\mathcal{P}_T(Q_T)$ is a penalty on the complexity of Q_T. Define $\widehat{\pi}_T = \arg\max_{\pi \in \Pi_T} \sum_{i=1}^n \widehat{Q}_T \{\boldsymbol{X}_{iT}, \pi_T(\boldsymbol{X}_{iT})\}$.

(S2) Recursively, for $t = T-1, \ldots, 1$ construct an estimator of Q_t in $\mathcal{Q}_t$, say $\widehat{Q}_t$, e.g.,

$$\widehat{Q}_t = \arg\min_{Q_t \in \mathcal{Q}_t} \sum_{i=1}^{n} \left\{ Y_{it} + \widehat{Q}_{t+1}\left\{ \boldsymbol{X}_{i(t+1)}, \widehat{\pi}_{t+1}(\boldsymbol{X}_{i(t+1)}) \right\} - Q_t(\boldsymbol{X}_{it}, A_{it}) \right\}^2 + \mathcal{P}_t(Q_t),$$

where $\mathcal{P}_t(Q_t)$ is a penalty on the complexity of Q_t. Define $\widehat{\pi}_t = \arg\max_{\pi_t \in \Pi_t} \sum_{i=1}^{n} \widehat{Q}_t\{\boldsymbol{X}_{it}, \pi_t(\boldsymbol{X}_{it})\}$.

Implementation of the preceding schematic requires a choice of models for the Q-functions, a means of constructing an estimator within this class, and an algorithm for computing $\arg\max_{\pi_t \in \Pi_t} \sum_{i=1}^{n} \widehat{Q}_t\{\boldsymbol{X}_{it}, \pi_t(\boldsymbol{X}_{it})\}$. In our implementation, we use kernel ridge regression with an extended Gaussian kernel to construct estimators of the Q-functions and a greedy stepwise algorithm to approximate $\widehat{\pi}_t$ from the estimated Q-functions.

2.2.2 Kernel Ridge Regression

We use kernel ridge regression to estimate the Q-functions. Starting with the last stage, let $K_T(\cdot,\cdot)$ be a symmetric and positive definite function from $\mathbb{R}^{d_T} \times \mathbb{R}^{d_T}$ to $\mathbb{R}$, and let $\mathbb{H}_T$ be the corresponding reproducing kernel Hilbert space (RKHS). In our implementation, we employ an extension of the Gaussian kernel that uses different scaling factors in different variables: $K_T(\boldsymbol{x}, \boldsymbol{z}) = \exp\left\{ -\sum_{j=1}^{d_T} \gamma_{Tj}(x_j - z_j)^2 \right\}$, where $\boldsymbol{\gamma}_T = (\gamma_{T1}, \ldots, \gamma_{Td_T})^{\mathrm{T}}$ is a tuning parameter and $\gamma_{Tj} > 0$ for all j. For each $a \in \mathcal{A}_T$, we estimate $Q_T(\cdot, a)$ via penalized least squares

$$\widehat{Q}_T(\cdot, a) = \arg\min_{f \in \mathbb{H}_T} n^{-1} \sum_{i \in \mathcal{I}_{Ta}} \{Y_{iT} - f(\boldsymbol{X}_{iT})\}^2 + \lambda_T \|f\|^2_{\mathbb{H}_T},$$

where $\mathcal{I}_{Ta} = \{i : A_{iT} = a\}$, and $\lambda_T > 0$ is a tuning parameter. Let $\boldsymbol{Y}_{Ta} = (Y_{iT})_{i \in \mathcal{I}_{Ta}}$ and $\boldsymbol{K}_{Ta} = \{K(\boldsymbol{X}_{iT}, \boldsymbol{X}_{jT})\}_{i,j \in \mathcal{I}_{Ta}}$. By the representer theorem (Kimeldorf and Wahba, 1971), $\widehat{Q}_T(\boldsymbol{x}, a) = \sum_{i \in \mathcal{I}_{Ta}} K_T(\boldsymbol{x}, \boldsymbol{X}_{iT}) \widehat{\beta}_{iTa}$, where $\widehat{\boldsymbol{\beta}}_{Ta} = (\widehat{\beta}_{iTa})_{i \in \mathcal{I}_{Ta}}$ satisfy $\widehat{\boldsymbol{\beta}}_{Ta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y}_{Ta} - \boldsymbol{K}_{Ta}\boldsymbol{\beta}\|^2 + n\lambda_T \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{K}_{Ta} \boldsymbol{\beta}$. Define $\widehat{\pi}_T = \arg\max_{\pi_T \in \Pi_T} \sum_{i=1}^n \widehat{Q}_T \{\boldsymbol{X}_{Ti}, \pi_T(\boldsymbol{X}_{Ti})\}$.

Similarly, for each $t < T$ let $\mathbb{H}_t$ be the RKHS induced by the kernel $K_t(\boldsymbol{x}, \boldsymbol{z}) = \exp\left\{-\sum_{j=1}^{d_t} \gamma_{tj}(x_j - z_j)^2\right\}$, and $\boldsymbol{\gamma}_t = (\gamma_{t1}, \ldots, \gamma_{td_t})^{\mathrm{T}}$ is a tuning parameter. Recursively, for each $t < T$, $a_t \in \mathcal{A}_t$, estimate $Q_t(\cdot, a)$ by

$$\widehat{Q}_t(\cdot, a) = \arg\min_{f \in \mathbb{H}_t} n^{-1} \sum_{i \in \mathcal{I}_{ta}} \left[Y_{it} + \widehat{Q}_{t+1}\{\boldsymbol{X}_{i,t+1}, \widehat{\pi}_{t+1}(\boldsymbol{X}_{i,t+1})\} - f(\boldsymbol{X}_{it})\right]^2 + \lambda_t \|f\|^2_{\mathbb{H}_t},$$

where $\mathcal{I}_{ta} = \{i : A_{it} = a\}$, and λ_t is a tuning parameter.

2.2.3 Construction of Decision Lists

In addition to a method for estimating the Q-functions, the proposed method requires a method for computing $\arg\max_{\pi_t \in \Pi_t} \sum_{i=1}^n \widehat{Q}_t \{\boldsymbol{X}_{ti}, \widehat{\pi}_t(\boldsymbol{X}_{ti})\}$ where Π_t is the space of list-based decision rules defined previously. Any element in Π_t can be expressed as $\{(R_{t\ell}, a_{t\ell})\}_{\ell=1}^{L_t}$, however, simultaneous optimization over all regions and treatments is not computationally feasible except in very small problems. Instead, we propose an algorithm that constructs $\widehat{\pi}_t$ using a greedy optimization procedure that optimizes one clause in $\widehat{\pi}_t$ at a time; unlike many greed algorithms, the proposed method is consistent for the global maximizer. To provide intuition, we describe in detail the first two steps of this greedy algorithm before stating it in more general terms.

Estimation of the First Clause

Define $\widehat{\pi}_t^Q$ to be map $\boldsymbol{x}_t \mapsto \arg\max_{a_t \mathcal{A}_t} \widehat{Q}_t(\boldsymbol{x}_t, a_t)$; thus, $\widehat{\pi}_t^Q$ is optimal estimated decision rule at stage t using non-parametric Q-learning. To estimate the first clause (R_{t1}, a_{t1}) in π_t, we consider the following decision-list parameterized by R and a:

$$\begin{aligned} &\texttt{If } \boldsymbol{x}_t \in R \texttt{ then } a; \\ &\texttt{else if } \boldsymbol{x}_t \in \mathcal{X}_t \texttt{ then } \widehat{\pi}_t^Q(\boldsymbol{x}_t). \end{aligned} \tag{2.3}$$

If all subjects follow (2.3), the estimated mean outcome is

$$\frac{1}{n}\sum_{i=1}^{n}\left[I(\boldsymbol{X}_{it} \in R)\widehat{Q}_t(\boldsymbol{X}_{it}, a) + I(\boldsymbol{X}_{it} \notin R)\widehat{Q}_t\{\boldsymbol{X}_{it}, \widehat{\pi}_t^Q(\boldsymbol{X}_{it})\}\right]. \tag{2.4}$$

Hence, we can pick the maximizer of (2.4) as the estimator of (R_{t1}, a_{t1}). Note that the difference between the estimated mean outcome under $\widehat{\pi}_t^Q$ and that under (2.3) is $n^{-1}\sum_{i=1}^{n} I(\boldsymbol{X}_{it} \in R)\left[\widehat{Q}_t\{\boldsymbol{X}_{it}, \widehat{\pi}_t^Q(\boldsymbol{X}_{it})\} - \widehat{Q}_t(\boldsymbol{X}_{it}, a)\right]$, which measures the decrease in the estimated mean outcome when some part of $\widehat{\pi}_t^Q$ is replaced with an if-then clause. This represents the price paid for interpretability, and by maximizing (2.4), we minimize this price.

To improve generalization performance, we add a complexity penalty to (2.3); in addition to encouraging parsimonious lists, we shall see that this penalty also ensures a unique maximizer. Define $V(R) \in \{0, 1, 2\}$ to be the number of covariates needed to check inclusion in R. We define $\widehat{R}_1$ and $\widehat{a}_1$ as the maximizers over R and a in

$$\frac{1}{n}\sum_{i=1}^{n}\left[I(\boldsymbol{X}_{it}\in R)\widehat{Q}_t(\boldsymbol{X}_{it},a)+I(\boldsymbol{X}_{it}\notin R)\widehat{Q}_t\{\boldsymbol{X}_{it},\widehat{\pi}_t^Q(\boldsymbol{X}_{it})\}\right] + \zeta\left\{\frac{1}{n}\sum_{i=1}^{n}I(\boldsymbol{X}_{it}\in R)\right\}+\eta\{2-V(R)\}, \quad (2.5)$$

where $\zeta,\eta>0$ are tuning parameters. Thus, the first penalty term rewards regions R with lots of mass relative to the distribution of $\boldsymbol{X}_t$ whereas the second term rewards regions that involve fewer covariates. Moreover, we impose the constraint $n^{-1}\sum_{i=1}^{n}I(\boldsymbol{X}_{it}\in R)>0$ to avoid searching over vacuous clauses.

Estimation of the Second Clause

To estimate the second clause we consider the following decision list parameterized by R and a

$$\begin{aligned}&\texttt{If } \boldsymbol{x}\in\widehat{R}_{t1} \texttt{ then } \widehat{a}_{t1};\\ &\texttt{else if } \boldsymbol{x}\in R \texttt{ then } a;\\ &\texttt{else if } \boldsymbol{x}\in\mathcal{X}_t \texttt{ then } \widehat{\pi}_t^Q(\boldsymbol{x}).\end{aligned} \quad (2.6)$$

If all the subjects follow the regime (2.6), the estimated mean outcome is

$$\frac{1}{n}\sum_{i=1}^{n}I(\boldsymbol{X}_{it}\in\widehat{R}_{t1})\widehat{Q}_t(\boldsymbol{X}_{it},\widehat{a}_{t1})+\frac{1}{n}\sum_{i=1}^{n}I(\boldsymbol{X}_{it}\notin\widehat{R}_{t1},\boldsymbol{X}_{it}\in R)\widehat{Q}_t(\boldsymbol{X}_{it},a) + \frac{1}{n}\sum_{i=1}^{n}I(\boldsymbol{X}_{it}\notin\widehat{R}_{t1},\boldsymbol{X}_{it}\notin R)\widehat{Q}_t\{\boldsymbol{X}_{it},\widehat{\pi}_t^Q(\boldsymbol{X}_{it})\}. \quad (2.7)$$

Note that the first term in (2.7) can be dropped during the optimization as it is independent of R and a. As in (2.5), we maximize the penalized criterion

$$\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_{it} \notin \widehat{R}_{t1}, \boldsymbol{X}_{it} \in R)\widehat{Q}_t(\boldsymbol{X}_{it}, a) + \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_{it} \notin \widehat{R}_{t1}, \boldsymbol{X}_{it} \notin R)\widehat{Q}_t\{\boldsymbol{X}_{it}, \widehat{\pi}_t^Q(\boldsymbol{X}_{it})\} + \zeta\left\{\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_{it} \notin \widehat{R}_{t1}, \boldsymbol{X}_{it} \in R)\right\} + \eta\{2 - V(R)\} \quad (2.8)$$

with respect to $R \in \mathcal{R}_t, a \in \mathcal{A}_t$ and subject to the constraint $n^{-1}\sum_{i=1}^{n} I(\boldsymbol{X}_{it} \notin \widehat{R}_{t1}, \boldsymbol{X}_{it} \in R) > 0$. We continue this procedure until either every subject gets a recommended treatment, namely $R_{t\ell} = \mathcal{X}_t$ for some ℓ, or the maximum length is reached, $\ell = L_{\max}$. If the maximum list length is reached, we set $R_{tL_{\max}} = \mathcal{X}_t$ to ensure that the regime applies to every subject and choose $\widehat{a}_{tL_{\max}}$ be the estimated best single treatment for all remaining subjects.

Estimation of All Clauses

An algorithmic description of the proposed algorithm is given below. Additional computational details, including the time complexity, are given in the next section.

Step 1. Initialize $\ell = 1$.

Step 2. If $\ell < L_{\max}$, compute

$$(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \underset{R\in\mathcal{R}_t, a\in\mathcal{A}_t}{\arg\max} \frac{1}{n}\sum_{i=1}^{n}\Big[I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \in R)\widehat{Q}_t(\boldsymbol{X}_{it}, a) + I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \notin R)\widehat{Q}\{\boldsymbol{X}_{it}, \widehat{\pi}_t^Q(\boldsymbol{X}_{it})\}\Big]$$

$$+\zeta\left\{\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_{it}\in\widehat{G}_{t\ell},\boldsymbol{X}_{it}\in R)\right\}+\eta\{2-V(R)\} \quad (2.9)$$

subject to $n^{-1}I(\boldsymbol{X}_{it}\in\widehat{G}_{t\ell},\boldsymbol{X}_{it}\in R)>0$, where $\widehat{G}_{t1}=\mathcal{X}_t$, $\widehat{G}_{t\ell}=\mathcal{X}_t\setminus\big(\bigcup_{k<\ell}\widehat{R}_{tk}\big)$ for $\ell\geq 2$, and $V(R)\in\{0,1,2\}$ is the number of variables used to define R. It is easy to verify that the objective function above reduces to (2.5) when $\ell=1$ and to (2.8) when $\ell=2$. If $\ell=L_{\max}$, set

$$(\widehat{R}_{t\ell},\widehat{a}_{t\ell})=\underset{R\in\mathcal{R}_t,a\in\mathcal{A}_t}{\arg\max}\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_{it}\in\widehat{G}_{t\ell})\widehat{Q}(\boldsymbol{X}_{it},a)+\eta\{2-V(R)\}. \quad (2.10)$$

The solution of (2.10) must satisfy $V(R)=0$ and hence $\widehat{R}_{t\ell}=\mathcal{X}_t$. Consequently the last clause does apply to all the rest subjects.

Step 3. If $\widehat{R}_{t\ell}=\mathcal{X}_t$ then go to Step 4; otherwise, increase ℓ by 1 and repeat Steps 2 and 3.

Step 4. Output $\widehat{\pi}_t=\{(\widehat{R}_{tk},\widehat{a}_{tk})\}_{k=1}^{\ell}$.

Implementation Details and Time Complexity

Computation of $(\widehat{R}_{t\ell},\widehat{a}_{t\ell})$ in (2.9) requires special attention because the objective function is non-differentiable and non-convex. We first argue that brute-force search can be used to obtain $(\widehat{R}_{t\ell},\widehat{a}_{t\ell})$. Although $\mathcal{R}_t$ contains infinitely many elements, because the objective function in (2.9) is piecewise linear, for each covariate it suffices to consider n thresholds located at the order statistics of that covariate. Hence, the number of thresholds to enumerate is of order n^2. In addition, there are $d_t(d_t+1)/2$ choices for variables in $R_{t\ell}$, m_t choices for $a_{t\ell}$, and each evaluation of (2.9) takes $O(n)$ operations. Therefore, the

time complexity for finding $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell})$ via brute-force search is $O(n^3 d_t^2 m_t)$. Unfortunately, the factor n^3 is overwhelming even when the sample size n is moderate.

Instead of brute-force search, we propose a novel algorithm to compute $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell})$, that substantially reduces the time complexity. Note that the n^3 factor is due to the enumeration of thresholds and the evaluation of the objective function in (2.9). By reorganizing the enumeration and evaluation, the proposed algorithm reduces the n^3 factor to $n \log n$. Thus, with this implementation, the proposed algorithm can be applied to large datasets; this is appealing in an era of 'big-data' where large data-bases are being mined generate hypotheses about precision medicine.

Proposition 1. *For each t and ℓ, the estimator $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell})$ in (2.9) can be computed within $O(n \log n\, d_t^2 m_t)$ operations.*

The proof of this result is constructive but technical so we provide a sketch of the main idea here and relegate the remaining details to the Supplemental Materials. Suppose R involves only one covariate: $R = \{\boldsymbol{x} : x_j \leq \tau\}$. For fixed t, j and a, we observe that, up to a constant independent of τ, the objective function in (2.9) is of the form $F(\tau) = n^{-1} \sum_{i=1}^{n} I(X_{ijt} \leq \tau) U_i + I(X_{ijt} > \tau) V_i$, where U_i and V_i are constants. As discussed previously, we only need to compute $F(\tau)$ for τ equal to observed covariate values, X_{ijt}. Let $i_1 < \cdots < i_n$ be a permutation of $1, \ldots, n$ such that $X_{i_1 jt} \leq \cdots \leq X_{i_n jt}$. Then, it can be shown that $F(X_{i_s jt}) = F(X_{i_{s-1} jt}) + U_{i_s} - V_{i_s}$, $s \geq 2$. Hence, one can enumerate all possible values for τ and evaluate $F(\tau)$ in $O(n)$ time, in contrast to $O(n^2)$ time for brute-force search. A similar recursive relationship can be established if R is of the form $\{\boldsymbol{x} : x_j > \tau\}$. When R involves two covariates, we combine this sorting technique with binary search tree (Cormen et al., 2009), which enables us to find the thresholds in

$O(n \log n)$ time.

Remark 3. The proposed algorithm differs from that in Zhang et al. (2015) in two important ways. First, the two algorithms maximize different objective functions. In Zhang et al. (2015), regime (2.3) is replaced by "if $\boldsymbol{x} \in R$ then a; else if $\boldsymbol{x} \in \mathcal{X}_t$ then a'", where R, a and a' are obtained by maximizing the estimated mean outcome under such a regime. However, this criteria fails to account for subsequent splits in the decision lists and can thereby get stuck in a local mode. In contrast, the proposed algorithm approximates the remaining list with the estimated optimal regime using non-parametric Q-learning. To illustrate the difference between the two objective functions, consider a scenario with $T = 1$ stage, a single covariate $S_1 \sim \text{Uniform}(-2, 2)$ and suppose that $\widehat{Q}_1(x, a) = Q_1(x, a) = ax(x - 1), a \in \{-1, 1\}$. Assume ζ and η are small but positive. Then the solution of (2.9) is $\widehat{R}_{11} = \{x : x \leq \tau\}$ and $\widehat{a}_{11} = 1$ with $\tau \approx 0$. Nevertheless, if the term $\widehat{\pi}_t^Q(\boldsymbol{X}_{it})$ were replaced by a fixed treatment $a' \neq a$, the solution would be $\widehat{R}_{11} = \mathcal{X}_1$ and $\widehat{a}_{11} = 1$, leading to a suboptimal regime. A second difference between the proposed algorithm and the one proposed in Zhang et al. (2015) is that the latter requires a pre-specified set of candidate thresholds for each predictor, and its time complexity is the same as brute-force search if we use all the unique values as candidate thresholds.

2.3 Theoretical Results

For each $\ell = 1, 2, \ldots,$ define the population analogs of (2.9) and (2.10) as follows, $(R^*_{t\ell}, a^*_{t\ell}) = \arg\max_{R \in \mathcal{R}_t, a \in \mathcal{A}_t} \Psi_{t\ell}(R, a)$, where

$$\Psi_{t\ell}(R,a) = E\left[I(\boldsymbol{X}_t \in G^*_{t\ell}, \boldsymbol{X}_t \in R)Q(X_t,a) + I(\boldsymbol{X}_t \in G^*_{t\ell}, \boldsymbol{X}_t \notin R)Q\left\{\boldsymbol{X}_t, \pi_t^Q(\boldsymbol{X}_t)\right\}\right] + \zeta \Pr(\boldsymbol{X}_t \in G^*_{t\ell}, \boldsymbol{X}_t \in R) + \eta\{2 - V(R)\}, \quad (2.11)$$

and $G^*_{t\ell} = \mathcal{X}_t$ if $\ell = 1$ and $G^*_{t\ell} = \mathcal{X}_t \setminus (\cup_{k<\ell} R^*_{tk})$ otherwise, until either $R^*_{t\ell} = \mathcal{X}_t$ or $\ell = L_{\max}$. In the latter case, instead of (2.11) we define

$$\Psi_{t\ell}(R,a) = E\left\{I(X_t \in G^*_{t\ell})Q(\boldsymbol{X}_t,a)\right\} + \eta\{2 - V(R)\}. \quad (2.12)$$

Let $L^*_t = \min\{\ell : R^*_{t\ell} = \mathcal{X}_t\}$ and $\pi^*_t = \left\{(R^*_{t\ell}, a^*_{t\ell})\right\}_{\ell=1}^{L^*_t}$. In (2.11) and (2.12), the Q-functions are defined as $Q_T(\boldsymbol{x},a) = E(Y_T|\boldsymbol{X}_T = \boldsymbol{x}, A_T = a)$, $Q_t(\boldsymbol{x},a) = E[Y_t + Q_{t+1}\{\boldsymbol{X}_{t+1}, \pi^*_{t+1}(\boldsymbol{X}_{t+1})\}|\boldsymbol{X}_t = \boldsymbol{x}, A_t = a]$ for $t = T-1, \ldots, 1$. Furthermore, let $\pi_t^Q(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}_t} Q_t(\boldsymbol{x},a)$ for all t.

We assume that all the covariates and outcomes are bounded. This is a common assumption in the context of nonparametric regression; the extension to include unbounded covariates is possible but at the expense of additional complexity.

Assumption 1. There exists $B > 0$ such that $\|\boldsymbol{X}_t\|_\infty \le B$ and $|Y_t| \le B$ with probability one for all $t = 1, \ldots, T$.

Under the boundedness assumption, it is natural to limit the value of $\widehat{Q}_t(\cdot,a)$, $a \in \mathcal{A}_t$ inside the interval $[-B, B]$. Namely, in the algorithm $\widehat{Q}_t(\cdot,a)$ is replaced by $\mathcal{T}_B\{\widehat{Q}_t(\cdot,a)\}$, where $\mathcal{T}_B$ is defined as

$$\mathcal{T}_B(f)(\boldsymbol{x}) = f(\boldsymbol{x})I\{-B \le f(\boldsymbol{x}) \le B\} + BI\{f(\boldsymbol{x}) > B\} + (-B)I\{f(\boldsymbol{x}) < -B\}.$$

Besides, we can also require $\zeta \in (0, B]$, since we have $\widehat{R}_{t\ell} = \mathcal{X}_t$ in equation (2.9) whenever

$\zeta \geq B$.

We also assume positivity (Robins, 2004), which ensures that $Q_t(\boldsymbol{x}, a)$ is well-defined for all $a \in \mathcal{A}_t$.

Assumption 2. For each t and $a \in \mathcal{A}_t$, $\Pr(A_t = a|\boldsymbol{X}_t) \geq \varpi$ almost surely for some positive constant ϖ.

A crucial intermediate step in deriving the asymptotic behavior of $\widehat{\pi}_t$'s is establishing convergence of $\widehat{Q}_t$ to Q_t; to facilitate this step we require a certain degree of smoothness in Q_t. A common means of imposing smoothness is to assume differentiability (see, e.g., Stone, 1982). However, the non-differentiable maximization operator that is implicit in the definition of the Q-functions forces us to consider a weaker notion of smoothness. Denote $B_t = [-b, b]^{d_t} \subset \mathbb{R}^{d_t}$. For any function $f : B_t \to \mathbb{R}$, define the r-th difference $\Delta^r_{\boldsymbol{h}}(f)$ by $\Delta^r_{\boldsymbol{h}}(f)(\boldsymbol{x}) = \sum_{i=0}^r \binom{r}{i}(-1)^{r-i} f(\boldsymbol{x} + i\boldsymbol{h})$ if $\boldsymbol{x} \in B_{t,r,\boldsymbol{h}}$ and 0 otherwise, where r is a positive integer, $\boldsymbol{h} = (h_1, \ldots, h_{d_t})^T$, $h_j \geq 0$ for all $j = 1, \ldots, d_t$, and $B_{t,r,\boldsymbol{h}} = \{\boldsymbol{x} \in B_t : \boldsymbol{x} + i\boldsymbol{h} \in B_t \text{ for all } i \leq r\}$. Define the r-th modulus of smoothness of f by $\omega_r(f, s) = \sup_{\|\boldsymbol{h}\|_2 \leq s} \sup_{\boldsymbol{x} \in B_t} |\Delta^r_{\boldsymbol{h}}(f)(\boldsymbol{x})|$. The definition above is similar to Eberts and Steinwart (2013, Definition 2.1), but replaces the L_p norm with the supremum norm. This modification allows us to drop the requirement that $\boldsymbol{X}_t$ have a density with respect to Lebesgue measure. Hence, our analysis applies when $\boldsymbol{X}_t$ contains discrete covariates.

The concept of modulus of smoothness generalizes the concept of differentiability. To see this, consider an example where $d = 1$. We observe that $\lim_{h \to 0} h^{-1} \Delta^1_h(f)(x) = f'(x)$. Suppose $|f'(x)|$ is bounded, then for sufficiently small h, there exists a constant C_f such that $|\Delta^1_h(f)(x)| \leq C_f |h|$. Hence, any continuously differentiable function f, defined on a finite interval, satisfies $\omega_1(f, s) = O(s)$, as $s \to 0$. Generally, if f is r-times continuously

differentiable, then $w_r(f, s) = O(s^r)$ as $s \to 0$. In addition, some non-differentiable functions also satisfy this condition. Consider $f(x) = |x|$ and $f(x) = \max(x, 0)$. It is easy to verify that $|\Delta_h^1(f)(x)| \leq |h|$ for any x. Thus $f(x) = |x|$ and $f(x) = \max(x, 0)$ also satisfy $\omega_1(f, s) = O(s)$ though f is not differentiable at 0. We make the following assumption regarding the smoothness of the Q-functions.

Assumption 3. For each t, there exists a positive integer r_t such that $\omega_r\{Q_t^*(\cdot, a), s\} = O(s^{r_t})$ as $s \to 0$, for any $a \in \mathcal{A}_t$.

In order to study the probabilistic convergence of $\widehat{\pi}_t$ to π_t^*, it is necessary to define an appropriate distance between $\widehat{R}_{t\ell}$ and $R_{t\ell}^*$. In view of Remark 2, the distance should incorporate the distribution of $\boldsymbol{X}_t$, thus, we define $\rho_t(R_1, R_2) = \Pr\{\boldsymbol{X}_t \in (R_1 \triangle R_2)\}$, where $C \triangle D$ denotes the symmetric set difference between sets C and D. It can be verified that $\rho_t(R_1, R_2)$ is non-negative, symmetric, and satisfies the triangle inequality. Note that $\rho_t(R_1, R_2) = 0$ only indicates that R_1 and R_2 refer to the same group of subjects with probability one with respect to $\boldsymbol{X}_t$ but does not imply $R_1 = R_2$. For example, suppose $X_{t\ell}$ takes values in $\{0, 1\}$, $R_1 = \{x : x \leq 0\}$ and $R_2 = \{x : x \leq 0.5\}$. Then $\rho_t(R_1, R_2) = 0$, as expected. Furthermore, the use of ρ_t helps to avoid the issue of non-unique representations of R when some covariates can be expressed using others. For example, if $X_{t1} = -X_{t2}$ and both are continuous, then $\rho_t(\{\boldsymbol{x} : x_1 \leq \tau\}, \{\boldsymbol{x} : x_2 > \tau\}) = 0$. Thus, our goal is to identify an equivalence class of clauses that each describe the same subset of patients. We require the following identifiability assumption on the equivalence class of optimal clauses.

Assumption 4. For each t and ℓ, the following inequalities hold:

(i) There exists a constant $\kappa > 0$ such that $\Psi_{t\ell}(R, a_{t\ell}^*) \leq \Psi_{t\ell}(R_{t\ell}^*, a_{t\ell}^*) - \kappa\rho_t^2(R, R_{t\ell}^*)$ as

$\rho_t(R, R^*_{t\ell}) \to 0$;

(ii) For any $\delta > 0$, there exists a constant $\epsilon > 0$ such that $\Psi_{t\ell}(R, a^*_{t\ell}) \leq \Psi_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \epsilon$ for all $R \in \mathcal{R}_t$ with $\rho_t(R, R^*_{t\ell}) > \delta$;

(iii) There exists a constant $\varsigma > 0$ such that $\Psi_{t\ell}(R, a) \leq \Psi_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \varsigma$ for all $R \in \mathcal{R}_t$ and $a \in \mathcal{A}_t \setminus \{a^*_{t\ell}\}$.

Assumption 4 guarantees the uniqueness of $(R^*_{t\ell}, a^*_{t\ell})$ in the sense that if $(\widetilde{R}^*_{t\ell}, \widetilde{a}^*_{t\ell})$ is another maximizer of $\Psi_{t\ell}(R, a)$, then $\rho(R^*_{t\ell}, \widetilde{R}^*_{t\ell}) = 0$ and $a^*_{t\ell} = \widetilde{a}^*_{t\ell}$. Moreover, condition (i) assumes that $\Psi(R, a^*_{t\ell})$ behaves like a quadratic function in a neighborhood of $R^*_{t\ell}$. When $\boldsymbol{X}_t$ has bounded density and R, $R^*_{t\ell}$ use the same covariates, $\Psi_{t\ell}$ can be viewed as a function of the threshold values and condition (i) implies that $\Psi_{t\ell}$ behaves like a quadratic function near the optimal threshold values, which is a common condition in parametric models.

Define the value of a decision rule at time t, say π_t, as $V_t(\pi_t) = \mathrm{E}\left[Q_t\{\boldsymbol{X}_t, \pi_t(\boldsymbol{X}_t)\}\right]$. Define the dissimilarity measure between π_t and π^*_t as $M_t(\pi_t) = \Pr\left\{\pi_t(\boldsymbol{X}_t) \neq \pi^*_t(\boldsymbol{X}_t)\right\}$. Our analysis focuses on how close $\widehat{\pi}_t$ is to π^*_t, and how well $\widehat{\pi}_t$ performs compared to π^*_t in terms of value. As $\boldsymbol{\gamma}_t$, λ_t, ζ and η may depend on n, we may write $\boldsymbol{\gamma}_{n,t}$, $\lambda_{n,t}$, ζ_n and η_n to emphasize such dependence. The following theorem establishes finite sample bounds. A proof is given in the Supplemental Materials.

Theorem 1. *Define* $\phi_T = (2/3)^{L^*_T} r_T/(2r_T+d_T)$, *and* $\phi_t = (2/3)^{L^*_t} \min\{r_t/(2r_t+d_t), \phi_{t+1}\}$ *for* $t = T-1, \ldots, 1$, *where* L^*_t *is the length of* π^*_t. *Assume* $\max_j \gamma_{n,t,j} = \overline{\theta}_{t,\gamma} n^{2/(2r_t+d_t)}$, $\min_j \gamma_{n,t,j} = \underline{\theta}_{t,\gamma} n^{2/(2r_t+d_t)}$, $\lambda_{n,t} = \theta_{t,\lambda} n^{-1}$, $\sup_n \zeta_n < \infty$, *and* $\sup_n \eta_n < \infty$, *where*

$\overline{\theta}_{t,\gamma}, \underline{\theta}_{t,\gamma}, \theta_{t,\lambda}$ are positive constants. Under Assumptions 1-4, for any $\tau > 0$, we have

$$\Pr\{M_t(\widehat{\pi}_t) \geq \tau\} \leq c_1 \exp(-c_2 n^{\phi_t - \xi}\tau),$$
$$\Pr\{V_t(\pi_t^*) - V_t(\widehat{\pi}_t) \geq \tau\} \leq c_3 \exp(-c_4 n^{\phi_t - \xi}\tau),$$

where c_i's are positive constants independent of n and τ, and ξ is an arbitrary positive number.

Also, let $\mathcal{S}_t$ be the indices of the signal variables defining the function Q_t, and define $\phi_t^{\mathcal{S}}$ in the same way as ϕ_t but with d_t replaced by $|\mathcal{S}_t|$. If $\max_{j\in\mathcal{S}} \gamma_{n,t,j} = \overline{\theta}_{\mathcal{S},t,\gamma} n^{2/(2r_t+|\mathcal{S}_t|)}$, $\min_{j\in\mathcal{S}} \gamma_{n,t,j} = \underline{\theta}_{\mathcal{S},t,\gamma} n^{2/(2r_t+|\mathcal{S}_t|)}$, $\max_{j\in\mathcal{S}^c} \gamma_{n,t,j} = \overline{\theta}_{\mathcal{S}^c,t,\gamma}$, where $\overline{\theta}_{\mathcal{S},t,\gamma}, \underline{\theta}_{\mathcal{S},t,\gamma}, \overline{\theta}_{\mathcal{S}^c,t,\gamma}$ are positive constants, then the inequalities above holds with ϕ_t replaced by $\phi_t^{\mathcal{S}}$.

The minimax convergence rate for a nonparametric regression estimator of an r_T-times continuously differentiable function is $O\{n^{-(2r_T)/(2r_T+d_T)}\}$ (Stone, 1982). By extending the technique in Eberts and Steinwart (2013), we show in the Supplemental Material that the estimated Q-function $\widehat{Q}_T$ converges to its true value Q_T at a nearly optimal rate $O\{n^{-(2r_T)/(2r_T+d_T)+\xi'}\}$, where $\xi' > 0$ can be arbitrarily small. The construction of $\widehat{\pi}_T$ involves estimating L_T^* pairs of parameters $(R_{t\ell}, a_{t\ell})$, one pair for each if-then clause. The estimation of each pair $(R_{t\ell}, a_{t\ell})$ reduces the convergence rate by an additional factor of 2/3. The underlying idea for this phenomenon is analogous to the problems analyzed in Kim and Pollard (1990). In earlier stages, the estimation of Q-functions is further complicated by the fact that $Q_{t+1}\{\boldsymbol{X}_{t+1}, \pi_{t+1}(\boldsymbol{X}_{t+1})\}$ is not observed but estimated via $\widehat{Q}_{t+1}\{\boldsymbol{X}_{t+1}, \widehat{\pi}_{t+1}(\boldsymbol{X}_{t+1})\}$.

When all the covariates are discrete, it can be shown that the convergence rate of

the estimated regime $\widehat{\pi}_t$ does not inherit the slow convergence rate from the underlying nonparametric regressions. The following result is proved in the Supplemental Material.

Theorem 2. *Assume that the distribution of $\boldsymbol{X}_t$ is discrete. Namely, for each t there exists a finite set $\widetilde{\mathcal{X}}_t$ such that $\Pr(\boldsymbol{X}_t \in \widetilde{\mathcal{X}}_t) = 1$. Assume $\max_j \gamma_{n,t,j} = \overline{\theta}_{t,\gamma} n^{2/(2r_t+d_t)}$, $\min_j \gamma_{n,t,j} = \underline{\theta}_{t,\gamma} n^{2/(2r_t+d_t)}$, $\lambda_{n,t} = \theta_{t,\lambda} n^{-1}$, $\sup_n \zeta_n < \infty$, and $\sup_n \eta_n < \infty$, where $\overline{\theta}_{t,\gamma}, \underline{\theta}_{t,\gamma}, \theta_{t,\lambda}$ are positive constants. Under Assumptions 1-4, we have*

$$\Pr\{M_t(\widehat{\pi}_t) > 0\} \leq c_1 \exp(-c_2 n),$$
$$\Pr\{V_t(\pi_t^*) - V_t(\widehat{\pi}_t) > 0\} \leq c_3 \exp(-c_4 n),$$

where c_i's are positive constants independent of n. In addition, the condition on $\boldsymbol{\gamma}_{n,t}$ and $\lambda_{n,t}$ can be relaxed as long as the kernel ridge regression maintains consistent.

In both theorems, the convergence rates are independent of ζ_n and η_n. However, the choice of ζ_n and η_n has an impact on the limiting DTR π_t^*. In practice, we suggest to tune $\boldsymbol{\gamma}_n$ and λ_n by minimizing the cross validated mean squared error in the kernel ridge regression, and tune ζ_n and η_n by maximizing the cross validated value of the regime.

2.4 Simulation Studies

We conducted a series of simulation experiments to examine the empirical performance of the proposed method. Five scenarios were considered. The first four came from Zhao et al. (2015a) and the fifth was adapted from Murphy (2003). Scenario I consists of two stages, two treatment options at each stage, and the covariates exhibit nonlin-

ear effects: $\boldsymbol{S}_1 = (S_{1,1}, \ldots, S_{1,50})^{\mathrm{T}}$, are independent standard normal random variables; A_1 is Uniform$\{-1, 1\}$; Y_1 is Normal$(\mu_1, 1)$, where $\mu_1 = 0.5S_{1,3}A_1$; S_2 is empty; A_2 is Uniform$\{-1, 1\}$; Y_2 is Normal$(\mu_2, 1)$ where $\mu_2 = \{(S_{1,1}^2 + S_{1,2}^2 - 0.2)(0.5 - S_{1,1}^2 - S_{1,2}^2) + Y_1\}A_2$. Scenario II consists of time-varying covariates. In this scenario: $\boldsymbol{S}_1$, A_1 and A_2 were generated in the same way as scenario I; Y_1 is Normal$(\mu_1, 1)$, where $\mu_1 = (1 + 1.5S_{1,3})A_1$, $\boldsymbol{S}_2 = (S_{2,1}, S_{2,2})^{\mathrm{T}}$; $S_{2,1}$ is Bernoulli with success probability $1 - \Phi(1.25S_{1,1}A_1)$; $S_{2,2}$ is Bernoulli with success probability $1 - \Phi(-1.75S_{1,2}A_1)$; and Y_2 is Normal$(\mu_2, 1)$, where $\mu_2 = (0.5 + Y_1 + 0.5A_1 + 0.5S_{2,1} - 0.5S_{2,2})A_2$. In Scenario III, A_1, A_2, A_3 are Uniform$\{-1, 1\}^3$; $S_{1,1}, S_{1,2}, S_{1,3}$ are *i.i.d.* Normal$(45, 15^2)$; S_2 is Normal$(1.5S_{1,1}, 10^2)$; S_3 is Normal$(0.5S_2, 10^2)$; $Y_1 = Y_2 = 0$, and Y_3 is Normal$(\mu_3, 1)$, where $\mu_3 = 20 - |0.6S_{1,1} - 40|\{I(A_1 > 0) - I(S_{1,1} > 30)\}^2 - |0.8S_2 - 60|\{I(A_2 > 0) - I(S_2 > 40)\}^2 - |1.4S_3 - 40|\{I(A_3 > 0) - I(S_3 > 40)\}^2$. Scenario IV is the same as Scenario III except that many noise variables were added. In addition to $S_{1,1}$, $S_{1,2}$ and $S_{1,3}$, we generated $S_{1,4}, \ldots, S_{1,50}$ *i.i.d.* from Normal$(45, 15^2)$. Scenario V involves ten stages and multiple treatment options at each stage. See Murphy (2003) for background and motivation for this scenario. For $t \in \{1, \ldots, 10\}$, treatments were coded as a pair of values $A_t = (A_{t1}, A_{t2})^{\mathrm{T}}$, generated as follows. First, A_{t1} is drawn from Uniform$\{0, 1\}$. Second, if $A_{t1} = 0$ then A_{t2} is drawn from Uniform$\{0, 1, 2, 3\}$, and otherwise A_{t2} is drawn uniformly from $\{1, 2, 3\}$. Thus, there are $m_t = 7$ treatment candidates at each decision point. In addition, $U_1, \ldots, U_{10}$ are *i.i.d.* Normal$(0, 0.01)$; $S_1 = 0.5 + U_1$; $S_t = 0.5 + 0.2S_{t-1} - 0.07A_{t-1,1}A_{t-1,2} - 0.01(1 - A_{t-1,1})A_{t-1,2} + U_t$ for $t \geq 2$; Y_t is Normal$(\mu_t, 0.64)$, where $\mu_t = 30I(t = 1) - 5U_t - 6\{A_{t1} - I(S_t > 5/9)\}^2 - 1.5A_{t1}(A_{t2} - 2S_t)^2 - 1.5(1 - A_{t1})(A_{t2} - 5.5S_t)^2$ for each t.

In each scenario, we considered sample sizes $n = 100, 200$, and 400. We generated

1000 data sets for each sample size and estimated the optimal DTR using the proposed method. In each stage, we tuned the scaling vector in the Gaussian kernel, $\boldsymbol{\gamma}_t$, as well as the amount of penalty, λ_t, via leave-one-out cross validation. The cross validated error was minimized via a Quasi-Newton type algorithm (Kim et al., 2010) with a random starting value. During the construction of decision lists, at each decision point we tuned ζ and η via five-fold cross validation over a pre-specified grid. We picked the combination that led to the largest cross validated outcome.

To form a basis for comparison, we also implemented Q-learning with linear models, non-parametric Q-learning with random forests, backward outcome weighted learning (BOWL), and simultaneous outcome weighted learning (SOWL; Zhao et al., 2015a). In Q-learning with linear Q-functions, we fit the working $Q_t(\boldsymbol{X}_t, A_t) = \sum_{a \in \mathcal{A}_t} I(A_t = a)\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{t,a}$. Motivated by Qian and Murphy (2011), we imposed an ℓ_1 penalty to reduce overfitting. The Q-functions were estimated by ℓ_1 regularized least squares, implemented in the `R` package `glmnet` (Friedman et al., 2010). The covariates were standardized to have mean zero and variance one before entering the model, and the tuning parameter was selected by five-fold cross validation. Our implementation of non-parametric Q-learning used the `R` package `randomForest` with default parameters settings (Liaw and Wiener, 2002). We implemented BOWL and SOWL according to the descriptions in Zhao et al. (2015a). Linear kernels were used, and the amount of regularization was chosen by five-fold cross validation. Note that BOWL and SOWL assume binary treatment options and thus are not applicable in Scenario V.

We measure the quality of an estimated DTR by the mean outcome under that DTR; we approximate this mean outcome using an independent test set of size 10^5. The results

are displayed in Table 1. In Scenario I, the second stage Q-function is highly nonlinear, and most methods tended to assign a single treatment to all patients in the second stage, leading to a mean outcome of 6.70. In contrast, the proposed method is able to correctly individualize treatment as the sample size increased and thus produce a higher mean outcome. In Scenario II, both Q-functions at the first and the second stages are linear. Hence, as expected, Q-learning with linear models performs best. Nevertheless, the proposed method and non-parametric Q-learning perform well and shows marked improvement over BOWL and SOWL. In Scenario III, Q-learning with linear models suffers from model misspecification whereas the proposed method and non-parametric Q-learning both perform well. Furthermore, although the Q-functions are complicated, the optimal DTR consists of linear functions of covariates. Hence, both BOWL and SOWL perform well in this scenario. Recall that scenario IV is the same as scenarioIII except for the addition of many noise variables. Thus, the results for scenario IV demosntrate a sensitivity to noise variables in BOWL and SOWL. One possible reason for this is that both BOWL and SOWL utilizes ℓ_2 penalties, which fails to exclude noise variables. In Scenario V, the proposed method outperforms competing methods, especially when the sample size is small. The reason might be due to the nonparametric estimation of Q-functions and the simple form of decision list compared to a random forest, as simpler DTRs tends to have better generalizability.

2.5 Data Analysis

As an illustration of the proposed method, we use data from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) to estimate an interpretable

Table 2.1 Simulation results. Given a scenario and a sample size, each method constructed 1000 DTRs, one per each simulated dataset. The number in each cell is the outcome under the estimated DTR, averaged over 1000 replications, with standard deviation in parentheses. In the header, n is the sample size, DL refers to the proposed decision list based approach, Q-lasso refers to the Q-learning approach with linear model and lasso penalty, Q-RF refers to the Q-learning approach using random forest

Scenario	n	DL	Q-lasso	Q-RF	BOWL	SOWL
1	100	6.63 (0.24)	6.55 (0.58)	6.70 (0.05)	6.70 (0.05)	6.70 (0.05)
1	200	6.73 (0.24)	6.64 (0.33)	6.70 (0.05)	6.70 (0.05)	6.70 (0.05)
1	400	6.94 (0.16)	6.66 (0.26)	6.70 (0.05)	6.70 (0.05)	6.70 (0.05)
2	100	3.66 (0.10)	3.68 (0.08)	3.41 (0.17)	3.15 (0.05)	2.77 (0.52)
2	200	3.71 (0.04)	3.73 (0.04)	3.62 (0.12)	3.22 (0.08)	2.84 (0.33)
2	400	3.73 (0.03)	3.75 (0.02)	3.71 (0.04)	3.37 (0.14)	2.91 (0.28)
3	100	14.49 (2.77)	5.42 (4.54)	12.94 (2.07)	10.65 (2.40)	10.27 (2.33)
3	200	17.42 (1.42)	7.88 (1.63)	15.79 (1.59)	13.09 (2.20)	12.98 (1.88)
3	400	18.60 (0.71)	8.41 (0.65)	18.02 (0.73)	15.33 (1.56)	16.22 (1.58)
4	100	13.38 (3.14)	4.54 (5.17)	11.47 (2.31)	6.72 (1.71)	6.04 (2.18)
4	200	17.33 (1.87)	7.69 (2.33)	14.82 (1.75)	8.90 (1.13)	8.34 (1.99)
4	400	18.84 (0.70)	8.61 (0.96)	17.04 (1.02)	10.75 (0.68)	9.38 (2.30)
5	100	23.68 (1.09)	12.97 (3.40)	17.83 (1.63)	–	–
5	200	25.94 (0.51)	13.80 (2.57)	21.60 (1.28)	–	–
5	400	26.80 (0.29)	16.65 (1.71)	24.73 (0.65)	–	–

DTR for treating bipolar disorder (Sachs et al., 2003). We focus on the randomized acute depression (RAD) pathway in STEP-BD, which is a Sequential Multiple Assignment Randomized Trial (SMART) and provides the data needed to build DTRs. One purpose of STEP-BD is to assess the effectiveness of adding antidepressants to mood stabilizers in treating patients with bipolar disorder. Although antidepressants were often assigned to supplement mood stabilizers in practice, it was found that the adjunctive antidepressant medication did not show much improvement over the use of mood stabilizers alone (Sachs et al., 2007). Thus, it is of scientific interest to tailor the use of antidepressants based on individual and time-dependent characteristics.

The RAD pathway in STEP-BD is a randomized trials with two stages. At both stages, patients always received one or more mood stabilizers chosen by their psychiatrists. In addition, they might receive one antidepressant in the form of bupropion or paroxetine. At week 0, patients were randomized to receive bupropion, paroxetine or placebo with probability 0.25, 0.25 and 0.5, respectively. After 6 weeks, patients returned to their psychiatrists for evaluation on response status. In another 6 weeks, responders continued their initial treatments, non-responders who received either bupropion or paroxetine initially were offered an increased dose, and non-responders who received placebo initially were randomized to received bupropion or paroxetine with equal probability. At week 12, patients returned to their psychiatrists for final measurements.

In this clinical trial, the covariate $\boldsymbol{X}_1$ of a patient consists of his/her age, gender, marital status, education level, employment status, bipolar type, nature of the episode prior to the current depressive episode, summary score for depression (SUM-D) at baseline, and summary score for mood elevation (SUM-ME) at baseline. The treatment A_1 takes

three values: bupropion, paroxetine and placebo. The covariate $\boldsymbol{X}_2$ consists of SUM-D at week 6, SUM-ME at week 6, and indicators for nine different adverse events at week 6. The treatment A_2 is either bupropion or paroxetine for non-responders who received placebo in the first stage. For other patients, A_2 is the same as A_1. The outcomes are $Y_1 = 0$ and $Y_2 =$ SUM-D at week 12. Note that smaller values of SUM-D and SUM-ME indicates better clinical status. A complete description of these variables is provided in the Supplementary Materials.

We apply the propose method to estimate an interpretable DTR. For simplicity, we only include patients with complete baseline and stage 1 information. And we use the last-value-carry-forward strategy if the SUM-D at week 12 is missing. The estimated optimal decision rule at the first stage is:

`If` SUM-D at week $0 > 8.625$ `then` bupropion;

`else if` SUM-D at week $0 \leq 4.875$ `and` race is not white `then` paroxetine;

`else` placebo.

The estimated regime suggests that the baseline SUM-D is informative in treatment selection. Recall that smaller values of SUM-D indicate lower symptoms. Hence, an interpretation of the estimated regime is: patients with severe depression symptoms should receive bupropion, while non-white patients with minor depression symptoms should receive paroxetine. Although applying an antidepressant medication to all patients did not lead to a better mean outcome relative to not applying antidepressants to any of the patients (Sachs et al., 2007), the estimated regime indicates that personalizing the use of

antidepressants based on SUM-D may improve the overall mean outcome. The estimated optimal decision rule at the second stage is:

`If` SUM-ME at week $6 \leq 0.875$ `then` bupropion;

`else if` SUM-D at week $6 > 8.5$ `then` bupropion;

`else` paroxetine.

From this rule it can be seen that patients with large SUM-D or low SUM-ME are assigned to buproprion.

2.6 Discussion

The current trend in methodological research for estimation of optimal treatment regimes seems to be the development of increasingly flexible models to mitigate risk of model misspecification. This trend is aligned with the notion that an estimated optimal regime will be used to make treatment decisions for future patients. However, in many settings an estimated optimal regime is not used to make treatment decisions but rather is used to generate hypotheses and inform future research. Indeed, our view is that the development of a precision medicine strategy should be the culmination of an iterative process of hypothesis generation and validation. With this perspective, the ability to interpret and estimated optimal regime in a domain context is paramount.

We used list-based regimes to ensure interpretability of the estimated regimes. Our proposed estimation algorithm combines non-parametric Q-learning with policy-search and consistently estimates the optimal regime under mild assumptions. In principle, the

proposed estimation framework could be used to estimate interpretable optimal regimes of other forms, e.g., more general tree structures or rule-based systems. Nevertheless, the simplicity of list-based regimes that ensures parsimony and interpretability also appears to have regularizing effect that improves generalization performance.

The recognition that estimated optimal regimes are often not used directly to select treatments for patients but instead are part of an iterative, collaborative process opens many new lines of research beyond estimation of interpretable regimes. These include methods for visualization, models for shared-decision making, models for patient preference and utility construction, and methods for constructing prediction sets for outcome trajectories in multistage decision problems. We are currently pursuing several of these research areas.

Chapter 3

R Package DTRList: Estimation of List-based Dynamic Treatment Regimes

3.1 Introduction

In personalized medicine or intervention science, people are interested in making individualized and time-adaptive treatment decisions. Dynamic treatment regimes (DTRs) formalize the treatment decision process as a sequence of rules, one per each stage. Each rule maps the available information to a recommended treatment. When the construction of DTRs is for exploratory purposes such as generating new hypotheses for future research or validation, the interpretability of the rules may become a concern. To address the interpretability issue, Zhang et al. (2015) propose a list-based approach to the esti-

mation of DTRs. The resulting rules are known as decision lists, which consist of a short list of "if-then" clauses and use thresholding conditions. They are readily interpretable and hence provide an appealing choice in many applications.

In this article, we describe the `R` package `DTRList`, which implements the estimation of list-based DTRs. The package employs Q-learning with policy search. It allows an arbitrary number of stages and an arbitrary number of treatment options at each stage. In Section 2, we briefly describe the statistical methodology of estimating list-based DTRs. In Section 3, we present a detailed example to illustrate the usage of the package. In Section 4, we consider the estimation of list-based rules for classification problems. The estimation of DTRs can be viewed as a reinforcement learning problem. Nevertheless, the method can be easily adapted to handle supervised learning problems. Hence we provide an example in Section 4.

3.2 Estimation of List-based Regimes

3.2.1 List-based Regimes

We are interested in list-based regimes. Formally, each decision rule π_t has the form:

$$
\begin{aligned}
&\texttt{If } \boldsymbol{x}_t \in R_{t1} \texttt{ then } a_{t1};\\
&\texttt{else if } \boldsymbol{x}_t \in R_{t2} \texttt{ then } a_{t2};\\
&\ldots\\
&\texttt{else if } \boldsymbol{x}_t \in R_{tL_t} \texttt{ then } a_{tL_t},
\end{aligned}
$$

where each $R_{t\ell}$ is a subset of $\mathcal{X}_t$ with the restriction that $R_{tL_t} = \mathcal{X}_t$; $a_{t\ell} \in \mathcal{A}_t$; $\ell = 1, \ldots, L_t$; and L_t is the length of π_t. To increase interpretability, we restrict $R_{t\ell}$ to clauses involving thresholding with at most two covariates, hence $R_{t\ell}$ is an element of

$$\begin{aligned}\mathcal{R}_t = \{&\mathcal{X}_t,\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1\},\\ &\{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} \leq \tau_2\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} > \tau_2\},\\ &\{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1 \text{ and } x_{j_2} \leq \tau_2\},\ \{\boldsymbol{x} \in \mathcal{X}_t : x_{j_1} > \tau_1 \text{ and } x_{j_2} > \tau_2\} :\\ &1 \leq j_1 < j_2 \leq d_t, \tau_1, \tau_2 \in \mathbb{R}\},\end{aligned}$$

where j_1, j_2 are indices and τ_1, τ_2 are thresholds. We also impose an upper bound, $L_{\max}$, on list length L_t for all t.

3.2.2 One-stage Problems

We assume that the observed data are $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$, which comprise n independent identically distributed observations, one for each subject in an randomized or observational study. Let $(\boldsymbol{X}, A, Y)$ be a generic observation. Here $\boldsymbol{X} \in \mathbb{R}^p$ are baseline subject covariates; $A \in \mathcal{A} = \{1, \ldots, m\}$ is the treatment received; and $Y \in \mathbb{R}$ is the outcome, coded so that higher values are better. A treatment regime, π, is a function from $\mathbb{R}^p$ into $\mathcal{A}$, so that under π a patient presenting with $\boldsymbol{X} = \boldsymbol{x}$ is recommended treatment $\pi(\boldsymbol{x})$. The value function is defined as $V(\pi) = \mathrm{E}^\pi Y$, which measures the expected value of the outcome if the population of subject followed the regime π.

Define $\omega(\boldsymbol{x}, a) = \Pr(A = a | \boldsymbol{X} = \boldsymbol{x})$ and $Q(\boldsymbol{x}, a) = \mathrm{E}(Y | \boldsymbol{X} = \boldsymbol{x}, A = a)$. It can be

shown that

$$V(\pi) = \mathrm{E}\left(\sum_{a=1}^{m} Q(\boldsymbol{X}, a) I\left\{\pi(\boldsymbol{X}) = a\right\}\right),$$

and

$$V(\pi) = \mathrm{E}\left(\sum_{a=1}^{m} \left[\frac{I(A=a)}{\omega(\boldsymbol{X}, a)}\left\{Y - Q(\boldsymbol{X}, a)\right\} + Q(\boldsymbol{X}, a)\right] I\left\{\pi(\boldsymbol{X}) = a\right\}\right).$$

They lead to two strategies of estimating $V(\pi)$, known as Q-learning and inverse probability weighting.

The Q-learning estimator is

$$\widehat{V}^{Q}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} Q(\boldsymbol{X}_i, a) I\left\{\pi(\boldsymbol{X}_i) = a\right\}.$$

The inverse probability weighting estimator is

$$\widehat{V}^{IPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} \left\{\frac{I(A_i = a)}{\omega(\boldsymbol{X}_i, a)} Y_i\right\} I\left\{\pi(\boldsymbol{X}_i) = a\right\}.$$

The augmented inverse probability weighting estimator is

$$\widehat{V}^{AIPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} \left[\frac{I(A_i = a)}{\omega(\boldsymbol{X}_i, a)}\left\{Y_i - Q(\boldsymbol{X}_i, a)\right\} + Q(\boldsymbol{X}_i, a)\right] I\left\{\pi(\boldsymbol{X}_i) = a\right\}.$$

All of them have a common form $\widehat{V}(\pi) = n^{-1}\sum_i \sum_a V_{ia} I\{\pi(\boldsymbol{X}_i) = a\}$. Next, we optimize $\widehat{V}(\pi)$ over the class of decision lists to construct the regime.

3.2.3 Multiple-stage Problems

We assume that the observed data are $\{(\boldsymbol{S}_{it}, A_{it}, Y_{it}) : t = 1, \ldots, T\}_{i=1}^{n}$, which comprise *i.i.d.* replicates of the form $\{(\boldsymbol{S}_t, A_t, Y_t) : t = 1, \ldots, T\}$. Here $\boldsymbol{S}_t \in \mathbb{R}^{p_t}$ is a vector of covariates measured at the beginning of the t-th stage, $A_t \in \mathcal{A}_t$ is the treatment actually received during the t-th stage, and $Y_t \in \mathbb{R}$ is a scalar outcome measured at the end of the t-th stage. The final outcome of interest is the sum of immediate outcomes, $Y = \sum_{t=1}^{T} Y_t$. We assume that larger values of Y are better. Let $\boldsymbol{X}_t$ denote the information available at stage t. Hence $\boldsymbol{X}_1 = \boldsymbol{S}_1$ and $\boldsymbol{X}_t = (\boldsymbol{X}_{t-1}^{\mathrm{T}}, A_{t-1}, Y_{t-1}, \boldsymbol{S}_t^{\mathrm{T}})^{\mathrm{T}}$ for $t > 1$. Denote the support of $\boldsymbol{X}_t$ by $\mathcal{X}_t$.

A treatment regime $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ is a sequence of functions $\pi_t : \mathcal{X}_t \to \mathcal{A}_t$ so that under $\boldsymbol{\pi}$ a patient presenting with $\boldsymbol{X}_t = \boldsymbol{x}_t$ at stage t is recommended treatment $\pi_t(\boldsymbol{x}_t)$.

To estimate $\boldsymbol{\pi}$, we employ Q-learning with policy search. It goes as follows.

(Step 1) We construct an estimator of Q_T. For example, we could use penalized least squares $\widehat{Q}_T = \arg\min_{Q_T} \sum_{i=1}^{n} \{Y_{iT} - Q_T(\boldsymbol{X}_{iT}, A_{iT})\}^2 + \mathcal{P}_T(Q_T)$, where $\mathcal{P}_T(Q_T)$ is a penalty on the complexity of Q_T. Define $\widehat{V}_T(\pi_T) = n^{-1} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_T} \widehat{Q}_T(\boldsymbol{X}_{iT}, a) I\{\pi_T(\boldsymbol{X}_{iT}) = a\}$. We then construct $\widehat{\pi}_T$ to maximize $\widehat{V}_T(\pi_T)$ over the class of decision lists.

(Step 2) Recursively, for $t = T-1, \ldots, 1$ we construct an estimator of Q_t. For example, we may take

$$\widehat{Q}_t = \arg\min_{Q_t} \sum_{i=1}^{n} \left\{ Y_{it} + \widehat{Q}_{t+1}\left\{\boldsymbol{X}_{i(t+1)}, \widehat{\pi}_{t+1}(\boldsymbol{X}_{i(t+1)})\right\} - Q_t(\boldsymbol{X}_{it}, A_{it}) \right\}^2 + \mathcal{P}_t(Q_t),$$

where $\mathcal{P}_t(Q_t)$ is a penalty on the complexity of Q_t. Define V_t and $\widehat{\pi}_t$ similarly to Step 1.

Note that for each t, the estimated value function still has the form

$$\widehat{V}_t(\pi) = n^{-1} \sum_i \sum_a V_{iat} I\{\pi(\boldsymbol{X}_{it}) = a\}.$$

3.2.4 Estimation of Decision Lists

Fix t. We would like to construct π_t so that $\widehat{V}_t(\pi_t)$ is as large as possible. Since simultaneous estimation of all "if-then" clauses in π_t incurs intensive computational burden, we will estimate of "if-then" clauses one-by-one. The procedure goes as follows.

Step 1. Initialize $\ell = 1$.

Step 2. If $\ell < L_{\max}$, compute

$$(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \underset{R \in \mathcal{R}_t, a \in \mathcal{A}_t}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \left[I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \in R) V_{iat} + I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \notin R) \overline{V}_{it} \right] + \zeta \left\{ \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \in R) \right\} + \eta\{2 - V(R)\}$$

subject to $n^{-1} I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}, \boldsymbol{X}_{it} \in R) > 0$, where $\overline{V}_{it} = \max_{a \in \mathcal{A}_t} V_{iat}$, $\widehat{G}_{t1} = \mathcal{X}_t$, $\widehat{G}_{t\ell} = \mathcal{X}_t \setminus \left(\bigcup_{k<\ell} \widehat{R}_{tk} \right)$ for $\ell \geq 2$, and $V(R) \in \{0, 1, 2\}$ is the number of variables used to define R. If $\ell = L_{\max}$, set

$$(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \underset{R \in \mathcal{R}_t, a \in \mathcal{A}_t}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}) V_{iat} + \eta\{2 - V(R)\}.$$

The solution must satisfy $V(R) = 0$ and hence $\widehat{R}_{t\ell} = \mathcal{X}_t$.

Step 3. If $\widehat{R}_{t\ell} = \mathcal{X}_t$ then go to Step 4; otherwise, increase ℓ by 1 and repeat

Steps 2 and 3.

Step 4. Output $\widehat{\pi}_t = \{(\widehat{R}_{tk}, \widehat{a}_{tk})\}_{k=1}^{\ell}$.

3.3 Using the `DTRList` Package

We illustrate the usage of the `DTRList` Package via a simulated dataset, which mimics data from an actual clinical trial which studies the effect of meal replacement shakes on adolescent obesity. The trial has two stages. At each stage, there are two treatment options: meal replacement (MR) and conventional diet (CD). At the beginning of the first stage, subject covariates including gender, race, parent BMI, baseline BMI were measured. At the beginning of the second stage, month4 BMI was measured. The primary outcome is month12 BMI, which was measured at the end of the second stage. The data was first included in the `iqLearn` package; see Linn et al. (2015) for more information.

After installing the `DTRList` package, we load the library as well as the BMI dataset.

```
> library(DTRList)
> data("bmiData")
```

Next we prepare the data and estimate a DTR. Here `x` is the collection of $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T$ and `stage.x` gives the number of stage that the covariate is measured. The treatment matrix `a` and the response matrix y should have T columns, one column per each stage.

```
> x <- as.matrix(bmiData[, c("gender", "race",
+ "parentBMI", "baselineBMI", "month4BMI")])
> stage.x <- c(1, 1, 1, 1, 2)
```

```
> a <- bmiData[, c("A1", "A2")]
> y <- cbind(0, bmiData[, "month12BMI"])
> dtr <- dtrList(y, a, x, stage.x)
> dtr
Stage 1:
If parentBMI > 34.83 then MR;
else if parentBMI <= 31.57 then CD;
else if baselineBMI <= 41.14 then MR;
else CD.

Stage 2:
If month4BMI > 34.91 then CD;
else MR.
```

For visualization, we can plot a flowchart.

```
> plot(dtr)
```

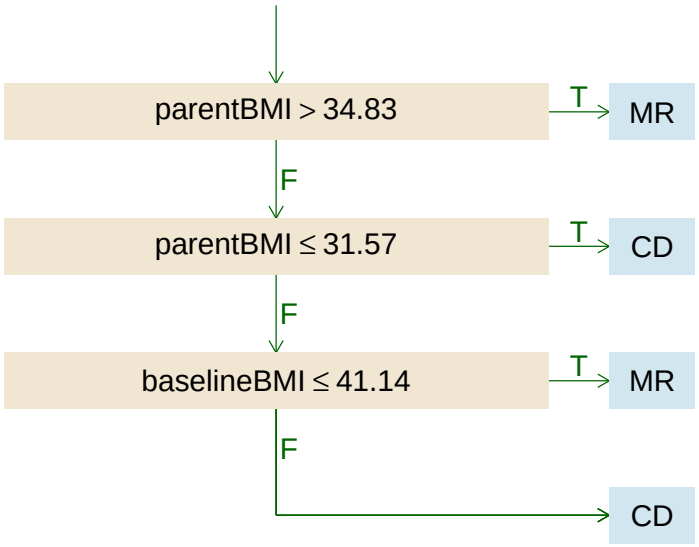


Figure 3.1 Estimated list-based DTR for the BMI dataset.

BIBLIOGRAPHY

Ashley, E. A. (2015). The precision medicine initiative: a new national effort. *Journal of the American Statistical Association*, 313(21):2119–2120.

Baker, S. G., Cook, N. R., Vickers, A., and Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):729–748.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford Univeristy Press.

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus de l'Académie des Sciences, Series I*, 334(6):495–500.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. New York: CRC Press.

Brusco, M. J. and Stahl, S. (2006). *Branch-and-Bound Applications in Combinatorial Data Analysis*. New York: Springer.

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT Press and McGraw-Hill, 3 edition.

Doove, L., Dusseldorp, E., Van Deun, K., and Van Mechelen, I. (2015). A novel method for estimating optimal tree-based treatment regimes in randomized clinical trials. Technical report.

Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7:1–42.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372.

Gail, M. H. (2009). Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute*, 101(13):959–963.

Gail, M. H., Costantino, J. P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K., and Vogel, V. (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute*, 91(21):1829–1846.

Gunter, L., Zhu, J., and Murphy, S. A. (2011). Variable selection for qualitative interactions. *Statistical Methodology*, 8(1):42–55.

Huang, Y., Laber, E. B., and Janes, H. (2015). Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics*, 16(2):383–399.

Jameson, J. L. and Longo, D. L. (2015). Precision medicinepersonalized, problematic, and promising. *New England Journal of Medicine*, 372(23):2229–2234.

Kang, C., Janes, H., and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70(3):695–707.

Keller, M. B., McCullough, J. P., Klein, D. N., Arnow, B., Dunner, D. L., Gelenberg, A. J., et al. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *New England Journal of Medicine*, 342(20):1462–1470.

Kim, D., Sra, S., and Dhillon, I. S. (2010). Tackling box-constrained optimization via a new projected quasi-newton approach. *SIAM Journal on Scientific Computing*, 32(6):3548–3563.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7):1163–1170.

Laber, E., Zhao, Y., Regh, T., Davidian, M., Tsiatis, A. A., Stanford, J. B., Zeng, D., and Kosorok, M. R. (2016). Sizing a phase ii trial to find a nearly optimal personalized treatment strategy. *Statistics in Medicine*, page in press.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014). Interactive model building for Q-learning. *Biometrika*, 101(4):831–847.

Laber, E. B. and Zhao, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. A. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8(1):21–48.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2012). Building interpretable classifiers with rules using Bayesian analysis. Technical Report TR609, Department of Statistics, University of Washington.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3):18–22.

Linn, K., Laber, E., and Stefanski, L. (2015). iqlearn: Interactive q-learning in r. *Journal of Statistical Software*, 64(1):1–25.

Marchand, M. and Sokolova, M. (2005). Learning with decision lists of data-dependent features. *Journal of Machine Learning Research*, 6:427–451.

Marlowe, D. B., Festinger, D. S., Dugosh, K. L., Benasutti, K. M., Fox, G., and Croft, J. R. (2012). Adaptive programming improves outcomes in drug court an experimental trial. *Criminal justice and behavior*, 39(4):514–532.

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4):629–645.

Moodie, E. E. M., Dean, N., and Sun, Y. R. (2013). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6:1–21.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *The International Journal of Biostatistics*, 6(2).

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180–1210.

Rivest, R. L. (1987). Learning decision lists. *Machine Learning*, 2(3):229–246.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In Lin, D. Y. and Heagerty, P. J., editors, *Proceedings of the Second Seattle Symposium in Biostatistics*, volume 179 of *Lecture Notes in Statistics*, pages 189–326. New York: Springer.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721.

Sachs, G. S., Nierenberg, A. A., Calabrese, J. R., Marangell, L. B., Wisniewski, S. R., Gyulai, L., Friedman, E. S., Bowden, C. L., Fossey, M. D., Ostacher, M. J., Ketter, T. A., Patel, J., Hauser, P., Rapport, D., Martinez, J. M., Allen, M. H., Miklowitz, D. J., Otto, M. W., Dennehy, E. B., and Thase, M. E. (2007). Effectiveness of adjunctive antidepressant treatment for bipolar depression. *New England Journal of Medicine*, 356(17):1711–1722.

Sachs, G. S., Thase, M. E., Otto, M. W., Bauer, M., Miklowitz, D., Wisniewski, S. R., Lavori, P., Lebowitz, B., Rudorfer, M., Frank, E., Nierenberg, A. A., Fava, M., Bowden, C., Ketter, T., Marangell, L., Calabrese, J., Kupfer, D., and Rosenbaum, J. F. (2003). Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biological Psychiatry*, 53(11):1028–1042.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 29(4):640–661.

Shiffman, R. N. (1997). Representation of clinical practice guidelines in conventional and augmented decision tables. *Journal of the American Medical Informatics Association*, 4(5):382–393.

Shortreed, S. M., Laber, E. B., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. (2011). Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine Learning*, 84(109–236):109–136.

Shortreed, S. M., Laber, E. B., Stroup, T. S., Pineau, J., and Murphy, S. A. (2014). A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine*, 33(24):4202–4214.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer-Verlag.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158.

Taylor, J. M. G., Cheng, W., and Foster, J. C. (2015). Reader reaction to "a robust method for estimating optimal treatment regimes" by Zhang et al. (2012). *Biometrics*, 71(1):267–273.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer-Verlag.

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731.

Wang, F. and Rudin, C. (2015). Falling rule lists. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 1013–1022.

Xu, Y., Müller, P., Wahed, A. S., and Thall, P. F. (2015a). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association*, page in press.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., and Shao, J. (2015b). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, 71(3):645–653.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012b). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012c). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015a). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.

Zhao, Y. Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015b). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Zhou, X. and Kosorok, M. R. (2016). Nearest neighbor rules for optimal treatment regimes. page under review.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, page in press.

APPENDICES

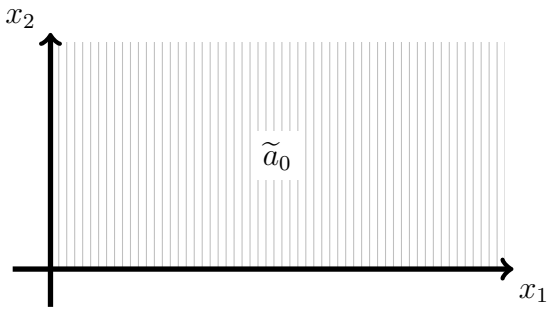
Appendix A

Supplementary Materials for Chapter 1

A.1 An Illustrative Run Through the Algorithm for Finding an Optimal Decision List

In this section, we illustrate how the proposed algorithm for finding an optimal decision list works. For simplicity, the patient covariate is assumed to be two-dimensional.

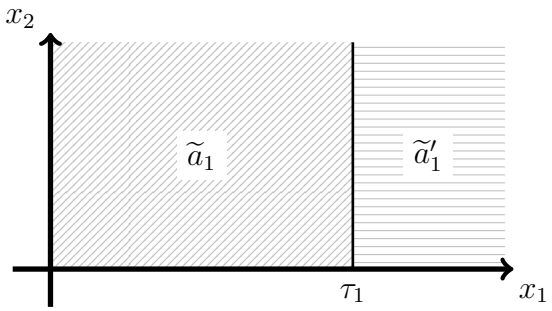
- The algorithm starts at Step 1.
  - We choose $L_{\max} = 5$ and $\alpha = 0.05$.
  - We compute $\widetilde{a}_0 = \arg\max_{a_0 \in \mathcal{A}} \widehat{R}\left[\{a_0\}\right]$. Suppose the maximum found is $\widehat{R}\left[\{\widetilde{a}_0\}\right] = 10$. Figure A.1 shows the decision list $\{\widetilde{a}_0\}$.
  - We set $\Pi_{\text{temp}} = \emptyset$ and $\Pi_{\text{final}} = \emptyset$.



Everyone $\widetilde{a}_0$.

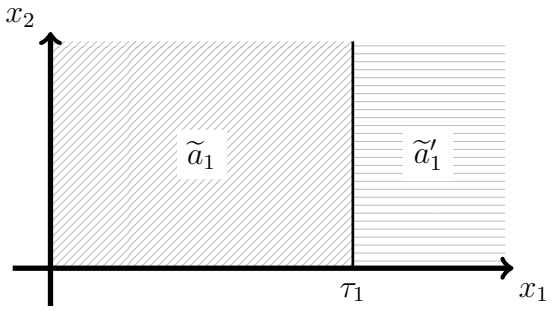
Figure A.1 Diagram and description of the decision list $\{\widetilde{a}_0\}$.

- The algorithm proceeds to Step 2.
  - The goal is to estimate the first clause (c_1, a_1).
  - We compute $(\widetilde{c}_1, \widetilde{a}_1, \widetilde{a}'_1) = \arg\max_{(c_1,a_1,a'_1)\in\mathcal{C}\times\mathcal{A}\times\mathcal{A}} \widehat{R}\left[\{(c_1, a_1), a'_1\}\right]$. This is done, conceptually, by evaluating $\widehat{R}(\cdot)$ at each element in $\mathcal{C}\times\mathcal{A}\times\mathcal{A}$. Suppose the maximum found is $\widehat{R}\left[\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}\right] = 15$ and the clause $\widetilde{c}_1$ has the form $x_1 \leq \tau_1$. Figure A.2 shows the decision list $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$.
  - We compute $\widehat{\Delta}_1 = \widehat{R}\left[\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}\right] - \widehat{R}\left[\{\widetilde{a}_0\}\right]$ and compare $\widehat{\Delta}_1$ to $z_{1-\alpha}\big\{\widehat{\mathrm{Var}}\big(\widehat{\Delta}_1\big)\big\}^{1/2}$. In this case $\widehat{\Delta}_1 = 15 - 10 = 5$. Suppose we get $\widehat{\mathrm{Var}}\big(\widehat{\Delta}_1\big) = 4$ after calculations. Since $5 > z_{0.95} \times 4^{1/2}$, we add two decision lists, $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$ and $\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}$, into the set Π_{temp} and proceed to estimate the second clause (c_2, a_2).
  - We make a remark on non-uniqueness here. The decision list $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$ can be equivalently expressed as $\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}$, where $\widetilde{c}'_1$ is the negation of $\widetilde{c}_1$. Since these two decision lists provide the same treatment recommendation to every patient, we have $\widehat{R}[\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}] = \widehat{R}[\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}] = 15$. However, their first clauses are different and may lead to considerably different final decision lists.



If $x_1 \leq \tau_1$ **then** $\widetilde{a}_1$;
else $\widetilde{a}'_1$.

Figure A.2 Diagram and description of the decision list $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$.



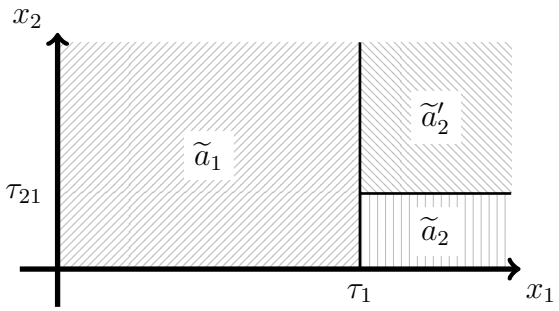
If $x_1 > \tau_1$ **then** $\widetilde{a}'_1$;
else $\widetilde{a}_1$.

Figure A.3 Diagram and description of the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}$.

Currently it is impossible to determine whether $(\widetilde{c}_1, \widetilde{a}_1)$ or $(\widetilde{c}'_1, \widetilde{a}'_1)$ should be used in the first clause. Thus we add both decision lists into Π_{temp}, and move on to building the second clause while keeping in mind that there are two possibilities, $(\widetilde{c}_1, \widetilde{a}_1)$ and $(\widetilde{c}'_1, \widetilde{a}'_1)$, for the first clause. Figure A.3 shows the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}$. The diagram is the same as in Figure A.2 while the description is different.

- The algorithm proceeds to Step 3.

  – We pick an element $\overline{\pi}$ from Π_{temp}. Currently Π_{temp} contains two decision lists: $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$ and $\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}$. Suppose we get $\overline{\pi} = \{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$. We remove



If $x_1 \leq \tau_1$ **then** $\widetilde{a}_1$;
else if $x_2 \leq \tau_{21}$ **then** $\widetilde{a}_2$;
else $\widetilde{a}_2'$.

Figure A.4 Diagram and description of the decision list $\{(\widetilde{c}_1, \widetilde{a}_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}_2'\}$. It is possible that $\widetilde{a}_2 = \widetilde{a}_1$ or $\widetilde{a}_2' = \widetilde{a}_1$.

$\overline{\pi}$ from Π_{temp}.

- We compute $(\widetilde{c}_2, \widetilde{a}_2, \widetilde{a}_2') = \arg\max_{(c_2, a_2, a_2') \in \mathcal{C} \times \mathcal{A} \times \mathcal{A}} \widehat{R}\left[\{(\widetilde{c}_1, \widetilde{a}_1), (c_2, a_2), a_2'\}\right]$. During the maximization $(\widetilde{c}_1, \widetilde{a}_1)$ is held fixed. Intuitively, this is to partition $\mathcal{T}(\widetilde{c}_1)^c$ while keeping $\mathcal{T}(\widetilde{c}_1)$ fixed. Suppose the maximum found is $\widehat{R}[\{(\widetilde{c}_1, \widetilde{a}_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}_2'\}] = 16$ and the clause $\widetilde{c}_2$ has the form $x_2 \leq \tau_{21}$. Figure A.4 shows the decision list $\{(\widetilde{c}_1, \widetilde{a}_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}_2'\}$.
- We compute $\widehat{\Delta}_2 = \widehat{R}\left[\{(\widetilde{c}_1, \widetilde{a}_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}_2'\}\right] - \widehat{R}\left[\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}_1'\}\right]$ and compare $\widehat{\Delta}_2$ to
  $z_{1-\alpha}\left\{\widehat{\text{Var}}\left(\widehat{\Delta}_2\right)\right\}^{1/2}$. In this case $\widehat{\Delta}_2 = 16 - 15 = 1$. Suppose we get $\widehat{\text{Var}}\left(\widehat{\Delta}_2\right) = 2.25$ after calculations. Since $\widehat{\Delta}_2 < z_{0.95}\left\{\widehat{\text{Var}}\left(\widehat{\Delta}_2\right)\right\}^{1/2}$, the simpler, more parsimonious decision list $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}_1'\}$ is preferred and added to Π_{final}, while $\{(\widetilde{c}_1, \widetilde{a}_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}_2'\}$ is discarded.

- The algorithm repeats Step 3.
  - Step 3 is repeated since Π_{temp} contains another element $\overline{\pi} = \{(\widetilde{c}_1', \widetilde{a}_1'), \widetilde{a}_1\}$. We

remove $\overline{\pi}$ from Π_{temp}.

- We compute $(\widetilde{c}_2, \widetilde{a}_2, \widetilde{a}'_2) = \arg\max_{(c_2,a_2,a'_2)\in\mathcal{C}\times\mathcal{A}\times\mathcal{A}} \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (c_2, a_2), a'_2\}\right]$. During the maximization $(\widetilde{c}'_1, \widetilde{a}'_1)$ is held fixed. Intuitively, this is to partition $\mathcal{T}(\widetilde{c}_1)$ while keeping $\mathcal{T}(\widetilde{c}_1)^c$ fixed. Suppose the maximum found is $\widehat{R}[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\})] = 18$ and the clause $\widetilde{c}_2$ has the form $x_2 \leq \tau_{22}$. Figure A.5 shows the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$.
- We compute $\widehat{\Delta}_2 = \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}\right] - \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), \widetilde{a}_1\}\right]$ and compare $\widehat{\Delta}_2$ to
  $z_{1-\alpha}\left\{\widehat{\text{Var}}\left(\widehat{\Delta}_2\right)\right\}^{1/2}$. In this case $\widehat{\Delta}_2 = 18 - 15 = 3$. Suppose we get $\widehat{\text{Var}}\left(\widehat{\Delta}_2\right) = 2$ after calculations. Then we have $\widehat{\Delta}_2 > z_{0.95}\left\{\widehat{\text{Var}}\left(\widehat{\Delta}_2\right)\right\}^{1/2}$, which means that the second clause significantly improves the performance of the decision list. Thus we add decision lists $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$ and $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$ to Π_{temp}.
- Here the non-uniqueness comes into play again. Consequently, although the decision lists $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$ and $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$ are equivalent, it is important to have both of them added to Π_{temp}.

- The algorithm repeats Step 3.

  - Now Π_{temp} contains two decision lists while Π_{final} contains one. Thus Step 3 is repeated. We first pick and remove an element $\overline{\pi}$ from Π_{temp}, say $\overline{\pi} = \{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$.
  - Next, we will build a decision list of length 3 and the first two clauses being $(\widetilde{c}'_1, \widetilde{a}'_1)$ and $(\widetilde{c}_2, \widetilde{a}_2)$. We compute $(\widetilde{c}_3, \widetilde{a}_3, \widetilde{a}'_3) = \arg\max_{(c_3,a_3,a'_3)\in\mathcal{C}\times\mathcal{A}\times\mathcal{A}}$

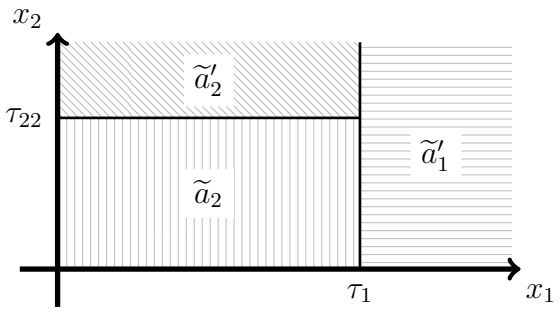


Figure A.5 Diagram and description of the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$. It is possible that $\widetilde{a}_2 = \widetilde{a}'_1$ or $\widetilde{a}'_2 = \widetilde{a}'_1$.

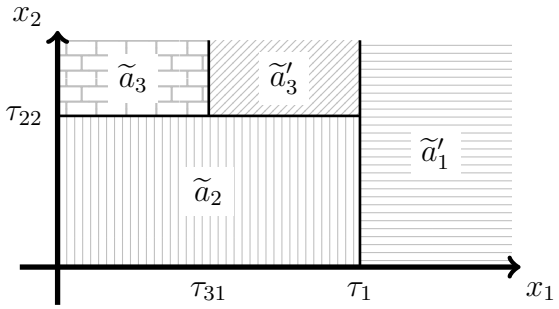
$\widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (c_3, a_3), a'_3\}\right]$. During the maximization $(\widetilde{c}'_1, \widetilde{a}'_1)$ and $(\widetilde{c}_2, \widetilde{a}_2)$ are held fixed. Suppose the maximum found is $\widehat{R}[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}] = 20$ and the clause $\widetilde{c}_3$ has the form $x_1 \leq \tau_{31}$. Figure A.6 shows the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$.

– We then compute $\widehat{\Delta}_3 = \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}\right] - \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}\right]$ and compare $\widehat{\Delta}_3$ to $z_{1-\alpha}\left\{\widehat{\mathrm{Var}}\left(\widehat{\Delta}_3\right)\right\}^{1/2}$. In this case $\widehat{\Delta}_3 = 20 - 18 = 2$. Suppose we get $\widehat{\mathrm{Var}}\left(\widehat{\Delta}_3\right) = 3$ after calculations. Then we have $\widehat{\Delta}_3 < z_{0.95}\left\{\widehat{\mathrm{Var}}\left(\widehat{\Delta}_3\right)\right\}^{1/2}$. Thus the simpler, more parsimonious, decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$ is preferred. So we add $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$ to Π_{final} and drop $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$.

- The algorithm repeats Step 3.

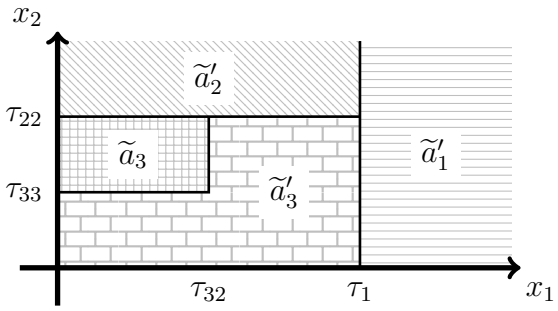
  – Since Π_{temp} contains one element $\overline{\pi} = \{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$, we repeat Step 3 once again. We remove $\overline{\pi}$ from Π_{temp}.

  – We compute $(\widetilde{c}_3, \widetilde{a}_3, \widetilde{a}'_3) = \arg\max_{(c_3, a_3, a'_3) \in \mathcal{C} \times \mathcal{A} \times \mathcal{A}} \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (c_3, a_3), a'_3\}\right]$ while keeping $(\widetilde{c}'_1, \widetilde{a}'_1)$ and $(\widetilde{c}'_2, \widetilde{a}'_2)$ fixed. Suppose $\widehat{R}(\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}) =$



If $x_1 > \tau_1$ **then** $\widetilde{a}'_1$;
else if $x_2 \leq \tau_{22}$ **then** $\widetilde{a}_2$;
else if $x_1 \leq \tau_{31}$ **then** $\widetilde{a}_3$;
else $\widetilde{a}'_3$.

Figure A.6 Diagram and description of the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$. Some of the values of $\widetilde{a}'_1, \widetilde{a}_2, \widetilde{a}_3, \widetilde{a}'_3$ can be equal.



If $x_1 > \tau_1$ **then** $\widetilde{a}'_1$;
else if $x_2 > \tau_{22}$ **then** $\widetilde{a}'_2$;
else if $x_1 \leq \tau_{32}$ and $x_2 > \tau_{33}$ **then** $\widetilde{a}_3$;
else $\widetilde{a}'_3$.

Figure A.7 Diagram and description of the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$. Some of the values of $\widetilde{a}'_1, \widetilde{a}'_2, \widetilde{a}_3, \widetilde{a}'_3$ can be equal.

20.5 and the clause $\widetilde{c}_3$ has the form $x_1 \leq \tau_{32}$ and $x_2 > \tau_{33}$. Figure A.7 shows the decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$.

– We compute $\widehat{\Delta}_3 = \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}\right] - \widehat{R}\left[\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}\right]$ and compare $\widehat{\Delta}_3$ to $z_{1-\alpha}\big\{\widehat{\mathrm{Var}}\big(\widehat{\Delta}_3\big)\big\}^{1/2}$. In this case $\widehat{\Delta}_1 = 20.5 - 18 = 2.5$. Suppose we get $\widehat{\mathrm{Var}}\big(\widehat{\Delta}_1\big) = 2.56$ after calculations. Then we have $\widehat{\Delta}_3 < z_{0.95}\big\{\widehat{\mathrm{Var}}\big(\widehat{\Delta}_3\big)\big\}^{1/2}$. So the simpler, more parsimonious, decision list $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$ is preferred. Consequently, we add $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$ to Π_{final} and discard $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), (\widetilde{c}_3, \widetilde{a}_3), \widetilde{a}'_3\}$.

- The algorithm finishes Step 4, because Π_{temp} contains no element now.
- The algorithm proceeds to Step 5.
  - We would like to pick a decision list from Π_{final} that maximizes $\widehat{R}(\cdot)$.
  - In this example, we have three decision lists in Π_{final}: $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$ with estimated value 15, $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$ with estimated value 18, and $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}'_2, \widetilde{a}'_2), \widetilde{a}_2\}$ with estimated value 18.
  - We then choose the one with the maximal estimated value (with ties broken using the first encountered). Therefore, the estimated optimal decision list $\widetilde{\pi}$ is described by $\{(\widetilde{c}'_1, \widetilde{a}'_1), (\widetilde{c}_2, \widetilde{a}_2), \widetilde{a}'_2\}$, as shown in Figure A.5.

A.2 Asymptotic Properties of $\widehat{R}(\pi)$ for a Given π

We shall derive some asymptotic properties of the doubly robust estimator $\widehat{R}(\pi)$ introduced in Section 2.2 in the main paper. In the next section, we will use these properties to derive an estimator for $\text{Var}\left\{\widehat{R}(\pi_1) - \widehat{R}(\pi_2)\right\}$, which is used by our proposed algorithm for finding an optimal decision list.

Hereafter denote the observed data for the ith subject by $O_i = (X_i^{\mathrm{T}}, A_i, Y_i)^{\mathrm{T}}$.

We first derive an i.i.d. representation of $\widehat{\gamma} = (\widehat{\gamma}_1^{\mathrm{T}}, \ldots, \widehat{\gamma}_{m-1}^{\mathrm{T}})^{\mathrm{T}}$, the maximum likelihood estimator of $\gamma = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_{m-1}^{\mathrm{T}})^{\mathrm{T}}$ in the multinomial logistic regression model:

$$\Pr(A = a | X = x) = \exp\left(u^{\mathrm{T}}\gamma_a\right) \Big/ \left\{1 + \sum_{j=1}^{m-1} \exp\left(u^{\mathrm{T}}\gamma_j\right)\right\}.$$

If $u = u(x) \equiv 1$, then the maximum likelihood estimator of $\omega(x, a) = \Pr(A = a | X = x)$

reduces to $\mathrm{E}_n\, I(A = a)$. Thus the multinomial logistic regression model includes the sample proportion as its special case. The log-likelihood function is

$$\begin{aligned}\ell_{\mathrm{t}}(\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a=1}^{m-1} I(A_i = a)U_i^{\mathrm{T}}\gamma_a - \log\left\{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\gamma_a)\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a=1}^{m-1} I(A_i = a)U_i^{\mathrm{T}}\Phi_a\gamma - \log\left\{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\right\}\right],\end{aligned}$$

where $U_i = u(X_i)$, q is the dimension of U_i, and $\Phi_1 = \big(I_q \mid 0_{q\times(m-2)q}\big), \Phi_2 = \big(0_{q\times q} \mid I_q \mid 0_{q\times(m-3)q}\big), \ldots, \Phi_{m-1} = \big(0_{q\times(m-2)q} \mid I_q\big)$ are $(m-1)$ matrices of size $q \times (m-1)q$ satisfying $\Phi_a\gamma = \gamma_a$. Hence we have

$$\begin{aligned}\frac{\partial\ell_{\mathrm{t}}(\gamma)}{\partial\gamma} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{a=1}^{m-1} I(A_i = a)\Phi_a^{\mathrm{T}}U_i - \frac{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\Phi_a^{\mathrm{T}}U_i}{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)}\right\}, \\ \frac{\partial^2\ell_{\mathrm{t}}(\gamma)}{\partial\gamma\partial\gamma^{\mathrm{T}}} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\Phi_a^{\mathrm{T}}U_iU_i^{\mathrm{T}}\Phi_a}{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{\{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\Phi_a^{\mathrm{T}}U_i\}\{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)U_i^{\mathrm{T}}\Phi_a\}}{\{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\}^2}. \qquad \text{(A.1)}\end{aligned}$$

Denote γ_0 as the maximizer of $\mathrm{E}\,\ell_{\mathrm{t}}(\gamma)$. By the likelihood theory, we have

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = -\sqrt{n}\left[\mathrm{E}\left\{\frac{\partial^2\ell_{\mathrm{t}}(\gamma_0)}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right\}\right]^{-1}\left\{\frac{\partial\ell_{\mathrm{t}}(\gamma_0)}{\partial\gamma}\right\} + o_p(1),$$

where the partial derivatives are given in (A.1), and $o_p(1)$ denotes a random quantity that convergences to zero in probability. Define

$$\varphi_\gamma(O_i) = -\left\{\mathrm{E}\left(\frac{\partial^2\ell_{\mathrm{t}}(\gamma_0)}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right)\right\}^{-1}\left\{\sum_{a=1}^{m-1} I(A_i = a)\Phi_a^{\mathrm{T}}U_i - \frac{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma_0)\Phi_a^{\mathrm{T}}U_i}{1 + \sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma_0)}\right\}.$$

Then we have

$$\sqrt{n}(\widehat{\gamma}-\gamma_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\gamma(O_i)+o_p(1).$$

Next we derive an i.i.d. representation of $\widehat{\beta}=(\widehat{\beta}_1^{\mathrm{T}},\ldots,\widehat{\beta}_m^{\mathrm{T}})^{\mathrm{T}}$, the maximum likelihood estimator of $\beta=(\beta_1^{\mathrm{T}},\ldots,\beta_m^{\mathrm{T}})^{\mathrm{T}}$ in the generalized linear model:

$$g\left\{\mathrm{E}(Y_i|X_i,A_i)\right\}=\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\beta_a.$$

We assume that Y_i given A_i and X_i has an distribution in the exponential family with density function

$$f_{Y_i}(y_i)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi}+c(y_i,\phi)\right\},$$

where θ_i and ϕ are parameters, and $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions. Note that for normal distribution ϕ is known as the dispersion parameter while for Bernoulli distribution ϕ is always equal to one. For simplicity we assume $g(\cdot)$ is a canonical link function hereafter. Then we have $b'(\cdot)\equiv g^{-1}(\cdot)$ and $\theta_i=\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\beta_a$. The log-likelihood function is

$$\begin{aligned}\ell_{\mathrm{o}}(\beta,\phi)&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\beta_a-b\left\{\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\beta_a\right\}}{\phi}+c(Y_i,\phi)\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\Psi_a\beta-b\left\{\sum_{a=1}^{m}I(A_i=a)Z_i^{\mathrm{T}}\Psi_a\beta\right\}}{\phi}+c(Y_i,\phi)\right],\end{aligned}$$

where r is the dimension of Z_i, and $\Psi_1=\left(I_q\mid 0_{q\times(m-1)q}\right)$, $\Psi_2=\left(0_{q\times q}\mid I_q\mid 0_{q\times(m-2)q}\right),\ldots,\Psi_m$

$= \left(0_{q\times(m-1)q} \mid I_q\right)$ are m matrices of size $q \times mq$ satisfying $\Psi_a\beta = \beta_a$. Then we have

$$\frac{\partial \ell_{\mathrm{o}}(\beta,\phi)}{\partial \beta} = \frac{1}{n\phi}\sum_{i=1}^{n}\left[Y_i - b'\left\{\sum_{a=1}^{m} I(A_i=a) Z_i^{\mathrm{T}}\Psi_a\beta\right\}\right]\left\{\sum_{a=1}^{m} I(A_i=a)\Psi_a^{\mathrm{T}} Z_i\right\},$$
$$\frac{\partial^2 \ell_{\mathrm{o}}(\beta,\phi)}{\partial \beta \partial \beta^{\mathrm{T}}} = -\frac{1}{n\phi}\sum_{i=1}^{n} b''\left\{\sum_{a=1}^{m} I(A_i=a) Z_i^{\mathrm{T}}\Psi_a\beta\right\}\left\{\sum_{a=1}^{m} I(A_i=a)\Psi_a^{\mathrm{T}} Z_i Z_i^{\mathrm{T}}\Psi_a\right\}.$$

By the property of the score function, we have

$$\mathrm{E}\left(\frac{\partial^2 \ell_{\mathrm{o}}(\beta_0,\phi_0)}{\partial\beta\partial\phi}\right) = -\frac{1}{\phi}\,\mathrm{E}\left(\frac{\partial \ell_{\mathrm{o}}(\beta_0,\phi_0)}{\partial\beta}\right) = 0.$$

Therefore, by the likelihood theory and the property of block diagonal matrix, we conclude that

$$\begin{aligned}\sqrt{n}(\widehat{\beta}-\beta_0) &= -\sqrt{n}\left[\mathrm{E}\left\{\frac{\partial^2 \ell_{\mathrm{o}}(\beta_0,\phi_0)}{\partial\beta\partial\beta^{\mathrm{T}}}\right\}\right]^{-1}\left\{\frac{\partial \ell_{\mathrm{o}}(\beta_0,\phi_0)}{\partial\beta}\right\} + o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\beta(O_i) + o_p(1),\end{aligned}$$

where

$$\begin{aligned}\varphi_\beta(O_i) = \left(\mathrm{E}\left[b''\left(\sum_{a=1}^{m} I(A_i=a) Z_i^{\mathrm{T}}\Psi_a\beta_0\right)\left\{\sum_{a=1}^{m} I(A_i=a)\Psi_a^{\mathrm{T}} Z_i Z_i^{\mathrm{T}}\Psi_a\right\}\right]\right)^{-1}\\ \cdot\left\{Y_i - b'\left(\sum_{a=1}^{m} I(A_i=a) Z_i^{\mathrm{T}}\Psi_a\beta_0\right)\right\}\left\{\sum_{a=1}^{m} I(A_i=a)\Psi_a^{\mathrm{T}} Z_i\right\}.\end{aligned}$$

Finally we derive an i.i.d. representation of $\widehat{R}(\pi)$. To emphasize the dependence of $\omega(x,a)$ and $\mu(x,a)$ on the parameters γ and β, in the following we write $\omega(x,a)$ as $\omega(x,a,\gamma)$ and $\mu(x,a)$ as $\mu(x,a,\beta)$. Thus we have $\widehat{\omega}(x,a) = \omega(x,a,\widehat{\gamma})$ and $\widehat{\mu}(x,a) =$

$\mu(x, a, \widehat{\beta})$. Note that

$$\omega(x, a, \gamma) = \frac{\exp(u^{\mathrm{T}}\Phi_a\gamma)}{\sum_{j=1}^m \exp(u^{\mathrm{T}}\Phi_j\gamma)},$$
$$\mu(x, a, \beta) = b'(z^{\mathrm{T}}\Psi_a\beta),$$

for $a = 1, \ldots, m$, where $\Phi_m = 0_{q\times(m-1)q}$. Hence we have

$$\frac{\partial\omega(x, a, \gamma)}{\partial\gamma} = \frac{\exp(u^{\mathrm{T}}\Phi_a\gamma)\left\{\sum_{j=1}^m \exp(u^{\mathrm{T}}\Phi_j\gamma)\cdot(\Phi_a^{\mathrm{T}} - \Phi_j^{\mathrm{T}})u\right\}}{\left\{\sum_{j=1}^m \exp(u^{\mathrm{T}}\Phi_j\gamma)\right\}^2},$$
$$\frac{\partial\mu(x, a, \beta)}{\partial\beta} = b''(z^{\mathrm{T}}\Psi_a\beta)\Psi_a^{\mathrm{T}}z. \tag{A.2}$$

By Taylor expansion, we have

$$\begin{aligned}
\widehat{R}(\pi) &= \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^m\left[\frac{I(A_i = a)}{\omega(X_i, a, \widehat{\gamma})}\left\{Y_i - \mu(X_i, a, \widehat{\beta})\right\} + \mu(X_i, a, \widehat{\beta})\right] I\{\pi(X_i) = a\} \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^m\left[\frac{I(A_i = a)}{\omega(X_i, a, \gamma_0)}\{Y_i - \mu(X_i, a, \beta_0)\} + \mu(X_i, a, \beta_0)\right] I\{\pi(X_i) = a\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^m\left[-\frac{I(A_i = a)}{\omega^2(X_i, a, \gamma_0)}\{Y_i - \mu(X_i, a, \beta_0)\} I\{\pi(X_i) = a\}\frac{\partial\omega(X_i, a, \gamma_0)}{\partial\gamma}\right]^{\mathrm{T}}(\widehat{\gamma} - \gamma_0) \\
&\quad + \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^m\left[\left\{-\frac{I(A_i = a)}{\omega(X_i, a, \gamma_0)} + 1\right\} I\{\pi(X_i) = a\}\frac{\partial\mu(X_i, a, \beta_0)}{\partial\beta}\right]^{\mathrm{T}}(\widehat{\beta} - \beta_0) + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^m\left[\frac{I(A_i = a)}{\omega(X_i, a, \gamma_0)}\{Y_i - \mu(X_i, a, \beta_0)\} + \mu(X_i, a, \beta_0)\right] I\{\pi(X_i) = a\} \\
&\quad + \mathrm{E}\left(\sum_{a=1}^m\left[-\frac{I(A_i = a)}{\omega^2(X_i, a, \gamma_0)}\{Y_i - \mu(X_i, a, \beta_0)\} I\{\pi(X_i) = a\}\frac{\partial\omega(X_i, a, \gamma_0)}{\partial\gamma}\right]\right)^{\mathrm{T}}(\widehat{\gamma} - \gamma_0)
\end{aligned}$$

$$+ \mathrm{E}\left(\sum_{a=1}^{m}\left[\left\{-\frac{I(A_i=a)}{\omega(X_i,a,\gamma_0)}+1\right\}I\{\pi(X_i)=a\}\frac{\partial\mu(X_i,a,\beta_0)}{\partial\beta}\right]\right)^{\mathrm{T}}(\widehat{\beta}-\beta_0)+o_p(1).$$

Recall that

$$R(\pi)=\mathrm{E}\left(\sum_{a=1}^{m}\left[\frac{I(A_i=a)}{\omega(X_i,a,\gamma_0)}\{Y_i-\mu(X_i,a,\beta_0)\}+\mu(X_i,a,\beta_0)\right]I\{\pi(X_i)=a\}\right).$$

Define

$$\begin{aligned}\varphi_R(O_i)=&\sum_{a=1}^{m}\left[\frac{I(A_i=a)}{\omega(X_i,a,\gamma_0)}\{Y_i-\mu(X_i,a,\beta_0)\}+\mu(X_i,a,\beta_0)\right]I\{\pi(X_i)=a\}\\&-\mathrm{E}\left(\sum_{a=1}^{m}\left[\frac{I(A_i=a)}{\omega(X_i,a,\gamma_0)}\{Y_i-\mu(X_i,a,\beta_0)\}+\mu(X_i,a,\beta_0)\right]I\{\pi(X_i)=a\}\right)\\&+\mathrm{E}\left(\sum_{a=1}^{m}\left[-\frac{I(A_i=a)}{\omega^2(X_i,a,\gamma_0)}\{Y_i-\mu(X_i,a,\beta_0)\}I\{\pi(X_i)=a\}\frac{\partial\omega(X_i,a,\gamma_0)}{\partial\gamma}\right]\right)^{\mathrm{T}}\varphi_\gamma(O_i)\\&+\mathrm{E}\left(\sum_{a=1}^{m}\left[\left\{-\frac{I(A_i=a)}{\omega(X_i,a,\gamma_0)}+1\right\}I\{\pi(X_i)=a\}\frac{\partial\mu(X_i,a,\beta_0)}{\partial\beta}\right]\right)^{\mathrm{T}}\varphi_\beta(O_i),\end{aligned}\tag{A.3}$$

where $\partial\omega/\partial\gamma$ and $\partial\mu/\partial\beta$ are given in (A.2). Then we have

$$\sqrt{n}\left\{\widehat{R}(\pi)-R(\pi)\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_R(O_i)+o_p(1).$$

Therefore, by the central limit theorem and the Slutsky's theorem, we conclude that

$$\sqrt{n}\left\{\widehat{R}(\pi)-R(\pi)\right\}\xrightarrow{d}N(0,\mathrm{E}\left\{\varphi_R^2(O_i)\right\}),\tag{A.4}$$

where $\xrightarrow{d}$ denotes convergence in distribution.

To estimate the asymptotic variance, we use the plug-in method. Namely, define $\widehat{\varphi}_R(O_i)$ as in (A.3) except that expectations are replaced with sample averages and true values are replaced with corresponding estimates. Then $\mathrm{Var}\left(\widehat{R}(\pi)\right)$ can be estimated by $\sum_{i=1}^n \widehat{\varphi}_R^2(O_i)/n^2$.

A.3 Asymptotic Properties of $\widehat{R}(\pi_1) - \widehat{R}(\pi_2)$

Define $\varphi_{R1}(O_i)$ as in (A.3) with π replaced by π_1. Define $\varphi_{R2}(O_i)$ as in (A.3) with π replaced by π_2. Define $\widehat{\varphi}_{R1}(O_i)$ and $\widehat{\varphi}_{R2}(O_i)$ similarly. Then we have

$$\begin{aligned}\sqrt{n}\left[\left\{\widehat{R}(\pi_1) - \widehat{R}(\pi_2)\right\} - \{R(\pi_1) - R(\pi_2)\}\right] &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\varphi_{R1}(O_i) - \varphi_{R2}(O_i)\} + o_p(1) \\ &\xrightarrow{d} N(0, \mathrm{E}\{\varphi_{R1}(O_i) - \varphi_{R2}(O_i)\}^2).\end{aligned}$$

Therefore, we can estimate $\mathrm{Var}\left\{\widehat{R}(\pi_1) - \widehat{R}(\pi_2)\right\}$ by

$$\widehat{\mathrm{Var}}\left\{\widehat{R}(\pi_1) - \widehat{R}(\pi_2)\right\} = \frac{1}{n^2}\sum_{i=1}^n \{\widehat{\varphi}_{R1}(O_i) - \widehat{\varphi}_{R2}(O_i)\}^2. \tag{A.5}$$

The variance estimator $\widehat{\mathrm{Var}}(\Delta_j)$ used in the algorithm in Section 2.4.1 in the main paper can be obtained via (A.5) with $\pi_1 = \left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}_j, \widetilde{a}_j), \widetilde{a}'_j\right\}$ and $\pi_2 = \left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), \overline{a}'_{j-1}\right\}$.

A.4 Implementation Details of Finding an Optimal Decision List

A.4.1 Algorithm Description

We give an equivalent version of the proposed algorithm for finding an optimal decision list. Compared to the algorithm presented in the main paper, this version makes use of recursive calls to avoid explicit constructions of sets Π_{temp} and Π_{final}, and facilitates the analysis of time complexity. The algorithm is as follows.

Input: $\widehat{R}(\cdot)$, $L_{\max}$, α
Output: a decision list $\widetilde{\pi}$ that maximize $\widehat{R}(\cdot)$
$\widetilde{a}_0 = \arg\max_{a_0 \in \mathcal{A}} \widehat{R}\left[\{a_0\}\right]$;
$\widetilde{\pi} = \texttt{FindList}(1, \{\}, \widetilde{a}_0)$;

The function `FindList` is defined below. When $j = 1$, we treat $(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1})$ as an empty array. Thus when $j = 1$, $\left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}_j, \widetilde{a}_j), \widetilde{a}'_j\right\}$ is the same as $\{(\widetilde{c}_1, \widetilde{a}_1), \widetilde{a}'_1\}$ and $\left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), \overline{a}'_{j-1}\right\}$ is the same as $\{\overline{a}'_0\}$.

In the `FindList` function, a crucial step is to compute $(\widetilde{c}_j, \widetilde{a}_j, \widetilde{a}'_j)$. A straightforward implementation that involves a brute-force search over $\mathcal{C} \times \mathcal{A} \times \mathcal{A}$ can be time consuming. We provide an efficient implementation below.

We observe that some calculations can be performed only once at the beginning of

Function `FindList` $(j,\ \{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1})\},\ \overline{a}'_{j-1})$

$(\widetilde{c}_j, \widetilde{a}_j, \widetilde{a}'_j) = \arg\max_{(c_j, a_j, a'_j) \in \mathcal{C} \times \mathcal{A} \times \mathcal{A}} \widehat{R}\left[\left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (c_j, a_j), a'_j\right\}\right]$;

$\widehat{\Delta}_j =$

$\widehat{R}\left[\left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}_j, \widetilde{a}_j), \widetilde{a}'_j\right\}\right] - \widehat{R}\left[\left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), \overline{a}'_{j-1}\right\}\right]$;

if $\widehat{\Delta}_j < z_{1-\alpha}\left\{\widehat{\mathrm{Var}}\left(\widehat{\Delta}_j\right)\right\}^{1/2}$ **then**

  $\widetilde{\pi} = \left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), \overline{a}'_{j-1}\right\}$;

else if $j = L_{\max}$ **then**

  $\widetilde{\pi} = \left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}_j, \widetilde{a}_j), \widetilde{a}'_j\right\}$;

else

  $\widetilde{\pi}_1 =$ `FindList`$(j+1,\ \{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}_j, \widetilde{a}_j)\},\ \widetilde{a}'_j)$;

  $\widetilde{\pi}_2 =$ `FindList`$(j+1,\ \left\{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_{j-1}, \overline{a}_{j-1}), (\widetilde{c}'_j, \widetilde{a}'_j)\right\},\ \widetilde{a}_j)$,

  where $\widetilde{c}'_j =$ negation of $\widetilde{c}_j$;

  $\widetilde{\pi} = \arg\max_{\pi \in \{\widetilde{\pi}_1, \widetilde{\pi}_2\}} \widehat{R}(\pi)$;

end

return $\widetilde{\pi}$;

end

the algorithm. First, define

$$\widehat{\xi}_{ia} = \frac{I(A_i = a)}{\omega(X_i, a, \widehat{\gamma})} \left\{Y_i - \mu(X_i, a, \widehat{\beta})\right\} + \mu(X_i, a, \widehat{\beta}).$$

Then we have

$$\widehat{R}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{\pi(X_i) = a\}.$$

Second, for the ith subject, denote x_{ij} as the observed value of his/her jth covariate. For the jth baseline covariate, there are $s_k = \#\mathcal{X}_j$ possible candidate cutoff values $\tau_{j1} \leq \cdots \leq \tau_{js_j}$, which divides the real line into $s_k + 1$ intervals:

$$(-\infty, \tau_{j1}],\ (\tau_{j1}, \tau_{j2}],\ \ldots,\ (\tau_{j(s_j-1)}, \tau_{js_j}],\ (\tau_{js_j}, \infty).$$

Then we code the observed values $x_{1j}, \ldots, x_{nj}$ into indices $b_{1j}, \ldots, b_{nj}$ according to which interval they fall.

In order to reduce the number of evaluations of $\widehat{R}(\cdot)$ when searching for the maximizer over $\mathcal{C} \times \mathcal{A} \times \mathcal{A}$, we organize the intermediate results as shown below. Let $\mathcal{I} = \{i : X_i \in \mathcal{T}(\overline{c}_\ell)^c \text{ for all } \ell < j\}$. Then $\mathcal{I}$ contains all the subjects that have not had treatment recommendations up to the jth clause. Since we have

$$n\widehat{R}(\pi) = \sum_{i \in \mathcal{I}} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{\pi(X_i) = a\} + \sum_{i \in \mathcal{I}^c} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{\pi(X_i) = a\}$$

and $\sum_{i \in \mathcal{I}^c} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{\pi(X_i) = a\}$ is constant during the maximization, we focus on maximizing $\sum_{i \in \mathcal{I}} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{\pi(X_i) = a\}$, which reduces to maximizing

$$\sum_{i \in \mathcal{I}} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{i \in \mathcal{T}(c_j), a = a_j\} + \sum_{i \in \mathcal{I}} \sum_{a=1}^{m} \widehat{\xi}_{ia} I\{i \notin \mathcal{T}(c_j), a = a'_j\}. \tag{A.6}$$

To identify the maximizer of (A.6), we first loop over all possible pairs of covariates. For each pair of covariates, say the kth and the ℓth covariates, define D, a three-dimensional array of size $m \times (s_k + 1) \times (s_\ell + 1)$, as $D_{auv} = \sum_{i \in \mathcal{I}} \widehat{\xi}_{ia} I(b_{ik} = u, b_{il} = v)$. Next, we loop over all possible cutoff values and construct the corresponding c_j. The values of a_j and a'_j that maximizes (A.6) for a given c_j can be easily obtained due to the additive structure. After enumerating all the possible conditions that c_j may take, we can find out $(\widetilde{c}_j, \widetilde{a}_j, \widetilde{a}'_j)$.

A.4.2 Time Complexity Analysis

Since computing $\widehat{\xi}_{ia}$s requires $O(nm)$ time and computing b_{ij}s requires $O(np)$ time. The calculations at the beginning of the algorithm take $O(nm + np)$ time in total.

The algorithm first computes $\widetilde{a}_0$, which requires $O(nm)$ time. Then it invokes a function call `FindList`$(1, \{\}, \widetilde{a}_0)$. Due to the recursive nature of the `FindList` function, we will compute the time complexity by establishing a recurrence relation between $T(j)$ and $T(j+1)$, where $T(j)$ is the time complexity of the function call `FindList` $(j, \{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_j, \overline{a}_j)\})$.

Suppose a call `FindList` $(j, \{(\overline{c}_1, \overline{a}_1), \ldots, (\overline{c}_j, \overline{a}_j)\})$ is invoked. The running time can be computed by going through the algorithm of the `FindList` function step-by-step as follows.

First, the function computes $(\widetilde{c}_j, \widetilde{a}_j, \widetilde{a}'_j)$. A naive implementation would involve looping over all the covariates, all the possible cutoff values and all the treatment options, whose running time is $O(nmp^2s^2)$, where $s = \max_j s_j$. However, the running time is greatly reduced if we use the efficient implementation described previously. For a given pair of covariates, we can compute D_{auv}s in $O(nm)$ time. Then we can find out the maximum of (A.6) in $O(ms^2)$ time by looping over all possible cutoff values. Therefore, the total time for computing $(\widetilde{c}_j, \widetilde{a}_j, \widetilde{a}'_j)$ is $O\{(n + s^2)mp^2\}$.

Second, the function computes $\widehat{\Delta}_j$, which takes $O(n)$ time.

Third, the function computes $\widehat{\mathrm{Var}}\big(\widehat{\Delta}_j\big)$, whose running time is $O(nmq + nmr)$, where q is the dimension of U_i and r is the dimension of Z_i. Since both U_i and Z_i are known feature vectors constructed from X_i, for most cases q and r are of the same order as p. So this step takes $O(nmp)$ time.

Fourth, the function executes the "if-then" statement. In the worst case, the function makes two recursive calls, taking $2T(j+1)$ time.

Combining these four steps, we have $T(j) = O\{(n+s^2)mp^2\}+2T(j+1)$. The boundary condition is $T(L_{\max}) = O\{(n+s^2)mp^2\}$. Using backward induction, we get $T(0) = O\{2^{L_{\max}}(n+s^2)mp^2\}$. Recall that $s = \max_j \#\mathcal{X}_j$.

Combining $T(0)$ with the running time before invoking `FindList`$(1, \{\}, \widetilde{a}_0)$, we obtain that the time complexity of the entire algorithm is $O[2^{L_{\max}}mp^2\{n + (\max_j \#\mathcal{X}_j)^2\}]$.

A.5 Implementation Details of Finding an Equivalent Decision List with Minimal Cost

In this section we give an algorithmic description of the proposed method for finding an equivalent decision list with minimal cost. Recall that two decision lists are called equivalent if they give the same treatment recommendation for every patient in the population.

Input: a decision list $\bar{\pi}$
Output: an equivalent decision list $\pi_{\min}$ with minimal cost $N_{\min}$
Identify atoms in $\bar{\pi}$ as $d_1, \ldots, d_K$;
Compute $\mathcal{I}_a = \{i : \bar{\pi}(X_i) = a\}$ for each $a \in \mathcal{A}$;
Set $\pi_{\min} = \{\}$ and $N_{\min} = \infty$;
`FindMinCost` (0, {}, $\pi_{\min}$, $N_{\min}$);

The function `FindMinCost` is defined below.

Function `FindMinCost` *(j, $\{(c_1, a_1), \ldots, (c_j, a_j)\}$, π_{min}, N_{min})*

Compute a lower bound of the cost as $N_{\text{bd}} = \mathcal{N}_\ell \sum_{\ell=1}^{j} \Pr_n(X \in \mathcal{R}_\ell) + \mathcal{N}_j \Pr_n(X \in \cap_{\ell=1}^{j} \mathcal{R}_\ell^c)$, where $\Pr_n$ denotes the empirical probability measure;

if $N_{\text{bd}} \geq N_{\min}$ **then return**;

$\mathcal{I} = \{i : X_i \in \mathcal{T}(c_\ell)^c \text{ for all } \ell \leq j\}$;

if $\mathcal{I} \subset \mathcal{I}_{a_0}$ *for some* a_0 **then**

  if $N[\{(c_1, a_1), \ldots, (c_j, a_j), a_0\}] < N_{\min}$ **then**

    $\pi_{\min} = \{(c_1, a_1), \ldots, (c_j, a_j), a_0\}$;

    $N_{\min} = N(\pi_{\min})$;

  end

else

  for $1 \leq k_1 < k_2 \leq K$ **do**

    Let $\mathcal{C}_{k_1,k_2}$ be the set consisting of all the logical clauses involving d_{k_1} or d_{k_2} or both using conjunction, disjunction, and/or negation;

    for $c_{j+1} \in \mathcal{C}_{k_1,k_2}$ **do**

      $\mathcal{J}_{j+1} = \{i \in \mathcal{I} : X_i \in \mathcal{T}(c_{j+1})\}$;

      if $\mathcal{J}_{j+1}$ *is non-empty and* $\mathcal{J}_{j+1} \subset \mathcal{I}_{a_{j+1}}$ *for some* $a_{j+1} \in \mathcal{A}$ **then**

        `FindMinCost` $(j+1, \{(c_1, a_1), \ldots, (c_j, a_j), (c_{j+1}, a_{j+1})\}, \pi_{\min}, N_{\min})$;

      end

    end

end

end

A.6 Point Estimate and Prediction Interval for $R(\widehat{\pi})$ with Bootstrap Bias Correction

In this section, we show how to estimate the value of the estimated treatment regime, $R(\widehat{\pi})$, and how to construct a prediction interval for it.

A.6.1 Methodology

To measure how well the estimated treatment regime $\widehat{\pi}$ performs, it is often of interest to construct an estimator of and a prediction interval for $R(\widehat{\pi})$. Since a natural candidate for estimating $R(\widehat{\pi})$ is $\widehat{R}(\widehat{\pi})$, it may be tempting to construct a prediction interval centering at $\widehat{R}(\widehat{\pi})$. However, $\widehat{R}(\widehat{\pi})$ is generally too optimistic to serve as an honest estimator of $R(\widehat{\pi})$. It has an upward bias due to the maximization process. As a remedy, we suggest using B bootstraps to correct this bias. Specifically, the perturbed version of $\widehat{R}(\pi)$ in the bth bootstrapping sample is

$$\widehat{R}_b^*(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(W_i\sum_{a=1}^{m}\left[\frac{I(A_i=a)}{\omega(X_i,a,\widehat{\gamma}^*)}\left\{Y_i-\mu(X_i,a,\widehat{\beta}^*)\right\}+\mu(X_i,a,\widehat{\beta}^*)\right]I\{\pi(X_i)=a\}\right),$$

where $W_1,\ldots,W_n$ are identically and independently distributed with standard exponential distribution, $\widehat{\gamma}^*$ is the solution to

$$\sum_{i=1}^{n}W_i\left\{\sum_{a=1}^{m-1}I(A_i=a)\Phi_a^{\mathrm{T}}U_i-\frac{\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)\Phi_a^{\mathrm{T}}U_i}{1+\sum_{a=1}^{m-1}\exp(U_i^{\mathrm{T}}\Phi_a\gamma)}\right\}=0,$$

and $\widehat{\beta}^*$ is the solution to

$$\sum_{i=1}^{n} W_i \left[Y_i - b' \left\{ \sum_{a=1}^{m} I(A_i = a) Z_i^{\mathrm{T}} \Psi_a \beta \right\} \right] \left\{ \sum_{a=1}^{m} I(A_i = a) \Psi_a^{\mathrm{T}} Z_i \right\} = 0.$$

Let $\widehat{\pi}_b^*$ be the maximizer of $\widehat{R}_b^*(\pi)$ over Π. Then the actual bias $\widehat{R}(\widehat{\pi}) - R(\widehat{\pi})$ can be estimated by the corresponding bias in the bootstrap world: $\widehat{\mathrm{Bias}} = \sum_{b=1}^{B} \{\widehat{R}_b^*(\widehat{\pi}_b^*) - \widehat{R}(\widehat{\pi}_b^*)\}/B$, where B is the number of bootstrap samples. The final estimator of $R(\widehat{\pi})$ is $\widehat{R}_{\mathrm{c}}(\widehat{\pi}) = \widehat{R}(\widehat{\pi}) - \widehat{\mathrm{Bias}}$.

To construct a prediction interval for $R(\widehat{\pi})$, we treat $\widehat{\pi}$ as a non-random quantity, and then utilize the asymptotic normality of $\widehat{R}(\widehat{\pi})$ given in (A.4). Let z_ρ be the 100ρ percentile of a standard normal distribution and $\widehat{\sigma}^2 = \widehat{\mathrm{Var}}\big\{\widehat{R}(\widehat{\pi})\big\}$. Then a $(1-\alpha)\times 100\%$ prediction interval for $R(\widehat{\pi})$ is

$$\big[\widehat{R}_{\mathrm{c}}(\widehat{\pi}) + z_{\alpha/2}\widehat{\sigma},\ \widehat{R}_{\mathrm{c}}(\widehat{\pi}) + z_{1-\alpha/2}\widehat{\sigma}\big]. \tag{A.7}$$

A potential drawback of this interval is, though, that it ignores the variation introduced by $\widehat{\pi}$. Nevertheless, our numerical experiences suggest that this extra variation is generally small and the coverage probability is close to the nominal level. Taking into account the variability of $\widehat{\pi}$ has to deal with the associated non-regularity issue, which is beyond the scope of this paper.

As a final remark, for binary outcome we suggest to conduct the bias correction and construct the prediction interval based on $\mathrm{logit}\{\widehat{R}(\cdot)\}$ first and then transform back to the original scale, where $\mathrm{logit}(v) = \log\{v/(1-v)\}$.

Table A.1 Point estimate and coverage probabilities of prediction intervals with and without bootstrap bias correction. Plain-PI refers to the coverage probability of the plain prediction interval, and Corrected-PI refers to the coverage probability of the bias-corrected prediction interval.

p	Setting	Continuous response				Binary response			
		$R(\widehat{\pi})$	$\widehat{R}_{\mathrm{c}}(\widehat{\pi})$	Plain-PI	Corrected-PI	$R(\widehat{\pi})$	$\widehat{R}_{\mathrm{c}}(\widehat{\pi})$	Plain-PI	Corrected-PI
10	I	2.78	2.78	0.95	0.94	0.77	0.76	0.97	0.96
	II	2.70	2.73	0.93	0.95	0.71	0.72	0.89	0.97
	III	2.59	2.61	0.95	0.95	0.73	0.74	0.88	0.96
	IV	2.89	2.98	0.88	0.94	0.71	0.72	0.63	0.98
	V	2.90	2.90	0.95	0.95	0.75	0.75	0.76	0.96
	VI	3.98	4.01	0.93	0.95	0.79	0.79	0.97	0.99
	VII	3.22	3.27	0.86	0.94	0.77	0.77	0.77	1.00
50	I	2.76	2.75	0.94	0.94	0.76	0.76	0.82	0.98
	II	2.70	2.72	0.93	0.94	0.71	0.71	0.80	0.96
	III	2.59	2.59	0.94	0.95	0.73	0.73	0.63	0.98
	IV	2.89	2.96	0.88	0.94	0.71	0.72	0.48	0.98
	V	2.87	2.87	0.93	0.94	0.74	0.74	0.33	0.96
	VI	3.95	3.99	0.91	0.94	0.78	0.79	0.89	0.99
	VII	3.21	3.27	0.88	0.94	0.76	0.77	0.63	0.99

A.6.2 Simulations

We present the point estimate and the coverage probabilities of the plain prediction interval and the prediction interval with bootstrap bias correction in Table A.1. The setting used here is exactly the same as that in Section 3 in the main paper. We can see that the bias correction improves the coverage probability substantially in finite samples, especially as the number of covariates gets larger. Besides, the bootstrap prediction interval is prone to overcoverage for the binary response.

A.7 Accuracy of Variable Selection

Consider the simulated experiments in the main paper. To quantify variable selection accuracy, we compute the true positive rate, the number of signal variables included in the decision list divided by the number of signal variables, and the false positive rate, the number of noise variables included in the decision list divided by the number of noise variables. A variable is called a signal variable if it appears in $\phi(x, a)$ and is a noise variable otherwise, irrespective of the actual functional form.

Table A.2 presents the true positive rates and the false positive rates under different settings. The proposed method consistently identifies signal variables and screens out noise variables in most settings. The only exception is setting IV, where the optimal regime is far away from being well approximated by decision lists. Thus the proposed approach loses power in detecting useful covariates due to misspecifying the form of the regime.

A.8 Impact of the Tuning Parameter in the Stopping Criterion

In the algorithm discussed in Section 2.4.1 in the main paper, we use a tuning parameter α to control the building process of the decision list and we suggest to fix α at 0.95. In the following we show that the final decision list is insensitive to the choice of α via simulation study. The setting used here is exactly the same as that in Section 3 in the main paper. We varied α among $\{0.9, 0.95, 0.99\}$.

Table A.3 shows the impact of α on the value and the cost of the estimated regime.

Table A.2 Accuracy of variable selection using decision list. TPR is the true positive rate and FPR is the false positive rate.

p	Setting	Continuous response		Binary response	
		TPR	FPR	TPR	FPR
10	I	1.00	0.00	1.00	0.07
	II	1.00	0.00	1.00	0.04
	III	1.00	0.00	1.00	0.11
	IV	0.93	0.00	0.79	0.07
	V	1.00	0.07	1.00	0.20
	VI	1.00	0.05	1.00	0.10
	VII	0.94	0.00	0.98	0.04
50	I	1.00	0.01	1.00	0.04
	II	1.00	0.00	1.00	0.02
	III	1.00	0.00	0.99	0.04
	IV	0.93	0.00	0.73	0.02
	V	1.00	0.02	0.99	0.06
	VI	1.00	0.02	1.00	0.03
	VII	0.94	0.00	0.97	0.02

We can see that the value and the cost as well as the accuracy of variable selection, averaged over 1000 replications, are very stable across different choices of α. Table A.4 shows the impact of α on the estimated regime. It is clear that α has little impact on the treatment recommendation made by the estimated regime.

A.9 Chronic Depression Data

In the application considered in Section 4.2 in the main paper, we applied the proposed method to construct an interpretable and parsimonious treatment regime. We follow Gunter et al. (2011) and Zhao et al. (2012), and use the following 50 covariates:

1. Gender: 1 if female, 0 if male;
2. Race: 1 if white, 0 otherwise;
3. Marital status I: 1 if single, 0 otherwise;
4. Marital status II: 1 if married or living with someone, 0 otherwise;
5. Body mass index: continuous;
6. Age at screening: continuous;
7. Having difficulty in planning family activity: 1 if strongly agree, 2 if agree, 3 if disagree, 4 if strongly disagree;
8. Supporting each other in the family: 1 if strongly agree, 2 if agree, 3 if disagree, 4 if strongly disagree;

Table A.3 The impact of α on the value and the cost of the estimated regime. In the header, α is the tuning parameter in the stopping criterion; $R(\widehat{\pi})$ is the mean outcome under the estimated regime $\widehat{\pi}$, computed on a test set of 10^6 subjects; $N(\widehat{\pi})$ is the cost of implementing the estimated regime $\widehat{\pi}$, computed on the same test set; TPR is the true positive rate, namely, the number of signal variables involved in $\widehat{\pi}$ divided by the number of signal variables; FPR is the false positive rate, namely, the number of noise variables involved in $\widehat{\pi}$ divided by the number of noise variables. Recall that p is the dimension of patient covariates.

p	Setting	$\alpha = 0.9$				$\alpha = 0.95$				$\alpha = 0.99$			
		$R(\widehat{\pi})$	$N(\widehat{\pi})$	TPR	FPR	$R(\widehat{\pi})$	$N(\widehat{\pi})$	TPR	FPR	$R(\widehat{\pi})$	$N(\widehat{\pi})$	TPR	FPR
Continuous response													
10	I	2.78	1.65	1.00	0.01	2.78	1.65	1.00	0.00	2.78	1.65	1.00	0.00
	II	2.71	1.66	1.00	0.00	2.70	1.66	1.00	0.00	2.69	1.66	1.00	0.00
	III	2.59	1.69	1.00	0.00	2.59	1.69	1.00	0.00	2.59	1.69	1.00	0.00
	IV	2.89	2.51	0.93	0.00	2.89	2.51	0.93	0.00	2.89	2.51	0.92	0.00
	V	2.90	1.91	1.00	0.07	2.90	1.91	1.00	0.07	2.90	1.91	1.00	0.07
	VI	3.98	1.61	1.00	0.06	3.98	1.61	1.00	0.05	3.98	1.61	1.00	0.05
	VII	3.22	2.56	0.94	0.00	3.22	2.56	0.94	0.00	3.21	2.56	0.94	0.00
50	I	2.75	1.96	1.00	0.01	2.76	1.96	1.00	0.01	2.78	1.96	1.00	0.00
	II	2.70	1.66	1.00	0.00	2.70	1.66	1.00	0.00	2.69	1.66	1.00	0.00
	III	2.58	1.75	1.00	0.00	2.59	1.75	1.00	0.00	2.59	1.75	1.00	0.00
	IV	2.89	2.54	0.93	0.00	2.89	2.54	0.93	0.00	2.89	2.54	0.92	0.00
	V	2.87	2.19	1.00	0.03	2.87	2.19	1.00	0.02	2.88	2.19	1.00	0.02
	VI	3.95	1.70	1.00	0.02	3.95	1.70	1.00	0.02	3.95	1.70	1.00	0.02
	VII	3.22	2.56	0.94	0.00	3.21	2.56	0.94	0.00	3.21	2.56	0.93	0.00
Binary response													
10	I	0.76	2.16	1.00	0.12	0.77	2.16	1.00	0.07	0.77	2.16	1.00	0.02
	II	0.71	1.75	1.00	0.06	0.71	1.75	1.00	0.04	0.71	1.75	1.00	0.02
	III	0.73	2.24	1.00	0.15	0.73	2.24	1.00	0.11	0.74	2.24	0.99	0.04
	IV	0.71	2.48	0.81	0.08	0.71	2.48	0.79	0.07	0.71	2.48	0.70	0.06
	V	0.75	2.64	1.00	0.24	0.75	2.64	1.00	0.20	0.75	2.64	1.00	0.16
	VI	0.79	2.11	1.00	0.11	0.79	2.11	1.00	0.10	0.79	2.11	1.00	0.10
	VII	0.77	2.87	0.98	0.06	0.77	2.87	0.98	0.04	0.76	2.87	0.97	0.02
50	I	0.75	2.87	1.00	0.05	0.76	2.87	1.00	0.04	0.76	2.87	1.00	0.02
	II	0.71	1.93	1.00	0.02	0.71	1.93	1.00	0.02	0.71	1.93	0.99	0.01
	III	0.72	2.68	0.99	0.04	0.73	2.68	0.99	0.04	0.73	2.68	0.99	0.02
	IV	0.71	2.65	0.75	0.02	0.71	2.65	0.73	0.02	0.71	2.65	0.66	0.02
	V	0.73	3.32	0.99	0.07	0.74	3.32	0.99	0.06	0.74	3.32	0.99	0.05
	VI	0.78	2.47	1.00	0.03	0.78	2.47	1.00	0.03	0.78	2.47	1.00	0.02
	VII	0.76	3.04	0.97	0.02	0.76	3.04	0.97	0.02	0.76	3.04	0.95	0.01

Table A.4 The impact of α on the estimated regime. In the header, α is the tuning parameter in the stopping criterion and $\widehat{\pi}_\alpha$ is the regime such obtained. For each pair of regimes $\widehat{\pi}_\alpha$ and $\widehat{\pi}_{\alpha'}$, we report the probability that they recommend the same treatment for a randomly selected patient in the population. Mathematically, this is to compute $\Pr\{\widehat{\pi}_\alpha(X) = \widehat{\pi}_{\alpha'}(X)|\widehat{\pi}_\alpha, \widehat{\pi}_{\alpha'}\}$ and then average over 1000 replications, where X is generated in the same way as in Section 3 in the main paper.

p	Setting	$\widehat{\pi}_{0.9}$ vs. $\widehat{\pi}_{0.95}$	$\widehat{\pi}_{0.95}$ vs. $\widehat{\pi}_{0.99}$	$\widehat{\pi}_{0.9}$ vs. $\widehat{\pi}_{0.99}$
Continuous response				
10	I	0.998	0.998	0.996
	II	0.986	0.975	0.961
	III	0.993	0.992	0.986
	IV	0.997	0.997	0.993
	V	0.999	1.000	0.998
	VI	0.998	0.997	0.995
	VII	0.991	0.988	0.979
50	I	0.984	0.985	0.970
	II	0.986	0.977	0.962
	III	0.988	0.989	0.976
	IV	0.998	0.996	0.994
	V	0.993	0.995	0.988
	VI	0.997	0.997	0.993
	VII	0.991	0.989	0.980
Binary response				
10	I	0.971	0.969	0.941
	II	0.978	0.964	0.944
	III	0.971	0.952	0.926
	IV	0.983	0.955	0.941
	V	0.973	0.969	0.944
	VI	0.992	0.993	0.985
	VII	0.976	0.968	0.944
50	I	0.973	0.946	0.920
	II	0.980	0.958	0.942
	III	0.971	0.947	0.925
	IV	0.985	0.962	0.947
	V	0.965	0.939	0.913
	VI	0.985	0.985	0.969
	VII	0.974	0.955	0.930

9. Having problems with primary support group: 1 if yes, 0 if no;
10. Having problems related to the social environment: 1 if yes, 0 if no;
11. Having occupational problems: 1 if yes, 0 if no;
12. Having economic problems: 1 if yes, 0 if no;
13. Receiving psychotherapy for current depression: 1 if yes, 0 if no or don't know;
14. Receiving medication for current depression: 1 if yes, 0 if no or don't know;
15. Having received psychotherapy for past depressions: 1 if yes, 0 if no or don't know;
16. Having received medication for past depressions: 1 if yes, 0 if no or don't know;
17. Number of major depressive disorder (MDD) episodes: 1 if one, 2 if two, 3 if at least three;
18. Length of current MDD episode (in years): continuous;
19. Age at MDD onset: continuous;
20. MDD severity: 1 if mild, 2 if moderate, 3 if severe;
21. MDD type I: 1 if melancholic, 0 otherwise;
22. MDD type II: 1 if atypical, 0 otherwise;
23. Number of dysthymia episodes: 0 if zero, 1 if one, 2 if at least two;
24. Generalized anxiety: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

25. Anxiety disorder (not otherwise specified): 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

26. Panic disorder: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

27. Social phobia: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

28. Specific phobia: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

29. Obsessive compulsive: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

30. Body dysmorphic disorder: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

31. Post-traumatic stress disorder: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

32. Anorexia nervosa: 0 if absent or inadequate information, 1 if subthreshold, 2 if threshold;

33. Alcohol abuse: 0 if absent, 1 if abuse, 2 if dependent;

34. Drug abuse (including cannabis, stimulant, opioid, cocaine, hallucinogen): 0 if absent, 1 if abuse, 2 if dependent;

35. Global assessment of functioning: continuous;

36. Chronic depression diagnosis I: 1 if no antecedent dysthymia and continuous full-syndrome type;

37. Chronic depression diagnosis II: 1 if no antecedent dysthymia and incomplete recovery type;

38. Chronic depression diagnosis III: 1 if superimposed on antecedent dysthymia;

39. Chronic depression severity: integer between 1 (normal) and 7 (extremely ill);

40. Hamilton anxiety rating scale (HAM-A) total score: continuous;

41. HAM-A psychic anxiety score: continuous;

42. HAM-A somatic anxiety score: continuous;

43. Hamilton depression rating scale (HAM-D) total score: continuous;

44. HAM-D anxiety/somatic score: continuous;

45. HAM-D cognitive disturbance score: continuous;

46. HAM-D retardation score: continuous;

47. HAM-D sleep disturbance: continuous;

48. Inventory of Depressive Symptoms - Self Report (IDS-SR) anxiety/arousal score: continuous;

49. IDS-SR general/mood cognition score: continuous;

50. IDS-SR sleep score: continuous.

A.10 Consistency of the Decision List

Since the consistency of the decision list is difficult to analyze theoretically, we present some empirical evidence that the decision list tends to be consistent. We follow the simulated experiments considered in Section 4 in the main paper but increase the sample size. We consider settings I and V only as the optimal regime in other settings cannot be representable as a decision list.

The sample sizes considered and the associated results are presented in Table A.5. For continuous response, the proposed method correctly identifies the form and the important covariates for treatment decision. As n increases, the loss in value decreases and the probability of recommending the best treatment increases. Also, the mean squared error of estimating the cutoff values decreases at the rate of n^{-1}. Results for binary response is qualitatively similar. Nevertheless, we may need a even larger sample size for the asymptotics to work. Therefore, the simulation results provides evidence that the proposed method is consistent.

Table A.5 Consistency of the decision list. In the header, n is the sample size; p is the number of predictors. Loss is $R(\pi^{\text{opt}}) - R(\widehat{\pi})$, namely, the difference between the the value under the estimated regime and the value under the optimal regime. Pr(best) is $\Pr\{\widehat{\pi}(X) = \pi^{\text{opt}}(X)|\widehat{\pi}\}$, namely, the probability that the treatment recommended by the estimate regime coincides with the treatment recommended by the optimal regime. Loss and Pr(best) are averaged over 1000 replications. Correct is the proportion of $\widehat{\pi}$ having the same form and covariates as π^{opt} among 1000 replications; MSE_1 is the mean squared error of the estimated cutoff for X_1; MSE_2 is the mean squared error of the estimated cutoff for X_2.

Setting	n	p	Loss	Pr(best)	Correct	$\text{MSE}_1(\times n)$	$\text{MSE}_2(\times n)$
Continuous response							
I	10^4	10	0.0023	0.9982	1.00	4.24	6.87
I	10^5	10	0.0006	0.9995	1.00	4.30	6.50
I	10^6	10	0.0002	0.9998	1.00	4.06	6.32
I	10^4	50	0.0022	0.9982	1.00	4.60	6.91
I	10^5	50	0.0006	0.9995	1.00	4.24	6.49
I	10^6	50	0.0002	0.9998	1.00	4.22	6.48
V	10^4	10	0.0039	0.9975	1.00	6.08	5.27
V	10^5	10	0.0010	0.9994	1.00	5.86	4.70
V	10^6	10	0.0003	0.9998	1.00	5.46	4.54
V	10^4	50	0.0036	0.9977	1.00	6.10	5.33
V	10^5	50	0.0010	0.9994	1.00	5.96	4.54
V	10^6	50	0.0003	0.9998	1.00	5.69	4.51
Binary response							
I	10^4	10	0.0007	0.9966	1.00	8.13	10.93
I	10^5	10	0.0001	0.9994	1.00	5.71	6.05
I	10^6	10	0.0000	0.9999	1.00	5.27	5.61
I	10^4	50	0.0007	0.9965	1.00	9.72	11.15
I	10^5	50	0.0001	0.9994	1.00	5.71	6.29
I	10^6	50	0.0000	0.9998	1.00	5.41	5.78
V	10^4	10	0.0094	0.9447	0.79	7.81	22.66
V	10^5	10	0.0081	0.9547	0.96	4.14	6.33
V	10^6	10	0.0079	0.9563	0.97	3.89	5.03
V	10^4	50	0.0099	0.9418	0.70	6.63	14.36
V	10^5	50	0.0078	0.9567	0.96	3.97	6.14
V	10^6	50	0.0081	0.9550	0.97	3.96	5.10

Appendix B

Supplementary Materials for Chapter 2

B.1 Proofs

B.1.1 Notations

For vector $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, define component-wise operations $\boldsymbol{u}^p = (u_1^p, \ldots, u_d^p)^T$, $p \in \mathbb{R}$, and $\boldsymbol{u} \circ \boldsymbol{v} = (u_1 v_1, \ldots, u_d v_d)^T$ where p is some real number. For $V \subset \mathbb{R}^d$, define $u \circ V = \{u \circ v : v \in V\}$. In addition, $\boldsymbol{u}$ is said to be positive if its every component is positive.

Let $\boldsymbol{O}_i$ be the collection of random variables associated with the ith subject. For any function f, define $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(\boldsymbol{O}_i)$.

For any measurable function f defined on $D \subset \mathbb{R}^d$, we write $\|f\|_2 = \left(\int_D f^2 d\mu \right)^{1/2}$ and $\|f\|_\infty = \inf\{t \in \mathbb{R} : \mu(|f| > t) = 0\}$, where μ is the Lebesgue measure on D.

Let (T, d) be a metric space and S be a subset of T. For $\varepsilon > 0$, the ε-covering number

of S is defined by $\mathcal{N}(S, d, \varepsilon) = \inf\{n \geq 1 : \text{ there exists } t_1, \cdots, t_n \in T \text{ such that } S \subset \bigcup_{i=1}^{n} B(t_i, \varepsilon)\}$, where $\inf \emptyset = \infty$ and $B(t, \varepsilon) = \{u \in T : d(u, t) \leq \varepsilon\}$ is the a ball with center t and radius ε. If $(T, \|\cdot\|)$ is a normed vector space, the ε-covering number is defined by viewing T as a metric space with induced metric $d(s, t) = \|s - t\|$.

Let $(T, \|\cdot\|)$ be a normed vector space. The unit ball of T is defined by $\mathcal{B}_T = \{t : \|t\| \leq 1\}$. Given a scalar $w \in \mathbb{R}$ and a set $S \subset T$, define $wV = \{ws : s \in S\}$.

In the following proofs, c and c_i's denote constants, which may vary from occurrence to occurrence.

B.1.2 Concentration inequalities

We first state the Talagrand's inequality (Bousquet, 2002, Theorem 2.3; see also Massart, 2000, Theorem 3 and Boucheron et al., 2013, Theorem 12.5).

Proposition 2. *Let $\mathcal{F}$ be a countable set of functions. Suppose $\mathrm{E}\, f = 0$, $\mathrm{E}\, f^2 \leq V$, $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Denote $Z = \sup_{f \in \mathcal{F}} |\, \mathbb{P}_n f|$. Then for all $\tau > 0$, we have*

$$\Pr\left[Z \geq \mathrm{E}\, Z + \left\{\frac{2V\tau + 4B\tau(\mathrm{E}\, Z)}{n}\right\}^{1/2} + \frac{B\tau}{3n}\right] \leq e^{-\tau}.$$

Corollary 3. *Under the conditions in Proposition 2, we have*

$$\Pr\left\{Z \geq 2\,\mathrm{E}\, Z + \left(\frac{2V\tau}{n}\right)^{1/2} + \frac{2B\tau}{n}\right\} \leq e^{-\tau}.$$

Proof. It is clear that

$$\left\{\frac{2V\tau + 4B(\mathrm{E}\, Z)\tau}{n}\right\}^{1/2} \leq \left(\frac{2V\tau}{n}\right)^{1/2} + \left\{\frac{4B\tau(\mathrm{E}\, Z)}{n}\right\}^{1/2} \leq \left(\frac{2V\tau}{n}\right)^{1/2} + \frac{B\tau}{n} + \mathrm{E}\, Z.$$

Note that we use a larger constant for simplicity. □

When the variance of f is unavailable, we have the following proposition (Boucheron et al., 2013, Theorem 12.1).

Proposition 4. *Let $\mathcal{F}$ be a countable set of functions. Suppose $\mathrm{E}\, f = 0$, $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Denote $Z = \sup_{f\in\mathcal{F}} |\, \mathbb{P}_n f|$. Then for all $\tau > 0$, we have*

$$\Pr\left\{Z \geq \mathrm{E}\, Z + \left(\frac{2B^2\tau}{n}\right)^{1/2}\right\} \leq e^{-\tau}.$$

Next, we establish bounds on $\mathrm{E}\, Z$.

Proposition 5. *Let $\mathcal{F}$ be a countable set of functions and $\mathcal{F}$ contains the zero function. Assume*

$$\sup_Q \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^2(Q)}, \varepsilon) \leq \psi(\varepsilon)$$

for some function $\psi(\cdot)$, where the supremum is taken over all discrete probability measure Q. Suppose $\mathrm{E}\, f = 0$, $\mathrm{E}\, f^2 \leq V$, $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Denote $Z = \sup_{f\in\mathcal{F}} |\, \mathbb{P}_n f|$. Then we have

$$\mathrm{E}\, Z \leq 1024\left(\frac{BJ_V}{n}\right) + 64\left(\frac{VJ_V}{n}\right)^{1/2},$$

where $J_V = \int_0^1 \psi(V^{1/2}\varepsilon)\, d\varepsilon$.

Proof. Without loss of generality, we assume $B = 1$. The general case can be obtained

by scaling f. The proof extends the idea in Boucheron et al. (2013, Lemma 13.5).

Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. Rademacher random variable, namely $\Pr(\sigma = 1) = \Pr(\sigma = -1) = 1/2$. By the symmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.1), we have $\mathrm{E}(n^{1/2} Z) \leq 2\,\mathrm{E} \sup_f |n^{1/2}\,\mathbb{P}_n\,\sigma f|$.

Conditional on all random variables except σ_is, by Hoeffding's inequality, the process $n^{1/2}\,\mathbb{P}_n\,\sigma f$ is subgaussian with respect to the metric $\|f - g\|_{L^2(\mathbb{P}_n)} = \{\mathbb{P}_n (f-g)^2\}^{1/2}$. Hence the chaining technique (van der Vaart and Wellner, 1996, Corollary 2.2.8) implies

$$\mathrm{E}_\sigma \sup_f |n^{1/2}\,\mathbb{P}_n\,\sigma f| \leq 4 \int_0^{\eta_n} \left\{ \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P}_n)}, \varepsilon) \right\}^{1/2} d\varepsilon,$$

where E_σ denote the expectation with respect to $\sigma_1, \ldots, \sigma_n$ only and $\eta_n^2 = \max\{\sup_f (\mathbb{P}_n\, f^2), V\}$. Hence, we obtain

$$\mathrm{E}_\sigma \sup_f |n^{1/2}\,\mathbb{P}_n\,\sigma f| \leq 4 \int_0^{\eta_n} \psi^{1/2}(\varepsilon)\, d\varepsilon = 4\eta_n \int_0^1 \psi^{1/2}(\eta_n \varepsilon)\, d\varepsilon \leq 4\eta_n \int_0^1 \psi^{1/2}(V^{1/2}\varepsilon)\, d\varepsilon.$$

Since $\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P}_n)}, \varepsilon) \leq \psi(\varepsilon)$ and $\psi(\varepsilon)$ is a decreasing function in ε.

Taking the other layer of expectation, we get

$$\mathrm{E}(n^{1/2} Z) \leq 8(\mathrm{E}\,\eta_n) \int_0^1 \psi^{1/2}(V^{1/2}\varepsilon)\, d\varepsilon \leq 8(\mathrm{E}\,\eta_n^2)^{1/2} J_V^{1/2}$$

by Jensen's inequality. Also, we have $\mathrm{E}\,\eta_n^2 \leq \mathrm{E} \sup_f |\,\mathbb{P}_n\, f^2 - \mathrm{E}\, f^2| + V$ since $\mathrm{E}\, f^2 \leq V$ for all f. By the symmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.1), we have $\mathrm{E} \sup_f |\,\mathbb{P}_n\, f^2 - \mathrm{E}\, f^2| \leq 2\,\mathrm{E} \sup_f |\,\mathbb{P}_n\, \sigma f^2|$. By the contraction inequality (van der Vaart and Wellner, 1996, Proposition A.3.2) and $\|f\|_\infty \leq 1$, we have $\mathrm{E} \sup_f |\,\mathbb{P}_n\, \sigma f^2| \leq$

$4\,\mathrm{E}\sup_f |\,\mathbb{P}_n \sigma f|$. By the desymmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.6), we have $\mathrm{E}\sup_f |\,\mathbb{P}_n \sigma f| \le 2\,\mathrm{E}\sup_f |\,\mathbb{P}_n f|$. Combine these inequalities, we have $\mathrm{E}\,\eta_n^2 \le 16\,\mathrm{E}\,Z + V$.

Therefore, we have

$$n^{1/2}\,\mathrm{E}\,Z \le 8(16\,\mathrm{E}\,Z + V)^{1/2} J_V^{1/2}.$$

Solving for $\mathrm{E}\,Z$, we have $\mathrm{E}\,Z \le (2n)^{-1}\{a + (a^2 + 4nb)^{1/2}\} \le n^{-1}a + n^{-1/2}b^{1/2}$ with $a = 1024 J_V$ and $b = 64 V J_V$. Hence we have $\mathrm{E}\,Z \le 1024 n^{-1} J_V + 64 n^{-1/2} V^{1/2} J_V^{1/2}$. □

Proposition 6. *Let $\mathcal{F}$ be a countable set of functions and $\mathcal{F}$ contains the zero function. Assume*

$$\sup_Q \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^2(Q)}, \varepsilon) \le \psi(\varepsilon)$$

for some function $\psi(\cdot)$, where the supremum is taken over all discrete probability measure Q. Suppose $\mathrm{E}\,f = 0$, $\|f\|_\infty \le B$ for all $f \in \mathcal{F}$. Denote $Z = \sup_{f\in\mathcal{F}} |\,\mathbb{P}_n f|$. Then we have

$$\mathrm{E}\,Z \le 8\left(\frac{B^2 J_B}{n}\right)^{1/2},$$

where $J_B = \int_0^1 \psi(B\varepsilon)\,d\varepsilon$.

Proof. Just apply the trivial bound $|\eta_n| \le B$ in the proof of Proposition 5. □

Though all the propositions in this subsection assume that $\mathcal{F}$ is countable, they all apply if $\mathcal{F}$ is uncountable and separable, because we have $\Pr\left(\sup_{f\in\mathcal{F}} |\,\mathbb{P}_n f| = \sup_{f\in\mathcal{F}'} |\,\mathbb{P}_n f|\right) = 1$ for some countable subset $\mathcal{F}' \subset \mathcal{F}$.

B.1.3 Properties of RKHS

We establish several useful properties of the RKHS $\mathbb{H}$ induced by the Gaussian kernel with individual scaling factor in each dimension

$$K_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{z}) = \exp\left\{-\sum_{j=1}^{d} \gamma_j (x_j - z_j)^2\right\},$$

where $\boldsymbol{x}, \boldsymbol{z} \in D \subset \mathbb{R}^d$. The lemmas below extend the properties of Gaussian kernel with a single scaling factor.

We may omit $\boldsymbol{\gamma}$ and write $K(\cdot, \cdot)$ when the value of $\boldsymbol{\gamma}$ is clear in the context. Similarly, to emphasize the dependence of $\mathbb{H}$ on the parameter $\boldsymbol{\gamma}$ and the domain D, we may write $\mathbb{H}_{\boldsymbol{\gamma}}$, $\mathbb{H}(D)$, or $\mathbb{H}_{\boldsymbol{\gamma}}(D)$, depending on the context.

The following lemma provides a feature map of the Gaussian kernel.

Lemma 7. *Define the function* $\phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}} : \mathbb{R}^d \to L^2(\mathbb{R}^d)$ *by*

$$\phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}(\boldsymbol{u}) = \left(\frac{4}{\pi}\right)^{d/4} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/4} \exp\left\{-\sum_{j=1}^{d} 2\gamma_j (x_j - u_j)^2\right\}, \ \boldsymbol{x} \in D, \ \boldsymbol{u} \in \mathbb{R}^d.$$

Then $\phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}$ *is a feature map of* $K_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{z})$.

Proof. Straightforward calculation similar to Steinwart and Christmann (2008, Lemma 4.45) gives $\langle \phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}, \phi_{\boldsymbol{\gamma}}^{\boldsymbol{z}} \rangle_{L^2(\mathbb{R}^d)} = K_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{z})$. By definition, $\phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}$ is a feature map. □

The following lemma shows that $\mathbb{H}_{\boldsymbol{\gamma}}(D)$ can be embedded into $\mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}(D)$ if $\gamma_j < \widetilde{\gamma}_j$ for all $j = 1, \ldots, d$.

Lemma 8. *Let $\boldsymbol{\gamma}$, $\widetilde{\boldsymbol{\gamma}}$ be two positive vectors satisfying $\gamma_j < \widetilde{\gamma}_j$ for all j. If $f \in \mathbb{H}_{\boldsymbol{\gamma}}$, then $f \in \mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}$ and $\|f\|_{\mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}} \leq \left(\prod_{j=1}^{d} \widetilde{\gamma}_j\right)^{1/4} \left(\prod_j \gamma_{j=1}^{d}\right)^{-1/4} \|f\|_{\mathbb{H}_{\boldsymbol{\gamma}}}$.*

Proof. We follow the strategy in Steinwart and Christmann (2008, Theorem 4.46). Since $f \in \mathbb{H}_{\boldsymbol{\gamma}}$, by Steinwart and Christmann (2008, Theorem 4.21), there exists $g \in L^2\left(\mathbb{R}^d\right)$ such that $f(\boldsymbol{x}) = \langle \phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}, g \rangle_{L^2(\mathbb{R}^d)}$ for all $\boldsymbol{x} \in D$.

Given $\boldsymbol{s} \in \mathbb{R}^d$ with $s_j > 0$ for all j, define the operator $W_s : L^2\left(\mathbb{R}^d\right) \to L^2\left(\mathbb{R}^d\right)$ by

$$(W_{\boldsymbol{s}} g)(\boldsymbol{v}) = \int_{\mathbb{R}^d} \pi^{-d/2} \left(\prod_{j=1}^{d} s_j\right)^{-1/2} \exp\left\{ -\sum_{j=1}^{d} s_j^{-1}(v_j - u_j)^2 \right\} g(\boldsymbol{u}) d\boldsymbol{u}, \text{ for } \boldsymbol{v} \in \mathbb{R}^d.$$

For any $g \in L^2(\mathbb{R}^d)$ and any $\boldsymbol{v} \in \mathbb{R}^d$, straightforward calculation using properties of normal densities shows $(W_{\boldsymbol{s}_1} W_{\boldsymbol{s}_2} g)(\boldsymbol{v}) = (W_{\boldsymbol{s}_1 + \boldsymbol{s}_2} g)(\boldsymbol{v})$. Hence, we have $W_{\boldsymbol{s}_1} W_{\boldsymbol{s}_2} = W_{\boldsymbol{s}_1 + \boldsymbol{s}_2}$.

Define $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\tau}} = (\widetilde{\tau}_1, \ldots, \widetilde{\tau}_d)^{\mathrm{T}}$, where $\tau_j = 1/\gamma_j$ and $\widetilde{\tau}_j = 1/\widetilde{\gamma}_j$. The assumption $\gamma_j < \widetilde{\gamma}_j$ implies $\tau_j > \widetilde{\tau}_j$. We observe that

$$f = \langle \phi_{\boldsymbol{\gamma}}^{\boldsymbol{x}}, g \rangle_{L^2(\mathbb{R}^d)} = (W_{\boldsymbol{\tau}/2} g) \cdot \pi^{d/4} \left(\prod_{j=1}^{d} \gamma_j\right)^{-1/4}.$$

Since

$$W_{\boldsymbol{\tau}/2} g = W_{\widetilde{\boldsymbol{\tau}}/2} W_{(\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}})/2} g = \langle \phi_{\widetilde{\boldsymbol{\gamma}}}^{\boldsymbol{x}}, W_{(\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}})/2} g \rangle_{L^2(\mathbb{R}^d)} \cdot \pi^{-d/4} \left(\prod_{j=1}^{d} \widetilde{\gamma}_j\right)^{1/4},$$

we have

$$f = \langle \phi_{\widetilde{\boldsymbol{\gamma}}}^{\boldsymbol{x}}, W_{(\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}})/2} g \rangle_{L^2(\mathbb{R}^d)} \cdot \left(\prod_{j=1}^{d} \gamma_j\right)^{-1/4} \left(\prod_{j=1}^{d} \widetilde{\gamma}_j\right)^{1/4}.$$

By Steinwart and Christmann (2008, Theorem 4.21), we conclude $f \in \mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}$.

Moreover, $\|f\|_{\mathbb{H}_{\boldsymbol{\gamma}}} = \|g\|_{L^2(\mathbb{R}^d)}$ and $\|f\|_{\mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}} = \|W_{(\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}})/2} g\|_{L^2(\mathbb{R}^d)} \cdot \left(\prod_{j=1}^{d} \gamma_j\right)^{-1/4} \left(\prod_{j=1}^{d} \widetilde{\gamma}_j\right)^{1/4}$.

By Young's inequality, $\|W_{(\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}})/2}g\|_{L^2(\mathbb{R}^d)} \leq \|g\|_{L^2(\mathbb{R}^d)}$. Hence, we have

$$\|f\|_{\mathbb{H}_{\widetilde{\boldsymbol{\gamma}}}} \leq \|g\|_{L^2(\mathbb{R}^d)} \left(\prod_{j=1}^d \gamma_j\right)^{-1/4} \left(\prod_{j=1}^d \widetilde{\gamma}_j\right)^{1/4} \leq \|f\|_{\mathbb{H}_{\boldsymbol{\gamma}}} \left(\prod_{j=1}^d \gamma_j\right)^{-1/4} \left(\prod_{j=1}^d \widetilde{\gamma}_j\right)^{1/4}.$$

□

The following lemma establishes isometric isomorphism between $\mathbb{H}_{\boldsymbol{\alpha}^{-2}\circ\boldsymbol{\gamma}}(\boldsymbol{\alpha}\circ D)$ and $\mathbb{H}_{\boldsymbol{\gamma}}(D)$ for any fixed $\boldsymbol{\alpha}$.

Lemma 9. *Let $\boldsymbol{\alpha}$ be an arbitrary positive vector. We define a mapping $\tau_{\boldsymbol{\alpha}} : L^\infty(D) \to L^\infty(\boldsymbol{\alpha}\circ D)$ as follows: given a function $f \in L^\infty(D)$, let $\tau_{\boldsymbol{\alpha}}(f)(\boldsymbol{x}) = f(\boldsymbol{\alpha}^{-1}\circ\boldsymbol{x})$ for $\boldsymbol{x} \in \boldsymbol{\alpha}\circ D$. Then for all $f \in \mathbb{H}_{\boldsymbol{\gamma}}(D)$, we have $\tau_{\boldsymbol{\alpha}}(f) \in \mathbb{H}_{\boldsymbol{\alpha}^{-2}\circ\boldsymbol{\gamma}}(\boldsymbol{\alpha}\circ D)$ and $\|\tau_{\boldsymbol{\alpha}}(f)\|_{\mathbb{H}_{\boldsymbol{\alpha}^{-2}\circ\boldsymbol{\gamma}}(\boldsymbol{\alpha}\circ D)} = \|f\|_{\mathbb{H}_{\boldsymbol{\gamma}(D)}}$.*

Proof. It is easy to verify that the arguments in Steinwart and Christmann (2008, Proposition 4.37) remains valid when scalar multiplication is replaced by component-wise multiplication between vectors. □

The following lemma computes the covering number of the unit ball in $\mathbb{H}_{\boldsymbol{\gamma}}(D)$.

Lemma 10. *Suppose $D \subset s\mathcal{B}_{\mathbb{R}^d}$. For any integer $m \geq 1$, we have*

$$\log \mathcal{N}\{\mathcal{B}_{\mathbb{H}_{\boldsymbol{\gamma}}(D)}, \|\cdot\|_\infty, \varepsilon\} \leq c_{m,d,s} \prod_{j=1}^d (1+\gamma_j)^{1/2} \varepsilon^{-d/m},$$

where $c_{m,d,s}$ is a constant that depends on m, d and s only.

Proof. Let $\mathbf{1}$ be the vector of ones. By Lemma 3, $\mathbb{H}_{\boldsymbol{\gamma}}(D)$ is isometric isomorphic to $\mathbb{H}_{\mathbf{1}}(\boldsymbol{\gamma}^{1/2}\circ D)$. Thus it suffices to compute the covering number for $\mathbb{H}_{\mathbf{1}}(\boldsymbol{\gamma}^{1/2}\circ D)$.

Define $\widetilde{D} = \boldsymbol{\gamma}^{1/2} \circ D$. It is shown that $\mathbb{H}_{\mathbf{1}}(\widetilde{D})$ can be embedded into $\mathbb{C}^m(\widetilde{D})$ (Steinwart and Christmann, 2008, Theorem 6.26). By Steinwart and Christmann (2008, Corollary 4.36), the embedding map from $\mathbb{H}_{\mathbf{1}}(\widetilde{D})$ to $\mathbb{C}^m(\widetilde{D})$ is continuous, and hence bounded. Thus, there exists a constant c_1 which depends only on m such that $\|f\|_{\mathbb{C}^m(\widetilde{D})} \leq c_1 \|f\|_{\mathbb{H}_{\mathbf{1}}(\widetilde{D})}$ for all $f \in \mathbb{H}_{\mathbf{1}}(\widetilde{D})$. Hence we have

$$\mathcal{N}\{\mathcal{B}_{\boldsymbol{H}_{\mathbf{1}}(\widetilde{D})}, \|\cdot\|_\infty, \varepsilon\} \leq \mathcal{N}(c_1 \mathcal{B}_{\mathbb{C}^m(\widetilde{D})}, \|\cdot\|_\infty, \varepsilon\} = \mathcal{N}(\mathcal{B}_{\mathbb{C}^m(\widetilde{D})}, \|\cdot\|_\infty, \varepsilon/c_1).$$

By van der Vaart and Wellner (1996, Theorem 2.7.1), there exists a constant c_2 which depends only on m and d such that

$$\log \mathcal{N}\{\mathcal{B}_{\mathbb{C}^m(\widetilde{D})}, \|\cdot\|_\infty, \varepsilon) \leq c_2 \mu(\{\boldsymbol{x} : \|\boldsymbol{x} - \widetilde{D}\| \leq 1\}) \varepsilon^{-d/m},$$

where μ is the Lebesgue measure on $\mathbb{R}^d$. Since $D \subset s\mathcal{B}_{\mathbb{R}^d}$ and $(1+su^{1/2}) \leq (1+s)(1+u)^{1/2}$ for all $u \geq 0$, we have

$$\mu(\{\boldsymbol{x} : \|\boldsymbol{x} - \widetilde{D}\| \leq 1\}) \leq \prod_{j=1}^{d} (1 + s\lambda_j^{1/2}) \leq (1+s)^d \prod_{j=1}^{d} (1+\lambda_j)^{1/2}.$$

□

B.1.4 Approximation error in kernel ridge regression

Define $\widetilde{Y}_T = Y_T$ and $\widetilde{Y}_t = Y_t + Q_{t+1}\{\boldsymbol{X}_{t+1}, \pi_{t+1}^*(\boldsymbol{X}_{t+1})\}$ for $t < T$. Then we have $Q_t(\boldsymbol{x}, a) = E(\widetilde{Y}_t | \boldsymbol{X}_t = \boldsymbol{x}, A_t = a)$ for all t. Fix a stage t and a treatment $a \in \mathcal{A}_t$. For notation simplicity, we shall omit the subscripts t and a afterwards. Given a function $f \in L^\infty(D)$,

we define

$$\mathcal{L}(f) = \mathrm{E}\left[I(A=a)\{\widetilde{Y} - f(\boldsymbol{X})\}^2\right].$$

We also define

$$f_0 = \underset{f:D\to\mathbb{R},\text{ measurable}}{\arg\min} \mathcal{L}(f).$$

Simple calculations show that $f_0(\boldsymbol{x}) = E(\widetilde{Y}|\boldsymbol{X}=\boldsymbol{x}, A=a)$ almost surely with respect to $P_{\boldsymbol{X}}$, where $P_{\boldsymbol{X}}$ is the distribution of $\boldsymbol{X}$. Hence f_0 is exactly $Q_t(\cdot, a)$. Besides, we have

$$\mathcal{L}(f) - \mathcal{L}(f_0) = E\left[I(A=a)\left\{f(\boldsymbol{X}) - f_0(\boldsymbol{X})\right\}^2\right].$$

It is likely that f_0 doesn't belong to the RKHS $\mathbb{H}_{\boldsymbol{\gamma}}$. Nevertheless, the estimator must live in $\mathbb{H}_{\boldsymbol{\gamma}}$. The following proposition shows that it is always possible to find an $f \in \mathbb{H}_{\boldsymbol{\gamma}}$ such that f and f_0 is close. The following proposition is a stronger version of Eberts and Steinwart (2013, Theorems 2.2 and 2.3) which handles multiple scaling factors and make the separation between signal variables and noise variables.

Proposition 11. *Suppose f_0 satisfies the modules of smoothness condition $\omega(f_0, s) \le c_1 s^r$ for some positive integer r, and $\|f_0\|_\infty \le B$ for some constant B. Let $\mathcal{S}$ be the index of signal variables in f_0. Namely, the value of $f(\boldsymbol{x})$ only depends on $\boldsymbol{x}_{\mathcal{S}}$. Then there exists some $f \in \mathbb{H}_{\boldsymbol{\gamma}}$ such that*

$$\lambda\|f\|^2_{\mathbb{H}_{\boldsymbol{\gamma}}} + \|f - f_0\|^2_\infty \le c\left\{\lambda\big(\max_{j\in\mathcal{S}}\gamma_j\big)^{|\mathcal{S}|/2}\big(\max_{j\in\mathcal{S}^c}\gamma_j\big)^{|\mathcal{S}^c|/2} + \big(\min_{j\in\mathcal{S}}\gamma_j\big)^{-r}\right\}$$

and $\|f\|_\infty \le 2^r B$, where c is some constant that depends on c_1, r, B and $|\mathcal{S}|$ only.

Proof. Define

$$W(\boldsymbol{x},\boldsymbol{u}) = \sum_{i=1}^{r} \binom{r}{i} (-1)^{i-1} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} i^{-d} \exp\left\{-\sum_{j=1}^{d} 2\gamma_j (x_j - u_j)^2/i^2\right\},$$

where $\boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^d$. Let $f(\boldsymbol{x}) = \int_{\mathbb{R}^d} W(\boldsymbol{x},\boldsymbol{u}) f_0(\boldsymbol{u})\, d\boldsymbol{u}$, $\boldsymbol{x} \in D$.

Then, for every $\boldsymbol{x} \in D$, we have

$$f(\boldsymbol{x}) = \sum_{i=1}^{r} \binom{r}{i} (-1)^{i-1} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} \int_{\mathbb{R}^d} i^{-d} \exp\left\{-\sum_{j=1}^{d} 2\gamma_j (x_j - u_j)^2/i^2\right\} f_0(\boldsymbol{u}) d\boldsymbol{u}.$$

By a change of variable $h_j = (u_j - x_j)/i$, we have

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{r} \binom{r}{i} (-1)^{i-1} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} \int_{\mathbb{R}^d} \exp\left\{-\sum_{j=1}^{d} 2\gamma_j h_j^2\right\} f_0(\boldsymbol{x} + i\boldsymbol{h}) d\boldsymbol{h} \\ &= \int_{\mathbb{R}^d} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} \exp\left(-\sum_{j=1}^{d} 2\gamma_j h_j^2\right) \sum_{i=1}^{r} \binom{r}{i} (-1)^{i-1} f_0(\boldsymbol{x} + i\boldsymbol{h}) d\boldsymbol{h} \end{aligned}$$

Note that

$$f_0(\boldsymbol{x}) = \int_{\mathbb{R}^d} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} \exp\left(-\sum_{j=1}^{d} 2\gamma_j h_j^2\right) f_0(\boldsymbol{x}) d\boldsymbol{h}.$$

Hence, we have

$$|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \le \int_{\mathbb{R}^d} \left(\frac{2}{\pi}\right)^{d/2} \left(\prod_{j=1}^{d} \gamma_j\right)^{1/2} \exp\left(-\sum_{j=1}^{d} 2\gamma_j h_j^2\right) |\Delta_{\boldsymbol{h}}^r(f_0, \boldsymbol{x})|\, d\boldsymbol{h}.$$

Since $f_0(\boldsymbol{x}) = f_0^*(\boldsymbol{x}_{\mathcal{S}})$ for some function $f_0^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$, we observe that

$$|\Delta_{\boldsymbol{h}}^r(f_0, \boldsymbol{x})| = \left|\Delta_{\boldsymbol{h}_{\mathcal{S}}}^r(f_0^*, \boldsymbol{x}_{\mathcal{S}})\right| \leq \omega_r(f_0^*, \|\boldsymbol{h}_{\mathcal{S}}\|_2) = \omega_r(f_0, \|\boldsymbol{h}_{\mathcal{S}}\|_2).$$

Hence, we have

$$|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \leq \int_{\mathbb{R}^{|\mathcal{S}|}} \left(\frac{2}{\pi}\right)^{|\mathcal{S}|/2} \left(\prod_j \gamma_{\mathcal{S},j}\right)^{1/2} \exp\left(-\sum_j 2_{\mathcal{S},j} h_{\mathcal{S},j}^2\right) \omega_r(f_0, \|\boldsymbol{h}_{\mathcal{S}}\|_2) d\boldsymbol{h}_{\mathcal{S}}.$$

Since $\omega_r(f_0, t) \leq (1 + t/s)^r \omega_r(f_0, s)$ for all $s, t > 0$, we have

$$\begin{aligned}\omega_r(f_0, \|\boldsymbol{h}_{\mathcal{S}}\|_2) &\leq \left\{1 + \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{1/2} \|\boldsymbol{h}_{\mathcal{S}}\|_2\right\}^r \omega_r\left\{f_0, \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{-1/2}\right\} \\ &\leq \left(1 + \|\boldsymbol{\gamma}_{\mathcal{S}} \circ \boldsymbol{h}_{\mathcal{S}}\|_2\right)^r \omega_r\left\{f_0, \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{-1/2}\right\}.\end{aligned}$$

Combining these inequalities, we have

$$\begin{aligned}|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \leq\ & \omega_r\left\{f_0, \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{-1/2}\right\} \cdot \\ & \int_{\mathbb{R}^{|\mathcal{S}|}} \left(\frac{2}{\pi}\right)^{|\mathcal{S}|/2} \left(\prod_j \gamma_{\mathcal{S},j}\right)^{1/2} \exp\left\{-\sum_j 2\gamma_{\mathcal{S},j} h_{\mathcal{S},j}^2\right\} \left(1 + \|\boldsymbol{\gamma}_{\mathcal{S}} \circ \boldsymbol{h}_{\mathcal{S}}\|_2\right)^r d\boldsymbol{h}_{\mathcal{S}}\end{aligned}$$

Using a change of variable $t_j = \gamma_{\mathcal{S},j} h_{\mathcal{S},j}$, we can see that the integral above is a constant depending on $|\mathcal{S}|$ only. Denoted the integral by c_2. Thus we have

$$|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \leq c_2 \omega_r\left\{f_0, \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{-1/2}\right\} \leq c_1 c_2 \left(\min_{j\in\mathcal{S}} \gamma_j\right)^{-1/2}.$$

Note that $W(\boldsymbol{x},\boldsymbol{u}) = \sum_{i=1}^{r}\binom{r}{i}(-1)^{i-1}\pi^{-d/4}i^{-d/2}\left(\prod_{j=1}^{d}\gamma_j\right)^{1/4}\phi^{\boldsymbol{x}}_{\boldsymbol{\gamma}/i^2}(\boldsymbol{u})$, where ϕ is the feature map defined in Lemma 1. Let $g_i(\boldsymbol{x}) = \int_{\mathbb{R}^d}\phi^{\boldsymbol{x}}_{\boldsymbol{\gamma}/i^2}(\boldsymbol{u})f_0(\boldsymbol{u})d\boldsymbol{u}$, then $g_i \in \mathbb{H}_{\boldsymbol{\gamma}/i^2}$. By Lemma 2, we have $g_i \in \mathbb{H}_{\boldsymbol{\gamma}}$ and the $\mathbb{H}_{\boldsymbol{\gamma}}$ norm of g_i is at most $i^{d/2}$ times of its $\mathbb{H}_{\boldsymbol{\gamma}/i^2}$ norm. Hence, we have

$$\|f\|_{\mathbb{H}} \le \sum_{i=1}^{r}\binom{r}{i}\pi^{-d/4}\left(\prod_{j=1}^{d}\gamma_j\right)^{1/4}\|f_0\|_2 \le 2^r\pi^{-d/4}\big(\max_{j\in\mathcal{S}}\gamma_j\big)^{|\mathcal{S}|/4}\big(\max_{j\in\mathcal{S}^c}\gamma_j\big)^{|\mathcal{S}^c|/4}\|f_0\|_2.$$

Therefore, we have

$$\begin{aligned}\lambda\|f\|_{\mathbb{H}}^2 + \mathcal{L}(f) - \mathcal{L}(f_0) &= \lambda\|f\|_{\mathbb{H}}^2 + \mathrm{E}\left\{f(\boldsymbol{X}) - f_0(\boldsymbol{X})\right\}^2 \\ &\le 2^{2r}\pi^{-d/2}B^2\lambda\big(\max_{j\in\mathcal{S}}\gamma_j\big)^{|\mathcal{S}|/2}\big(\max_{j\in\mathcal{S}^c}\gamma_j\big)^{|\mathcal{S}^c|/2} + c_1^2c_2^2\big(\min_{j\in\mathcal{S}}\gamma_j\big)^{-1}.\end{aligned}$$

In addition, for any $\boldsymbol{x}\in D$, we have

$$\begin{aligned}|f(\boldsymbol{x})| &\le \sum_{i=1}^{r}\binom{r}{i}\int_{\mathbb{R}^d}\left(\frac{2}{\pi}\right)^{d/2}\left(\prod_{j=1}^{d}\gamma_j\right)^{1/2} i^{-d}\exp\left\{-\sum_{j=1}^{d}2\gamma_j(x_j-u_j)^2/i^2\right\}d\boldsymbol{u}\cdot\|f_0\|_\infty \\ &= \sum_{i=1}^{r}\binom{r}{i}\|f_0\|_\infty \le 2^rB.\end{aligned}$$

□

B.1.5 Risk bounds for kernel ridge regression

Recall that the truncation operator $\mathcal{T}_B : L^\infty(D)\to L^\infty(D)$ is defined as

$$\mathcal{T}_B(f)(\boldsymbol{x}) = f(\boldsymbol{x})I\{-B\le f(\boldsymbol{x})\le B\} + BI\{f(\boldsymbol{x})>B\} + (-B)I\{f(\boldsymbol{x})<-B\},\quad \boldsymbol{x}\in D.$$

For any function f, g, we have $|\mathcal{T}_B(f)(\boldsymbol{x}) - \mathcal{T}_B(g)(\boldsymbol{x})| \leq |f(\boldsymbol{x}) - g(\boldsymbol{x})|$. Hence, we have $\|\mathcal{T}_B(f) - \mathcal{T}_B(g)\|_\infty \leq \|f - g\|_\infty$. As a consequence, for any $B \geq \|f_0\|_\infty$, we have

$$\mathcal{L}\{\mathcal{T}_B(f)\} - \mathcal{L}(f_0) = \mathrm{E}\{\mathcal{T}_B(f)(\boldsymbol{X}) - f_0(\boldsymbol{X})\}^2 \leq \mathrm{E}\{f(\boldsymbol{X}) - f_0(\boldsymbol{X})\}^2 = \mathcal{L}(f) - \mathcal{L}(f_0).$$

Define $\widehat{Y}_T = Y_T$ and $\widehat{Y}_t = Y_t + \widehat{Q}_{t+1}\{\boldsymbol{X}_{t+1}, \widehat{\pi}_{t+1}(\boldsymbol{X}_{t+1})\}$ for $t < T$. Given sequences $\boldsymbol{\gamma}_n$ and λ_n, the estimator of the Q-function is $\widehat{Q}_t(\cdot, a) = \mathcal{T}_B(\widehat{q}_n)$, where

$$\widehat{f}_n = \arg\min_{f \in \mathbb{H}_{\boldsymbol{\gamma}}} \mathbb{P}_n \, I(A = a)\{\widehat{Y} - f(\boldsymbol{X})\}^2 + \lambda \|f\|^2_{\mathbb{H}_{\boldsymbol{\gamma}}}.$$

To facilitate our analysis, we define

$$q_n = \arg\min_{f \in \mathbb{H}_{\boldsymbol{\gamma}}} \mathbb{P}_n \, I(A = a)\{\widetilde{Y} - f(\boldsymbol{X})\}^2 + \lambda \|f\|^2_{\mathbb{H}_{\boldsymbol{\gamma}}}.$$

Note that we omit the subscript n in $\boldsymbol{\gamma}_n$ and λ_n for simplicity. The difference between $\widehat{f}_n$ and q_n is that we use $\widetilde{Y}_t = Y_t + Q_{t+1}\{\boldsymbol{X}_{t+1}, \pi^*_{t+1}(\boldsymbol{X}_{t+1})\}$ for $t < T$ when defining $\widehat{f}_n$, which is an unobserved quantity since it relies on π^*_{t+1} and Q_{t+1}. In contrast, we replace π^*_{t+1} and Q_{t+1} by their estimates $\widehat{\pi}_{t+1}$ and $\widehat{Q}_{t+1}$ to obtain $\widehat{Y}_t$. Hence $\widehat{Q}_t(\cdot, a)$ is based on observed quantities only.

In this subsection, we will show that the difference between $\mathcal{T}_B(\widehat{f}_n)$ and $f_0 = Q_t(\cdot, a)$ is small. To be precise, define $\mathcal{E}(f) = \lambda \|f\|^2_{\mathbb{H}_{\boldsymbol{\gamma}}} + \mathcal{L}\{\mathcal{T}_B(f)\} - \mathcal{L}(f_0)$. Our goal is to show that $\mathcal{E}(\widehat{f}_n)$ is small with large probability. The proof below follows the idea in Steinwart and Christmann (2008, Theorem 7.20). For notation convenience, we define $\overline{\gamma}_{\mathcal{S}} = 1 + \max_{j \in \mathcal{S}} \gamma_j$, $\underline{\gamma}_{\mathcal{S}} = \min_{j \in \mathcal{S}} \gamma_j$ and $\overline{\gamma}_{\mathcal{S}^c} = 1 + \max_{j \in \mathcal{S}^c} \gamma_j$. For any f, define $\ell_f = I(A =$

$a)\{\widetilde{Y} - f(\boldsymbol{X})\}^2$ and $h_f = \ell_f - \ell_{f_0}$. Then we have $\mathcal{L}(f) - \mathcal{L}(f_0) = \mathrm{E}\, h_f$. Thus $\mathrm{E}\, h_f \geq 0$ for all f.

Lemma 12. *For any $f \in \mathbb{H}_\gamma$, we have*

$$\mathcal{E}(\widehat{f}_n) \leq \lambda \|f\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, h_f - \mathbb{P}_n\, h_{\widehat{f}_n} + \mathrm{E}\, h_{\mathcal{T}_B(\widehat{f}_n)} + 2\,\mathbb{P}_n \left(\widehat{Y} - \widetilde{Y}\right)^2.$$

Proof. By the definition of $\widehat{f}_n$ and q_n, we have

$$\begin{aligned}
\lambda \|\widehat{f}_n\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, I(A = a)\{\widehat{Y} - \widehat{f}_n(\boldsymbol{X})\}^2 &\leq \lambda \|q_n\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, I(A = a)\{\widehat{Y} - q_n(\boldsymbol{X})\}^2, \\
\lambda \|q_n\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, I(A = a)\{\widetilde{Y} - q_n(\boldsymbol{X})\}^2 &\leq \lambda \|f\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, I(A = a)\{\widehat{Y} - f(\boldsymbol{X})\}^2.
\end{aligned}$$

Thus, we have

$$\lambda \|\widehat{f}_n\|^2_{\mathbb{H}_\gamma} \leq \lambda \|f\|^2_{\mathbb{H}_\gamma} + \mathbb{P}_n\, h_f - \mathbb{P}_n\, h_{q_n} + \mathbb{P}_n\, I(A = a) \left\{\widehat{Y} - q_n(\boldsymbol{X})\right\}^2 - \mathbb{P}_n\, I(A = a) \left\{\widehat{Y} - \widehat{f}_n(\boldsymbol{X})\right\}^2.$$

For any number a_1, a_2, b_1, b_2, we have

$$\begin{aligned}
(a_1 - b_1)^2 - (a_1 - b_2)^2 &= (2a_1 - b_1 - b_2)(b_2 - b_1) \\
&= (2a_2 - b_1 - b_2)(b_2 - b_1) + 2(a_1 - a_2)(b_2 - b_1) \\
&\leq (a_2 - b_1)^2 - (a_2 - b_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2.
\end{aligned}$$

Hence, we have

$$\mathbb{P}_n\, I(A = a) \left\{\widehat{Y} - q_n(\boldsymbol{X})\right\}^2 - \mathbb{P}_n\, I(A = a) \left\{\widehat{Y} - \widehat{f}_n(\boldsymbol{X})\right\}^2$$

$$\leq \mathbb{P}_n h_{q_n} - \mathbb{P}_n h_{\widehat{f}_n} + \mathbb{P}_n I(A=a)\left(\widehat{Y} - \widetilde{Y}\right)^2 + \mathbb{P}_n I(A=a)\left\{\widehat{f}_n(\boldsymbol{X}) - q_n(\boldsymbol{X})\right\}^2.$$

Let $\widehat{\boldsymbol{Y}}$ be the vector of $\widehat{Y}_i$, $i \in \mathcal{I}_a$, $\widetilde{\boldsymbol{Y}}$ the vector of $\widetilde{Y}_i$, $i \in \mathcal{I}_a$ and $\boldsymbol{K}$ the matrix of $K(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j \in \mathcal{I}_a$, where $\mathcal{I}_a = \{i : A_i = a\}$. By the representer theorem and the eigenvalues of $\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}$, we have

$$\|\{\widehat{f}_n(\boldsymbol{X}_i)\}_{i\in\mathcal{I}_a} - \{q_n(\boldsymbol{X}_i)\}_{i\in\mathcal{I}_a}\|_2 = \|\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}(\widehat{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}})\|_2 \leq \|\widehat{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}\|_2.$$

Thus we have

$$\mathbb{P}_n I(A=a)\left\{\widehat{f}_n(\boldsymbol{X}) - q_n(\boldsymbol{X})\right\}^2 \leq \mathbb{P}_n I(A=a)\left(\widehat{Y} - \widetilde{Y}\right)^2$$

and the final inequality follows.

□

Proposition 13. *Suppose* $\Pr\left\{\mathbb{P}_n(\widehat{Y} - \widetilde{Y})^2 \geq c_1 n^{-\alpha} + c_2 n^{-\beta}\tau\right\} \leq e^{-\tau}$. *For any* $\delta > 0$ *and* $\tau > 0$, *we have*

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{f}_n(\boldsymbol{X}) - f_0(\boldsymbol{X})\right\}^2 \leq c\left(\lambda \overline{\gamma}_{\mathcal{S}}^{|\mathcal{S}|/2} \overline{\gamma}_{\mathcal{S}^c}^{|\mathcal{S}^c|/2} + \underline{\gamma}_{\mathcal{S}}^{-r} + \overline{\gamma}_{\mathcal{S}}^{|\mathcal{S}|/2} \overline{\gamma}_{\mathcal{S}^c}^{|\mathcal{S}^c|/2} \lambda^{-\delta} n^{-1} + n^{-1}\tau + n^{-\alpha} + n^{-\beta}\tau\right)\right] \leq e^{-}$$

where c *is a constant that depends on* δ, d, r, B *and* ϖ *only, and* $\mathrm{E}_{\boldsymbol{X}}$ *denotes the expectation with respect to* $\boldsymbol{X}$ *only.*

Proof. By Proposition 11 and the inequality $E\left[I(A=a)\left\{f(\boldsymbol{X}) - f_0(\boldsymbol{X})\right\}^2\right] \leq \|f -$

$f_0\|_\infty^2$, there exists some function $f_n \in \mathbb{H}_\gamma$ such that

$$\lambda\|f_n\|_{\mathbb{H}_\gamma}^2 + \mathrm{E}\, h_{f_n} \le c\left\{\lambda\big(\max_{j\in\mathcal{S}}\gamma_j\big)^{|\mathcal{S}|/2}\big(\max_{j\in\mathcal{S}^c}\gamma_j\big)^{|\mathcal{S}^c|/2} + \big(\min_{j\in\mathcal{S}}\gamma_j\big)^{-r}\right\} \tag{B.1}$$

for some constant c independent of n, and $\|f\|_\infty \le 2^r B$.

By the property of the truncation operator and the fact that $\|Y\|_\infty \le B$ with probability 1, we have $\mathbb{P}_n\, h_{\mathcal{T}_B(\widehat{f}_n)} \le \mathbb{P}_n\, h_{\widehat{f}_n}$. We apply Lemma 12 with $f = f_n$ and obtain

$$\begin{aligned}\mathcal{E}(\widehat{f}_n) &\le \lambda\|f_n\|_{\mathbb{H}_\gamma}^2 + \mathbb{P}_n\, h_{f_n} - \mathbb{P}_n\, h_{\mathcal{T}_B(\widehat{f}_n)} + \mathrm{E}\, h_{\mathcal{T}_B(\widehat{f}_n)} + \mathbb{P}_n(\widehat{Y}-\widetilde{Y})^2 \\ &\le (\lambda\|f_n\|_{\mathbb{H}_\gamma}^2 + \mathrm{E}\, h_{f_n}) + |\mathbb{P}_n\, h_{f_n} - \mathrm{E}\, h_{f_n}| + |\mathrm{E}\, h_{\mathcal{T}_B(\widehat{f}_n)} - \mathbb{P}_n\, h_{\mathcal{T}_B(\widehat{f}_n)}| + \mathbb{P}_n(\widehat{Y}-\widetilde{Y})^2.\end{aligned}$$

Note that $\mathrm{E}\, h_{\mathcal{T}_B(\widehat{f}_n)}$ means to compute $h_{\mathcal{T}_B(f)}$ first and then plug in $f = \widehat{f}_n$. Hence $\mathrm{E}\, h_{\mathcal{T}_B(\widehat{f}_n)}$ is a random variable.

We will consider these three terms separately. The first term can be upper bounded using equation (B.1).

For the second term, we first observe that

$$|h_{f_n}| \le |\{Y - f_n(\boldsymbol{X})\}^2 - \{Y - f_0(\boldsymbol{X})\}^2| = |\{f_n(\boldsymbol{X}) + f_0(\boldsymbol{X}) - 2Y\}\{f_n(\boldsymbol{X}) - f_0(\boldsymbol{X})\}|.$$

Since $\|f_0\|_\infty \le \widetilde{B}$ and $\|f_n\|_\infty \le \widetilde{B}$ for $\widetilde{B} = 2^r B$, we have $Eh_{f_n}^2 \le 16\widetilde{B}^2\,\mathrm{E}\{f_n(\boldsymbol{X}) - f_0(\boldsymbol{X})\}^2 = 16\widetilde{B}^2\,\mathrm{E}\, h_{f_n}$ and $|h_{f_n}| \le 8\widetilde{B}^2$. By Bernstein's inequality (Steinwart and Christmann, 2008, Theorem 6.12), we obtain

$$\Pr\left(|\,\mathbb{P}_n\, h_{f_n} - \mathrm{E}\, h_{f_n}| \ge \frac{16\widetilde{B}^2\tau}{3n} + \left\{\frac{32\widetilde{B}^2\tau(\mathrm{E}\, h_{f_n})}{n}\right\}^{1/2}\right) \le 2e^{-\tau}.$$

Since $2(uv)^{1/2} \le u + v$, we have

$$\left\{ \frac{32\widetilde{B}^2 \tau (\mathrm{E}\, h_{f_n})}{n} \right\}^{1/2} \le \frac{8\widetilde{B}^2 \tau}{n} + \mathrm{E}\, h_{f_n} \le \frac{8\widetilde{B}^2 \tau}{n} + \mathrm{E}\, h_{f_n} + \lambda \|f_n\|_{\mathbb{H}_\gamma}^2.$$

Therefore, we have

$$\Pr\left(|\mathbb{P}_n h_{f_n} - \mathrm{E}\, h_{f_n}| \ge \frac{14\widetilde{B}^2 \tau}{n} + \mathrm{E}\, h_{f_n} + \lambda \|f_n\|_{\mathbb{H}_\gamma}^2 \right) \le 2e^{-\tau}. \tag{B.2}$$

Bounding the third term is a little bit more involved. Fix $s > 0$. For any $f \in \mathbb{H}_\gamma$, we define

$$m_f = \frac{h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}}{\mathcal{E}(f) + s} = \frac{h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}}{\lambda \|f\|_{\mathbb{H}_\gamma}^2 + \mathrm{E}\, h_{\mathcal{T}_B(f)} + s}.$$

Since $\|\mathcal{T}_B(f)\|_\infty \le B$, we have $\|m_f\|_\infty \le 16B^2/s$. Since $\mathrm{E}\, h_{\mathcal{T}_B(f)}^2 \le 16B^2\, \mathrm{E}\, h_{\mathcal{T}_B(f)}$, we have

$$\mathrm{E}\, m_f^2 \le \frac{\mathrm{E}\, h_{\mathcal{T}_B(f)}^2}{4s\, \mathrm{E}\, h_{\mathcal{T}_B(f)}} \le \frac{4B^2}{s},$$

where $\mathrm{E}\, h_{\mathcal{T}_B(f)} > 0$, and $\mathrm{E}\, h_{\mathcal{T}_B(f)}^2 = 0 \le 4B^2/s$ when $\mathrm{E}\, h_{\mathcal{T}_B(f)} = 0$.

Define $\mathcal{F}_s = \{f \in \mathbb{H}_\gamma : \mathcal{E}(f) \le s\} \cup \{0\}$, where 0 denotes the zero function. By Corollary 3, we have

$$\Pr\left\{ \sup_{f \in \mathcal{F}_s} |m_f| \ge 2\, \mathrm{E} \sup_{f \in \mathcal{F}_s} |m_f| + \left(\frac{8B^2\tau}{ns} \right)^{1/2} + \frac{32B^2\tau}{ns} \right\} \le e^{-\tau}.$$

We shall derive an upper bound for $\sup_{f \in \mathcal{F}_s} |m_f|$ based on an upper bound for $\mathrm{E} \sup_{f \in \mathcal{F}_s} |h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}|$. To this end, we compute the covering number for $\mathcal{G}_s = \{h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)} : f \in \mathcal{F}_s\}$.

For any $f \in \mathcal{F}_s$, we have $\|f\|_{\mathbb{H}_{\boldsymbol{\gamma}}} \le s^{1/2}\lambda^{-1/2}$. Hence we have

$$\mathcal{N}(\mathcal{F}_s, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}\{(s^{1/2}\lambda^{-1/2})\mathcal{B}_{\mathbb{H}_{\boldsymbol{\gamma}}}, \|\cdot\|_\infty, \varepsilon\} = \mathcal{N}(\mathcal{B}_{\mathbb{H}_{\boldsymbol{\gamma}}}, \|\cdot\|_\infty, s^{-1/2}\lambda^{1/2}\varepsilon).$$

By the fact that $\|\mathcal{T}_B(f) - \mathcal{T}_B(g)\|_\infty \le \|f - g\|_\infty$, we have

$$\|h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)} - h_{\mathcal{T}_B(g)} + \mathrm{E}\, h_{\mathcal{T}_B(g)}\|_\infty \le 8B\|f - g\|_\infty.$$

Hence we have $\mathcal{N}(\mathcal{G}_s, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}\{\mathcal{F}_s, \|\cdot\|_\infty, \varepsilon/(8B)\}$. Combining these inequalities and applying Lemma 10, we have

$$\log \mathcal{N}(\mathcal{G}_s, \|\cdot\|_\infty, \varepsilon) \le \log \mathcal{N}\{\mathcal{B}_{\mathbb{H}_{\boldsymbol{\gamma}}}, \|\cdot\|_\infty, (8B)^{-1}s^{-1/2}\lambda^{1/2}\varepsilon\} \le c_1 a_{\boldsymbol{\gamma}} s^{d/(2m)}\lambda^{-d/(2m)}\varepsilon^{-d/m},$$

where $m \ge 1$ is an arbitrary integer, c_1 is a constant that depends on m, d, B, r only, and $a_{\boldsymbol{\gamma}} = \prod_{j=1}^d (1 + \gamma_j)^{1/2} \le \overline{\gamma}_{\mathcal{S}}^{|\mathcal{S}|/2}\overline{\gamma}_{\mathcal{S}^c}^{|\mathcal{S}^c|/2}$.

For any $f \in \mathcal{F}_s$, we have $\|h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}\|_\infty \le 16B^2$ and $\operatorname{Var} h_{\mathcal{T}_B(f)} \le \mathrm{E}\, h^2_{\mathcal{T}_B(f)} \le 16B^2 s$. We apply Proposition 5 and obtain

$$\mathrm{E} \sup_{f \in \mathcal{F}_s} |h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}| \le 1024(16B^2 J/n) + 64(16B^2 Js/n)^{1/2},$$

where $J = \int_0^1 c_1 a_{\boldsymbol{\gamma}} (16B^2)^{d/(2m)}\lambda^{-d/(2m)}\varepsilon^{-d/m}\, d\varepsilon \le c_2 a_{\boldsymbol{\gamma}}\lambda^{-d/(2m)}$. Hence we have

$$\mathrm{E} \sup_{f \in \mathcal{F}_s} |h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}| \le c_3 \left\{ a_{\boldsymbol{\gamma}}\lambda^{-d/(2m)}n^{-1} + a_{\boldsymbol{\gamma}}^{1/2}\lambda^{-d/(4m)}s^{1/2}n^{-1/2} \right\}.$$

Hence, by the peeling technique (Steinwart and Christmann, 2008, Theorem 7.7), we

obtain

$$\mathrm{E}\sup_{f\in\mathcal{F}}|m_f| \le 4c_3\{a_{\boldsymbol{\gamma}}\lambda^{-d/(2m)}s^{-1}n^{-1} + a_{\boldsymbol{\gamma}}^{1/2}\lambda^{-d/(4m)}s^{-1/2}n^{-1/2}\}.$$

Combining the bound of $\mathrm{E}\sup_{f\in\mathcal{F}}|m_f|$ and the tail bound of $\sup_{f\in\mathcal{F}}|m_f|$, we have

$$\Pr\left[\sup_{f\in\mathcal{F}}\frac{|h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}|}{\mathcal{E}(f)+s} \ge c_4\left\{\frac{a_{\boldsymbol{\gamma}}}{\lambda^{d/(2m)}sn} + \frac{a_{\boldsymbol{\gamma}}^{1/2}}{\lambda^{d/(4m)}s^{1/2}n^{1/2}} + \frac{\tau^{1/2}}{s^{1/2}n^{1/2}} + \frac{\tau}{sn}\right\}\right] \le e^{-\tau},$$

where $c_4 > 0$ is some constant that depends on m, d, B, r only. Without loss of generality, we assume $c_4 \ge 1$.

Let

$$s = 64c_4^2\max\left\{\frac{a_{\boldsymbol{\gamma}}}{\lambda^{d/(2m)}n}, \frac{\tau}{n}\right\}.$$

Then we have

$$\frac{c_4^2 a_{\boldsymbol{\gamma}}}{\lambda^{d/(2m)}sn} \le \left(\frac{c_4^2 a_{\boldsymbol{\gamma}}}{\lambda^{d/(2m)}sn}\right)^{1/2} \le \frac{1}{8}, \quad \frac{c_4^2\tau}{sn} \le \left(\frac{c_4^2\tau}{sn}\right)^{1/2} \le \frac{1}{8}.$$

Therefore, we have

$$\Pr\left\{|h_{\mathcal{T}_B(f)} - \mathrm{E}\, h_{\mathcal{T}_B(f)}| \ge \mathcal{E}(f)/2 + s/2 \text{ for some } f\in\mathcal{F}\right\} \le e^{\tau}. \tag{B.3}$$

We plug in $f = \widehat{f}_n$ in equation (B.3). Then we combine equations (B.1), (B.2), (B.3) and the condition on $\mathbb{P}_n(\widehat{Y} - \widetilde{Y})^2$. We obtain

$$\Pr\left\{\mathcal{E}(\widehat{f}_n) \le c_6\left(\lambda\overline{\gamma}_{\mathcal{S}}^{|\mathcal{S}|/2}\overline{\gamma}_{\mathcal{S}^c}^{|\mathcal{S}^c|/2} + \underline{\gamma}_{\mathcal{S}}^{-r} + \overline{\gamma}_{\mathcal{S}}^{|\mathcal{S}|/2}\overline{\gamma}_{\mathcal{S}^c}^{|\mathcal{S}^c|/2}\lambda^{-d/(2m)}n^{-1} + n^{-1}\tau\right)\right\} \le e^{-\tau}.$$

Since m can be arbitrarily large, $\delta = d/(2m)$ can be arbitrarily small.

Finally, since

$$\mathrm{E}_{\boldsymbol{X}}\, I(A=a)\left\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\right\}^2 \le \mathcal{E}\left(\widehat{f}_n\right).$$

we observe that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{X}}\, I(A=a)\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\}^2 &= \mathrm{E}_{\boldsymbol{X}}\Pr(A=a|\boldsymbol{X})\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\}^2 \\ &\ge \varpi\,\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\right\}^2.\end{aligned}$$

□

We immediately obtain the following corollaries.

Corollary 14. *Suppose* $\overline{\gamma}_{\mathcal{S}} = \overline{\theta}_{\mathcal{S}} n^{2/(2r+d)}$, $\underline{\gamma}_{\mathcal{S}} = \underline{\theta}_{\mathcal{S}} n^{2/(2r+d)}$, $\overline{\gamma}_{\mathcal{S}^c} = \overline{\theta}_{\mathcal{S}^c} n^{2/(2r+d)}$, *and* $\lambda = \theta_\lambda n^{-1}$. *Then we have*

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\right\}^2 \ge c\left\{n^{-2r/(2r+d)+\delta}+n^{-1}\tau\right\}\right] \le e^{-\tau},$$

Corollary 15. *Suppose* $\overline{\gamma}_{\mathcal{S}} = \overline{\theta}_{\mathcal{S}} n^{2/(2r+|\mathcal{S}|)}$, $\underline{\gamma}_{\mathcal{S}} = \underline{\theta}_{\mathcal{S}} n^{2/(2r+|\mathcal{S}|)}$, $\overline{\gamma}_{\mathcal{S}^c} = \overline{\theta}_{\mathcal{S}^c}$, *and* $\lambda = \theta_\lambda n^{-1}$ *for some* $\xi \ge 0$. *Then we have*

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{f}_n(\boldsymbol{X})-f_0(\boldsymbol{X})\right\}^2 \ge c\left\{n^{-2r/(2r+|\mathcal{S}|)+\delta}+n^{-1}\tau\right\}\right] \le e^{-\tau}.$$

B.1.6 Useful inequalities for the analysis of decision lists

Define $U_t(\boldsymbol{x}, a) = \max_{a' \in \mathcal{A}_t} Q_t(\boldsymbol{x}, a') - Q_t(\boldsymbol{x}, a)$ and $\widehat{U}_t(\boldsymbol{x}, a) = \max_{a' \in \mathcal{A}_t} \widehat{Q}_t(\boldsymbol{x}, a') - Q_t(\boldsymbol{x}, a)$. Since

$$|\max_{a' \in \mathcal{A}_t} \widehat{Q}_t(\boldsymbol{x}, a') - \max_{a' \in \mathcal{A}_t} Q_t(\boldsymbol{x}, a)| \leq \max_{a' \in \mathcal{A}_t} |\widehat{Q}_t(\boldsymbol{x}, a') - Q_t(\boldsymbol{x}, a)|,$$

we have

$$|\widehat{U}_t(\boldsymbol{x}, a) - U_t(\boldsymbol{x}, a)| \leq 2 \max_{a' \in \mathcal{A}_t} |\widehat{Q}_t(\boldsymbol{x}, a') - Q_t(\boldsymbol{x}, a')|$$

. Thus we have

$$\left\{\widehat{U}_t(\boldsymbol{x}, a) - U_t(\boldsymbol{x}, a)\right\}^2 \leq 4 \sum_{a' \in \mathcal{A}_t} \left\{\widehat{Q}_t(\boldsymbol{x}, a') - Q_t(\boldsymbol{x}, a')\right\}^2. \tag{B.4}$$

Following the notations used in the algorithm description, define

$$\widehat{\Omega}_{t\ell}(R, a) = I(\boldsymbol{X}_t \in \widehat{G}_{t\ell}, \boldsymbol{X}_t \in R) \left\{\widehat{U}_t(\boldsymbol{X}_t, a) - \zeta\right\} - \eta \left\{2 - V(R)\right\}$$

and

$$\Omega_{t\ell}(R, a) = I(\boldsymbol{X}_t \in G^*_{t\ell}, \boldsymbol{X}_t \in R) \left\{U_t(\boldsymbol{X}_t, a) - \zeta\right\} - \eta \left\{2 - V(R)\right\}.$$

By the definition of $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell})$ in the main article, we have

$$\begin{aligned}(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \underset{R \in \mathcal{R}_t, a \in \mathcal{A}_t}{\arg\max}\ & \mathbb{P}_n\, I(\boldsymbol{X}_t \in \widehat{G}_{t\ell}) \widehat{Q}_t\{\boldsymbol{X}_t, \widehat{\pi}_t^Q(\boldsymbol{X}_t)\} \\ & - \mathbb{P}_n\, I(\boldsymbol{X}_t \in \widehat{G}_{t\ell}, \boldsymbol{X}_t \in R) \widehat{U}_t(\boldsymbol{X}_t, a) \\ & + \mathbb{P}_n\, \zeta I\{\boldsymbol{X}_t \in \widehat{G}_{t\ell}, \boldsymbol{X}_t \in R\} + \eta\{2 - V(R)\}.\end{aligned}$$

Thus we have $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \arg\max_{R\in\mathcal{R}_t, a\in\mathcal{A}_t} \mathbb{P}_n \widehat{\Omega}_{t\ell}(R,a)$. Similarly, swe have

$$(R^*_{t\ell}, a^*_{t\ell}) = \operatorname*{arg\,max}_{R\in\mathcal{R}_t, a\in\mathcal{A}_t} \Psi_{t\ell}(R,a) = \operatorname*{arg\,max}_{R\in\mathcal{R}_t, a\in\mathcal{A}_t} \mathrm{E}\,\Omega_{t\ell}(R,a).$$

Recall that $\mathcal{R}_t$ consists of rectangles in $\mathbb{R}^d$ defined using at most two variables. Hence $\mathcal{R}_t$ is a subset of the set of all intervals $\{(\boldsymbol{a},\boldsymbol{b}] : \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d\}$, where $(\boldsymbol{a},\boldsymbol{b}] = \{\boldsymbol{x}\in\mathbb{R}^d : a_j \le x_j \le b_j \text{ for all } j\}$. Hence $\mathcal{R}_t$ is a Vapnik-Cervonenkis class, or VC class for short (van der Vaart and Wellner, 1996, Example 2.6.1).

The following lemma gives an upper bound for $\sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \widehat{\Omega}_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)|$ for any given $a\in\mathcal{A}_t$.

Lemma 16. *We have*

$$\begin{aligned}\sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \widehat{\Omega}_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)| &\le \sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \Omega_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)| \\ &\quad + \left[\mathbb{P}_n\left\{\widehat{U}_t(\boldsymbol{X}_t,a) - U_t(\boldsymbol{X}_t,a)\right\}^2\right]^{1/2} + B\sum_{k<\ell}\mathbb{P}_n\, I\left(\boldsymbol{X}_t \in \widehat{R}_{tk} \triangle R^*_{tk}\right),\end{aligned}$$

and

$$\Pr\left\{\sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \Omega_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)| \ge c(n^{-1/2} + \tau^{1/2} n^{-1/2})\right\} \le e^{-\tau}.$$

Proof. We have

$$\begin{aligned}&\sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \widehat{\Omega}_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)| \\ &\qquad \le \sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \widehat{\Omega}_{t\ell}(R,a) - \mathbb{P}_n \Omega_{t\ell}(R,a)| + \sup_{R\in\mathcal{R}_t} |\,\mathbb{P}_n \Omega_{t\ell}(R,a) - \mathrm{E}\,\Omega_{t\ell}(R,a)|.\end{aligned}$$

For the first term, we observe that

$$\begin{aligned}
&\sup_{R\in\mathcal{R}_t}\left|\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R,a)-\mathbb{P}_n\,\Omega_{t\ell}(R,a)\right| \\
&\quad\le \sup_{R\in\mathcal{R}_t}\left|\mathbb{P}_n\,I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\cap G^*_{t\ell},\boldsymbol{X}_t\in R)\left\{\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right\}\right| \\
&\qquad+\sup_{R\in\mathcal{R}_t}\mathbb{P}_n\,I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\triangle G^*_{t\ell},\boldsymbol{X}_t\in R)\left|\widehat{U}_t(\boldsymbol{X}_t,a)-\zeta\right| \\
&\quad\le \mathbb{P}_n\left|\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right|+B\,\mathbb{P}_n\,I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\triangle G^*_{t\ell}),
\end{aligned}$$

By Jensen's inequality, we have

$$\mathbb{P}_n\left|\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right|\le\left[\mathbb{P}_n\left\{\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right\}^2\right]^{1/2}.$$

By the definition of $G_{t\ell}$, we have

$$\mathbb{P}_n\,I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\triangle G^*_{t\ell})\le\sum_{k<\ell}\mathbb{P}_n\,I(\boldsymbol{X}_t\in\widehat{R}_{tk}\triangle G^*_{tk}).$$

For the second term, by the property of VC class (van der Vaart and Wellner, 1996, Lemma 2.6.18), the set

$$\mathcal{F}=\{I(\boldsymbol{X}_t\in\mathbb{R})I(\boldsymbol{X}_t\in G^*_{t\ell})\{U_t(\boldsymbol{X}_t,a)-\zeta\}:R\in\mathcal{R}_t\}$$

is also a VC class. Let ν be its VC index. Then, by van der Vaart and Wellner (1996,

Theorem 2.6.7), we have

$$\sup_Q \mathcal{N}(\mathcal{G}, \|\cdot\|_{L^2(Q)}, \varepsilon) \le c_1 \varepsilon^{-2\nu},$$

where Q is any probability measure and c_1 is a constant that depends on ν only. For any $f \in \mathcal{F}$, we have $\|f\|_\infty \le B$. Thus, by Propositions 4 and 6, since $\int_0^1 \log(\varepsilon^{-2\nu}) < \infty$, we have

$$\Pr\left\{\sup_{f\in\mathcal{F}} |\,\mathbb{P}_n f - \mathrm{E}\, f| \ge c\left(\frac{B^2}{n}\right)^{1/2} + c\left(\frac{B^2\tau}{n}\right)^{1/2}\right\} \le e^{-\tau}$$

for any $\tau > 0$, where c is some constant that depends on ν.

□

Recall that $\rho_t(R_1, R_2) = \Pr(\boldsymbol{X}_t \in R_1 \triangle R_2)$. The following lemma gives an upper bound for

$$\sup_{R\in\mathcal{R}_t, \rho_t(R,R^*_{t\ell})\le\delta} \left|\left\{\mathbb{P}_n \widehat{\Omega}_{t\ell}(R, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R, a^*_{t\ell})\right\} - \left\{\mathbb{P}_n \widehat{\Omega}_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell})\right\}\right|.$$

Lemma 17. *We have*

$$\begin{aligned}
&\sup_{R\in\mathcal{R}_t, \rho_t(R,R^*_{t\ell})\le\delta} \left|\left\{\mathbb{P}_n \widehat{\Omega}_{t\ell}(R, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R, a^*_{t\ell})\right\} - \left\{\mathbb{P}_n \widehat{\Omega}_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell})\right\}\right| \\
\le &\sup_{R\in\mathcal{R}_t, \rho_t(R,R^*_{t\ell})\le\delta} \left|\{\mathbb{P}_n \Omega_{t\ell}(R, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R, a^*_{t\ell})\} - \{\mathbb{P}_n \Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell})\}\right| \\
&+ \left\{\sup_{R\in\mathcal{R}_t, \rho_t(R,R^*_{t\ell})\le\delta} \mathbb{P}_n\, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell})\right\}^{1/2} \left[\mathbb{P}_n \left\{\widehat{U}_t(\boldsymbol{X}_t, a) - U_t(\boldsymbol{X}_t, a)\right\}^2\right]^{1/2} \\
&+ B\left\{\sup_{R\in\mathcal{R}_t, \rho_t(R,R^*_{t\ell})\le\delta} \mathbb{P}_n\, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell})\right\}^{1/2} \left\{\sum_{k<\ell} \mathbb{P}_n\, I\left(\boldsymbol{X}_t \in \widehat{R}_{tk} \triangle R^*_{tk}\right)\right\}^{1/2}.
\end{aligned}$$

Besides, denote as J the first term in the right hand side of the equation above. We have

$$\Pr\left\{J \geq c\delta^{1/2-\beta}(n^{-1/2} + n^{-1/2}\tau^{1/2})\right\} \leq e^{-\tau}.$$

In addition, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_t,\rho_t(R,R^*_{t\ell})} \mathbb{P}_n\, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) \geq c\delta^{1-\beta}(1 + n^{-1}\tau)\right\} \leq e^{-\tau}.$$

Proof. We have

$$\begin{aligned}
&\sup_{R\in\mathcal{R}_t,\rho_t(R,R^*_{t\ell})\leq\delta} \left|\left\{\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R, a^*_{t\ell})\right\} - \left\{\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell})\right\}\right| \\
\leq &\sup_{R\in\mathcal{R}_t,\rho_t(R,R^*_{t\ell})\leq\delta} \left|\{\mathbb{P}_n\,\Omega_{t\ell}(R, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R, a^*_{t\ell})\} - \{\mathbb{P}_n\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell}) - \mathrm{E}\,\Omega_{t\ell}(R^*_{t\ell}, a^*_{t\ell})\}\right| \\
&+ \sup_{R\in\mathcal{R}_t,\rho_t(R,R^*_{t\ell})\leq\delta} \left|\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R, a) - \mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R^*_{t\ell}, a) - \mathbb{P}_n\,\Omega_{t\ell}(R, a) + \mathbb{P}_n\,\Omega_{t\ell}(R^*_{t\ell}, a)\right|.
\end{aligned}$$

The first term can be upper bounded using properties of VC class. For any $\delta > 0$, define

$$\begin{aligned}
\mathcal{F}_\delta = \big\{&I(\boldsymbol{X}_t \in R)I(\boldsymbol{X}_t \in G^*_{t\ell})\,\{U_t(\boldsymbol{X}_t, a) - \zeta\} \\
&- I(\boldsymbol{X}_t \in R^*_{t\ell})I(\boldsymbol{X}_t \in G^*_{t\ell})\,\{U_t(\boldsymbol{X}_t, a) - \zeta\} : R \in \mathcal{R}_t, \rho_t(R, R^*_{t\ell}) \leq \delta\big\}.
\end{aligned}$$

Since $\mathcal{R}_t$ is a VC class, $\mathcal{F}_\delta$ is a VC class for any δ. Besides, we have $\sup_Q \mathcal{N}(\mathcal{F}_\delta, \|\cdot\|_{L^2(Q)}, \varepsilon) \leq c_1\varepsilon^{-2\nu}$ for some constants c_1 and ν independent of δ.

For any $f \in \mathcal{F}_\delta$, we have $\|f\|_\infty \leq B$ and $\mathrm{E}\, f^2 \leq B^2\delta$. Thus, by Propositions 1 and 3,

we have

$$\Pr\left[\sup_{f\in\mathcal{F}_\delta}|\,\mathbb{P}_n f-\mathrm{E}\,f|\geq c_2\left\{\frac{\delta^{1/2}\log^{1/2}(1/\delta)}{n^{1/2}}+\frac{\log(1/\delta)}{n}+\frac{\delta^{1/2}\tau^{1/2}}{n^{1/2}}+\frac{\tau^{1/2}}{n^{1/2}}\right\}\right]\leq e^{-\tau},$$

where c_2 is some constant that depends on B. As $\delta\in(0,1]$, we have $\log(1/\delta)\leq c_3\delta^{-\beta}$ for any $\beta>0$, where c_3 is same constant that depends on β only. Thus, we have

$$\Pr\left\{\sup_{f\in\mathcal{F}_\delta}|\,\mathbb{P}_n f-\mathrm{E}\,f|\geq c_4\delta^{1/2-\beta}\left(\frac{1}{n^{1/2}}+\frac{1}{n\delta^{1/2}}+\frac{\delta^{\beta}\tau^{1/2}}{n^{1/2}}+\frac{\tau}{n\delta^{1/2}}\right)\right\}\leq e^{-\tau}.$$

Hence, when $\delta^{1/2}\geq n^{-1/2}\tau^{1/2}$, we have

$$\Pr\left\{\sup_{f\in\mathcal{F}_\delta}|\,\mathbb{P}_n f-\mathrm{E}\,f|\geq c_5\delta^{1/2-\beta}\left(n^{-1/2}+n^{-1/2}\tau^{1/2}\right)\right\}\leq e^{-\tau}.$$

For the second term, we observe that

$$\begin{aligned}
&\left|\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R,a)-\mathbb{P}_n\,\widehat{\Omega}_{t\ell}(R^*_{t\ell},a)-\mathbb{P}_n\,\Omega_{t\ell}(R,a)+\mathbb{P}_n\,\Omega_{t\ell}(R^*_{t\ell},a)\right|\\
=&\Big|\,\mathbb{P}_n\left\{I(\boldsymbol{X}_t\in R)-I(\boldsymbol{X}_t\in R^*_{t\ell})\right\}I(\boldsymbol{X}_t\in\widehat{G}_{t\ell})\left\{\widehat{U}_t(\boldsymbol{X}_t,a)-\zeta\right\}\\
&-\mathbb{P}_n\left\{I(\boldsymbol{X}_t\in R)-I(\boldsymbol{X}_t\in R^*_{t\ell})\right\}I(\boldsymbol{X}_t\in G^*_{t\ell})\left\{\widehat{U}_t(\boldsymbol{X}_t,a)-\zeta\right\}\Big|\\
\leq&\,\mathbb{P}_n\,I(\boldsymbol{X}_t\in R\triangle R^*_{t\ell})I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\cap G^*_{t\ell})\left|\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right|\\
&+\mathbb{P}_n\,I(\boldsymbol{X}_t\in R\triangle R^*_{t\ell})I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\triangle G^*_{t\ell})B
\end{aligned}$$

Using Cauchy-Schwarz inequality, we have

$$\mathbb{P}_n\,I(\boldsymbol{X}_t\in R\triangle R^*_{t\ell})I(\boldsymbol{X}_t\in\widehat{G}_{t\ell}\cap G^*_{t\ell})\left|\widehat{U}_t(\boldsymbol{X}_t,a)-U_t(\boldsymbol{X}_t,a)\right|$$

$$
\begin{aligned}
&\leq \mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) \left| \widehat{U}_t(\boldsymbol{X}_t, a) - U_t(\boldsymbol{X}_t, a) \right| \\
&\leq \{\mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell})\}^{1/2} \left[\mathbb{P}_n \left\{ \widehat{U}_t(\boldsymbol{X}_t, a) - U_t(\boldsymbol{X}_t, a) \right\}^2 \right]^{1/2},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) I(\boldsymbol{X}_t \in \widehat{G}_{t\ell} \triangle G^*_{t\ell}) B \\
\leq & B \, \{\mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell})\}^{1/2} \left\{ \mathbb{P}_n \, I(\boldsymbol{X}_t \in \widehat{G}_{t\ell} \triangle G^*_{t\ell}) \right\}^{1/2}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\sup_{R \in \mathcal{R}_t, \rho_t(R, R^*_{t\ell})} \left| \mathbb{P}_n \, \widehat{\Omega}_{t\ell}(R, a) - \mathbb{P}_n \, \widehat{\Omega}_{t\ell}(R^*_{t\ell}, a) - \mathbb{P}_n \, \Omega_{t\ell}(R, a) + \mathbb{P}_n \, \Omega_{t\ell}(R^*_{t\ell}, a) \right| \\
\leq & \left\{ \sup_{R \in \mathcal{R}_t, \rho_t(R, R^*_{t\ell})} \mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) \right\}^{1/2} \left[\mathbb{P}_n \left\{ \widehat{U}_t(\boldsymbol{X}_t, a) - U_t(\boldsymbol{X}_t, a) \right\}^2 \right]^{1/2} \\
&+ B \left\{ \sup_{R \in \mathcal{R}_t, \rho_t(R, R^*_{t\ell})} \mathbb{P}_n \, I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) \right\}^{1/2} \left\{ \mathbb{P}_n \, I\left(\boldsymbol{X}_t \in \widehat{G}_{t\ell} \triangle G^*_{t\ell} \right) \right\}^{1/2}.
\end{aligned}
$$

Finally, denote $\mathcal{G}_\delta = \{I(\boldsymbol{X}_t \in R \triangle R^*_{t\ell}) : R \in \mathcal{R}_t, \rho_t(R, R^*_{t\ell}) \leq \delta\}$. Then for any $g \in \mathcal{G}_\delta$, we have $\|g\|_\infty \leq 1$ and $\mathrm{E}\, g^2 \leq \delta$. Thus, by Propositions 1 and 3, we have

$$
\Pr \left[\sup_{g \in \mathcal{G}_\delta} |\, \mathbb{P}_n \, g - \mathrm{E}\, g| \geq c_6 \left\{ \frac{\delta^{1/2} \log^{1/2}(1/\delta)}{n^{1/2}} + \frac{\log(1/\delta)}{n} + \frac{\delta^{1/2} \tau^{1/2}}{n^{1/2}} + \frac{\tau}{n} \right\} \right] \leq e^{-\tau}
$$

Since $\sup_{g \in \mathcal{G}_\delta} \mathrm{E}\, g \leq \delta$ and $(\delta/n)^{1/2} \leq (\delta + 1/n)/2$, we have

$$
\Pr \left\{ \sup_{g \in \mathcal{G}_\delta} \mathbb{P}_n \, g \geq c_7 \delta^{1-\beta} \left(1 + \frac{\tau}{n} \right) \right\} \leq e^{-\tau}.
$$

□

The following lemma is useful for establishing the rate of convergence.

Lemma 18. *Let* $\{M_n(\theta) : \theta \in \Theta\}$ *be a stochastic process and* $M(\theta)$ *a deterministic function. Suppose* $M(\theta) - M(\theta_0) \leq -\kappa d^2(\theta, \theta_0)$ *for some non-negative function* $d : \Theta \times \Theta \to \mathbb{R}$ *and positive number* κ*. Let* c_0 *be some value that may depends on* n*. Suppose when* $\eta \geq c_0$*, we have*

$$\Pr\left\{\sup_{\theta: d(\theta,\theta_0)\leq\delta} |(M_\theta - M)(\theta) - (M_n - M)(\theta_0)| \geq c_1\delta^\xi\tau^{1/2}\right\} \leq e^{-\tau},$$

where $\xi \in (0, 1]$*,* c_1 *is a constant independent of* τ*.*

Let $\widehat{\theta}_n = \arg\max_{\theta\in\Theta} M_n(\theta)$*. Define*

$$\eta = \max\left\{4\kappa^{-1/(2-\xi)}c_1^{1/(2-\xi)}\tau^{1/(4-2\xi)}, c_0\right\}.$$

Then we have

$$\Pr\left\{d(\widehat{\theta}_n, \theta_0) \geq \eta\right\} \leq 3e^{-\tau}.$$

Proof. Fix $\eta > 0$, define $\eta_j = \eta 2^{-j}$, $j \geq 0$. We have

$$\Pr\left\{d(\widehat{\theta}_n - \theta_0) \geq \eta\right\} \leq \sum_{j=1}^{\infty} \Pr\left[\sup_{\theta:\eta_{j-1}\leq d(\theta,\theta_0)<\eta_j} \{M_n(\theta) - M_n(\theta_0)\} \geq 0\right].$$

We observe that

$$M_n(\theta) - M_n(\theta_0) = \{(M_n - M)(\theta) - (M_n - M)(\theta_0)\} + \{M(\theta) - M(\theta_0)\}$$

$$\leq |(M_n - M)(\theta) - (M_n - M)(\theta_0)| - \kappa d^2(\theta, \theta_0).$$

Hence, we have

$$\Pr\left[\sup_{\theta:\eta_{j-1}\leq d(\theta,\theta_0)<\eta_j} \{M_n(\theta) - M_n(\theta_0)\} \geq 0\right] \leq \Pr\left\{\sup_{\theta:d(\theta,\theta_0)\leq\eta_j} |(M_n - M)(\theta) - (M_n - M)(\theta_0)| \geq \kappa\eta_{j-1}^2\right\}$$

Let $\beta = 1/(2-\xi)$. Then $\eta = 4\kappa^{-\beta}c_1^{\beta}\tau^{\beta/2}$. Hence $\eta^{2-\xi} \geq 4\kappa^{-1}c_1\tau^{1/2}$. Since $j2^{-j} \leq 1$, $j \geq j^{1/2} \geq 1$ for all $j \geq 1$ and $\xi - 2 \leq -1$, we have

$$\eta^{2-\xi} \geq \kappa^{-1}2^{-j+2}jc_1\tau^{1/2} \leq \kappa^{-1}2^{j(\xi-2)+2}c_1 j^{1/2}\tau^{1/2}$$

That is, $\kappa\eta^2 2^{2j-2} \geq \eta^{\xi}2^{j\xi}c_1 j^{1/2}\tau^{1/2}$. By the definition of η_j and η_{j-1}, we have $\kappa\eta_{j-1}^2 \geq \eta_j^{\xi}c_1 j^{1/2}\tau^{1/2}$. By the condition on $M_n - M$, we have

$$\Pr\left\{\sup_{\theta:d(\theta,\theta_0)\leq\eta_j} |(M_n - M)(\theta) - (M_n - M)(\theta_0)| \geq \kappa\eta_{j-1}^2\right\} \leq e^{-j\tau}.$$

Therefore, we have $\Pr\left\{d(\widehat{\theta}_n, \theta_0) \geq \eta\right\} \leq \sum_{j=1}^{\infty} e^{-j\tau} = e^{-\tau}/(1 - e^{-\tau})$. Note that $e^{-\tau}/(1 - e^{-\tau}) \leq 3e^{-\tau}$ when $\tau \geq 1$ and $\Pr\left\{d(\widehat{\theta}_n, \theta_0) \geq \eta\right\} \leq 1 \leq 3e^{-\tau}$ when $\tau < 1$.

□

B.1.7 Proof of Theorem 1

In this subsection, ξ and β denote arbitrary positive numbers. The value of ξ or β may be different at each occurrence. We start from the last stage $t = T$. Define $\varphi_T = r_T/(2r_T + d_T)$. Since $\widehat{Y}_T = \widetilde{Y}_T$ for any $a \in \mathcal{A}_T$, under the conditions on $\boldsymbol{\gamma}_T$ and λ_T, by

Proposition 13 and its corollary, we have

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{Q}_T(\boldsymbol{X},a)-Q_T(\boldsymbol{X},a)\right\}^2 \geq c_1\left(n^{-2\varphi_T+\xi}+n^{-1}\tau\right)\right] \leq e^{-\tau}.$$

This establishes the consistency and convergence rate for $\widehat{Q}_T$.

Next, we consider $(\widehat{R}_{T\ell}, \widehat{a}_{T\ell})$ for $\ell = 1, 2, \cdots$. When $\ell = 1$, we have $\widehat{G}_{T1} = G^*_{T1} = \mathcal{X}_T$. Thus, for any $a \in \mathcal{A}_T$, by equation (B.4) and Lemma 16, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T} |\,\mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a) - \mathrm{E}\,\Omega_{T1}(R,a)| \geq c_1 n^{-\varphi_T+\xi}\tau\right\} \leq e^{-\tau}.$$

By Assumption 4 (iii), we have $\inf_{R\in\mathcal{R}_T, a\neq a^*_{T1}} \mathrm{E}\,\Omega_{T1}(R,a) \geq \mathrm{E}\,\Omega_{T1}(R^*_{T1}, a^*_{T1}) + \varsigma$. Thus we have

$$\begin{aligned}\Pr(\widehat{a}_{T1} \neq a^*_{T1}) &\leq \sum_{a\neq a^*_{T1}} \Pr\left\{\sup_{R\in\mathcal{R}_T} \mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a) \geq \mathbb{P}_n\,\widehat{\Omega}_{T1}(R^*_{T1}, a^*_{T1})\right\} \\ &\leq \sum_{a} \Pr\left\{\sup_{R\in\mathcal{R}_T} \left|\mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a) - \mathrm{E}\,\Omega_{T1}(R,a)\right| \geq \varsigma/2\right\}\end{aligned}$$

Hence we obtain

$$\Pr(\widehat{a}_{T1} \neq a^*_{T1}) \leq c_1 \exp(-c_2 n^{\varphi_T-\xi}),$$

where c_1 depends on $|\mathcal{A}_T|$ and c_2 depends on ς. Actually, as seen from the proof of Theorem 2, we are able to obtain a faster convergence rate for $\widehat{a}_{T1}$. However, this won't affect the final result because $\widehat{R}_{T1}$ converges at a much slower rate, as shown below.

We proceed to establish the convergence rate for $\widehat{R}_{T1}$. For any $\delta > 0$, by equation (B.4)

and Lemma 17, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T,\rho_T(R,R^*_{T1})\le\delta}\left|\mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a^*_{T1})-\mathbb{P}_n\,\widehat{\Omega}_{T1}(R^*_{T1},a^*_{T1})-\mathrm{E}\,\Omega_{T1}(R,a^*_{T1})+\mathrm{E}\,\Omega_{T1}(R^*_{T1},a^*_{T1})\right|\right.$$
$$\left.\ge c_1\delta^{1/2-\beta}n^{-\varphi_T+\xi}\tau\right\}\le e^{-\tau}.$$

Hence, by Lemma 18, we have

$$\Pr\left\{\rho_T(\widehat{R}_{T1},R^*_{T1})\ge c_1 n^{-(2/3)\varphi_T+\xi}\tau\right\}\le c_2e^{-\tau}.$$

Note that we take β sufficiently small so that it can be absorbed into ξ.

Next we proceed to $\ell=2$. By equation (B.4) and Lemma 16, for any $a\in\mathcal{A}_T$, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T}|\,\mathbb{P}_n\,\widehat{\Omega}_{T2}(R,a)-\mathrm{E}\,\Omega_{T2}(R,a)|\ge c_1n^{-(2/3)\varphi_T+\xi}\tau\right\}\le e^{-\tau}.$$

Similar to $\widehat{a}_{T1}$, we obtain

$$\Pr(\widehat{a}_{T2}\ne a^*_{T2})\le c_1\exp\left\{-c_2n^{(2/3)\varphi_T-\xi}\right\}.$$

By equation (B.4) and Lemma 17, for any $\delta>0$, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T,\rho_T(R,R^*_{T2})\le\delta}\left|\mathbb{P}_n\,\widehat{\Omega}_{T2}(R,a^*_{T2})-\mathbb{P}_n\,\widehat{\Omega}_{T2}(R^*_{T2},a^*_{T2})-\mathrm{E}\,\Omega_{T2}(R,a^*_{T2})+\mathrm{E}\,\Omega_{T2}(R^*_{T2},a^*_{T2})\right|\right.$$
$$\left.\ge c_1\delta^{1/2-\beta}n^{-(2/3)\varphi_T+\xi}\tau\right\}\le e^{-\tau}.$$

Hence, by Lemma 18, we have

$$\Pr\left\{\rho_T(\widehat{R}_{T2}, R^*_{T2}) \geq c_1 n^{-(2/3)^2\varphi_T+\xi}\tau\right\} \leq c_2 e^{-\tau}.$$

Again, β is chosen to be sufficiently small and is absorbed into ξ.

Using induction, for any ℓ, we obtain

$$\Pr(\widehat{a}_{T\ell} \neq a^*_{T\ell}) \leq c_1 \exp\left\{-c_2 n^{(2/3)^{\ell-1}\varphi_T}\right\}$$

and

$$\Pr\left\{\rho_T(\widehat{R}_{T\ell}, R^*_{T\ell}) \geq c_1 n^{-(2/3)^{\ell}\varphi_T}\tau\right\} \leq c_2 e^{-\tau}.$$

Make a change of variable $\tau \to c_1 n^{-(2/3)^{\ell}\varphi_T}\tau$, we obtain

$$\Pr\{\rho_T(\widehat{R}_{T\ell}, R^*_{T\ell}) \geq \tau\} \leq c_1 \exp\left\{-c_2 n^{(2/3)^{\ell}\varphi_T}\right\}.$$

Therefore, we have

$$\begin{aligned}\Pr\left\{M_T(\widehat{\pi}_T) \geq \tau\right\} &\leq \sum_{\ell=1}^{L^*_T} \Pr\left(\widehat{a}_{T\ell} \neq a^*_{T\ell}\right) + \sum_{\ell=1}^{L^*_T} \Pr\left\{\rho_T(\widehat{R}_{T\ell}, R^*_{T\ell}) \geq \tau/L^*_T\right\} \\ &\leq c_1 \exp(-c_2 n^{\phi_T-\xi}\tau),\end{aligned}$$

where $\phi_T = (2/3)^{L^*_T}\varphi_T$. Consequently, we have

$$\Pr\left\{V_T(\pi^*_T) - V_T(\widehat{\pi}_T) \geq \tau\right\} \leq \Pr\left\{M_T(\widehat{\pi}_T) \geq \tau/B\right\} \leq c_3 \exp(-c_4 n^{\phi_T-\xi}\tau).$$

Next, we proceed to the earlier stages. We consider the $(T-1)$th stage. By the risk bounds of $\widehat{Q}_T$ and $\widehat{\pi}_T$, we have

$$\Pr\left\{\mathbb{P}_n\left(\widehat{Y}_T - \widetilde{Y}_T\right)^2 \geq c_1 n^{-\phi_T+\xi}\tau\right\} \leq c_2 e^{-\tau}.$$

Hence, by Proposition 13, for any $a \in \mathcal{A}_{T-1}$, we have

$$\Pr\left[E_{\boldsymbol{X}}\left\{\widehat{Q}_{T-1}(\boldsymbol{X},a) - Q_{T-1}(\boldsymbol{X},a)\right\}^2 \geq c_1 n^{-\varphi_{T-1}+\xi}\tau\right] \leq c_2 e^{-\tau},$$

where $\varphi_{T-1} = \min\{\phi_T, r_{T-1}/(2r_{T-1}+d_{T-1})\}$. Namely, the convergence rate of $\widehat{Q}_{T-1}$ depends on the kernel regression convergence rate assuming the true response $\widetilde{Y}$ is observed and the convergence rate of the surrogate response $\widehat{Y}$ to its truth.

The analysis of $(\widehat{R}_{T-1,\ell}, \widehat{a}_{T-1,\ell})$s are the same as in the last stage. Hence we have

$$\Pr\left\{M_{T-1}(\widehat{\pi}_{T-1}) \geq \tau\right\} \leq c_1 \exp(-c_2 n^{\phi_{T-1}-\xi}\tau),$$

and

$$\Pr\left\{V_{T-1}(\pi^*_{T-1}) - V_{T-1}(\widehat{\pi}_{T-1}) \geq \tau\right\} \leq c_3 \exp(-c_4 n^{\phi_{T-1}-\xi}\tau),$$

where $\phi_{T-1} = (2/3)^{L^*_{T-1}}\varphi_{T-1}$. Using induction, these two inequalities hold when $T-1$ is replaced by every $t = T-2, \ldots, 1$.

B.1.8 Proof of Theorem 2

At the last stage, by Proposition 13, we have

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{Q}_T(\boldsymbol{X},a)-Q_T(\boldsymbol{X},a)\right\}^2\geq c_1\left(n^{-\varphi_T+\xi}+n^{-1}\tau\right)\right]\leq e^{-\tau},$$

where $\varphi_T = r_{T-1}/(2r_{T-1}+d_{T-1})$ and $\xi>0$ is arbitrary. By equation (B.4), we have

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{U}_T(\boldsymbol{X},a)-U_T(\boldsymbol{X},a)\right\}^2\geq c_1\left(n^{-\varphi_T+\xi}+n^{-1}\tau\right)\right]\leq e^{-\tau}.$$

Using a similar argument to the proof of Theorem 1, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T}|\,\mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a)-\mathrm{E}\,\Omega_{T1}(R,a)|\geq c_1\left(n^{-\varphi_T+\xi}+n^{-1/2}\tau^{1/2}\right)\right\}\leq e^{-\tau}.$$

and

$$\Pr(\widehat{a}_{T1}\neq a^*_{T1})\leq\sum_a\Pr\left\{\sup_{R\in\mathcal{R}_T}\left|\mathbb{P}_n\,\widehat{\Omega}_{T1}(R,a)-\mathrm{E}\,\Omega_{T1}(R,a)\right|\geq\varsigma/2\right\}.$$

Note that ς is a fixed number independent of n. Let $\tau^{1/2}=n^{1/2}\max(c_2\varsigma-n^{-\varphi_T+\xi},0)$ and choose c_2 such that $2c_1c_2<1$. Then we have

$$\Pr(\widehat{a}_{T1}\neq a^*_{T1})\leq c_3\exp(-c_4n)$$

as $\varphi_T\in(0,1)$.

Define $\vartheta=\inf_{R:\rho_T(R,R^*_{T1})>0}\rho_T(R,R^*_{T1})$. Since the covariates are discrete, ϑ is strictly positive. This is a major difference between the continuous covariates and the discrete

covariates. Then we have

$$\begin{aligned}\Pr\left\{\rho_T(\widehat{R}_{T1}, R^*_{T1}) > 0\right\} &\leq \Pr\left\{\sup_{R\in\mathcal{R}_T:\rho_T(R,R^*_{T1})>0} \mathbb{P}_n\,\widehat{\Omega}_{T1}(R, a^*_{T1}) \geq \mathbb{P}_n\,\widehat{\Omega}_{T1}(R^*_{T1}, a^*_{T1}) + \vartheta\right\} \\ &\leq \Pr\left\{\sup_{R\in\mathcal{R}_T}\left|\mathbb{P}_n\,\widehat{\Omega}_{T1}(R, a^*_{T1}) - \mathrm{E}\,\Omega_{T1}(R, a^*_{T1})\right| \leq \vartheta/2\right\} \\ &\leq c_5\exp(-c_6 n).\end{aligned}$$

We move forward to analyze $(\widehat{R}_{T2}, \widehat{a}_{T2})$. For any $a \in \mathcal{A}_T$, we have

$$\Pr\left\{\sup_{R\in\mathcal{R}_T} |\,\mathbb{P}_n\,\widehat{\Omega}_{T2}(R, a) - \mathrm{E}\,\Omega_{T2}(R, a)| \geq c_1\left(n^{-\varphi_T+\xi} + n^{-1/2}\tau^{1/2}\right)\right\} \leq e^{-\tau}.$$

Similar to $(\widehat{R}_{T1}, \widehat{a}_{T1})$, we have

$$\Pr(\widehat{a}_{T2} \neq a^*_{T2}) \leq c_1\exp(-c_2 n)$$

and

$$\Pr\left\{\rho_T(\widehat{R}_{T2}, R^*_{T2}) > 0\right\} \leq c_3\exp(-c_4 n).$$

As seen from the inequality, a notable difference is that the estimation error does not propagate along the list. The tail probability decays at the same exponential rate for every ℓ. Therefore, we have

$$\Pr\left\{M_T(\widehat{\pi}_T) > 0\right\} \leq \sum_{\ell=1}^{L^*_T}\Pr\left(\widehat{a}_{T\ell} \neq a^*_{T\ell}\right) + \sum_{\ell=1}^{L^*_T}\Pr\left\{\rho_T(\widehat{R}_{T\ell}, R^*_{T\ell}) > 0\right\} \leq c_1\exp(-c_2 n).$$

Thus we have

$$\Pr\left\{V_T(\pi_T^*) - V_T(\widehat{\pi}_T) > 0\right\} \leq \Pr\left\{M(\widehat{\pi}_T) > 0\right\} \leq c_1 \exp(-c_2 n).$$

We then move to the $(T-1)$th stage. The analysis is conditional on the event $\{M(\widehat{\pi}_T) = 0\}$, which occurs with probability $1 - c_1 \exp(-c_2 n)$. When $M(\widehat{\pi}_T) = 0$, we have

$$\Pr\left[\widehat{Q}\left\{\boldsymbol{X}_T, \widehat{\pi}_T(\boldsymbol{X}_T)\right\} = \widehat{Q}\left\{\boldsymbol{X}_T, \pi_T^*(\boldsymbol{X}_T)\right\}\right] = 1.$$

Hence

$$\Pr\left\{\mathbb{P}_n\left(\widehat{Y}_{T-1} - \widetilde{Y}_{T-1}\right)^2 \geq c_1\left(n^{-\varphi_T+\xi} + n^{-1}\tau\right)\right\} \leq e^{-\tau}.$$

Define $\varphi_{T-1} = \min\left\{r_{T-1}/(2r_{T-1} + d_{T-1}), \varphi_T\right\}$. By Proposition 13, we have

$$\Pr\left[\mathrm{E}_{\boldsymbol{X}}\left\{\widehat{Q}_{T-1}(\boldsymbol{X}, a) - Q_{T-1}(\boldsymbol{X}, a)\right\}^2 \geq c_1\left(n^{-\varphi_{T-1}+\xi} + n^{-1}\tau\right)\right] \leq e^{-\tau}.$$

Note that nothing is changed except that T is replaced by $T-1$. Hence, using the same approach as in the Tth stage, we obtain

$$\Pr\left\{M_{T-1}(\widehat{\pi}_{T-1}) > 0\right\} \leq c_1 \exp(-c_2 n),$$

and

$$\Pr\left\{V_{T-1}(\pi_{T-1}^*) - V_{T-1}(\widehat{\pi}_{T-1}) > 0\right\} \leq c_1 \exp(-c_2 n).$$

Using induction, we can establish similar inequalities for $t = T-2, \ldots, 1$.

B.2 Algorithm Details and Proof of Proposition 1

Fix an t and ℓ. Define

$$U_{iat\ell} = \left[\widehat{Q}_t\left\{\boldsymbol{X}_{it}, \widehat{\pi}_t^Q(\boldsymbol{X}_{it})\right\} - \widehat{Q}_t(\boldsymbol{X}_{it}, a) - \zeta\right] I\left(\boldsymbol{X}_{it} \in \widehat{G}_{t\ell}\right).$$

For notation simplicity, we shall omit the subscript t and ℓ and write U_{ia} and $\boldsymbol{X}_i$. By the definition of $(\widehat{R}_{t\ell}, \widehat{a}_{t\ell})$, we have

$$(\widehat{R}_{t\ell}, \widehat{a}_{t\ell}) = \underset{R\in\mathcal{R}_t, a\in\mathcal{A}_t}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} U_{ia} I(\boldsymbol{X}_i \in R) - \eta\{2 - V(R)\}.$$

We will first fix the treatment a and the covariates involved in R, and focus on the computing the optimal thresholds. Then we will loop over all covariate pairs and all treatment options.

Finding the threshold when R involves one covariate Without loss of generality, we assume $R = \{\boldsymbol{x} : x_j \leq \tau\}$. The other situation $R = \{\boldsymbol{x} : x_j > \tau\}$ can be handled similarly. We want to compute

$$\widehat{\tau} = \underset{\tau}{\arg\min} \sum_{i=1}^{n} U_{ita} I(X_{ij} \leq \tau),$$

where X_{ij} is the jth component of $\boldsymbol{X}_i$.

Let $i_1, \ldots, i_n$ be a permutation of $1, \ldots, n$ such that $X_{i_1 j} \leq \cdots \leq X_{i_n j}$. Since the

objective function is piecewise constant, we only need to compute

$$F(\tau) = \sum_{i=1}^{n} U_{ia} I(X_{ij} \leq \tau)$$

when τ equals to some $X_{i_s j}$. We observe that

$$F(X_{i_s j}) = \sum_{h \leq s} U_{i_h a}.$$

Thus it is clear that when $s \geq 2$

$$F(X_{i_s j}) = F(X_{i_{s-1} j}) + U_{i_s a}.$$

Hence, by starting at $s = 1$ and using the recursive relationship, we can compute $F(X_{i_s j})$ for all s and pick the smallest one in $O(n)$ time.

Dealing with ties If $X_{i_s j} = X_{i_{s+1} j}$ for some $s \geq 1$, then $F(X_{i_s j})$ shouldn't be counted when picking the minimum. This is because $F(X_{i_s j})$ hasn't included all subjects with $X_{ij} = X_{i_s j}$ yet.

To avoid this problem, when there are ties, we first aggregate the U_{ia} values for subjects having the same value of X_{ij}. Similar action can be taken when R involves two covariates, in which case the U_{ia} values for subjects having the same value for both covariates are aggregated.

Finding the threshold when R involves two covariates This situation is more complicated. Without loss of generality, we assume $R = \{\boldsymbol{x} : x_j \leq \tau \text{ and } x_k \leq \sigma\}$. We

want to compute

$$(\widehat{\tau}, \widehat{\sigma}) = \arg\min_{\tau,\sigma} \frac{1}{n} \sum_{i=1}^{n} U_{ia} I(X_{ij} \leq \tau, X_{ik} \leq \sigma).$$

We cannot utilize the idea for one covariate, since there is no natural ordering in two-dimensional space. Our solution is to sort in one dimension and to use binary tree for fast lookup and insertion in the other dimension.

We start with constructing a complete binary tree of at least n leaves. The height of such a tree is of order $O(\log_2 n)$.

Let $i_1, \ldots, i_n$ be a permutation of $1, \ldots, n$ such that $X_{i_1 j} \leq \ldots X_{i_n j}$. At each time s, we will insert $U_{i_s a}$ into the binary tree and search for the optimal threshold σ among $X_{ik}, i = 1, \ldots, n$. Note that at time s, values $U_{i_h a}, h \leq s$ are contained in the binary tree. So we are looking at the threshold $\tau = X_{i_s j}$. Specifically, if the rank of $X_{i_s k}$ among X_{ik}s is h, which means $X_{i_s k}$ is the hth smallest among X_{ik}s, then we put $U_{i_s a}$ in the hth leaf from the left in the tree.

In the tree, each node is associated with a subtree in which that node serves as the root. Each node contains two pieces of information. First, it computes the sum of all $U_{i_s a}$s in the associated subtree. Second, it computes the best thresholding sum in the associated subtree, which is the smallest value among the sum of all $U_{i_s a}$s that satisfies $X_{i_s k} \leq \sigma$ for some σ, where σ can take the value of any $X_{i_s k}$ in the associated subtree.

The binary tree structure enables us to update these two pieces of information effectively when a new value, $U_{i_s a}$, is inserted into the tree. We move from the leaf node to its parent, and then its ancestors, and finally the root. At each node, the sum of all $U_{i_s a}$s in the associated subtree is increased by $U_{i_s a}$. As for updating the best thresholding sum,

since the thresholding condition is $X_{i_sk} \leq \sigma$, the best thresholding sum of a node can only be either the best thresholding sum in its left child, or, the sum of all U_{i_sa} values in the left child plus the best thresholding sum in the right child, whichever is smaller.

Since the height of the tree is $O(\log_2 n)$, the updating process involves at most $O(\log_2 n)$ nodes and the time complexity at each node is constant. Therefore, when U_{i_sa} is inserted into the tree, we are able to find the optimal σ that minimizes $\sum_{h \leq s} U_{i_h a} I(X_{i_h k} \leq \sigma)$ in $O(\log_2 n)$ time.

Then we let s run from 1 to n, and find the s that gives the minimum. In this way, we find the minimum of $\sum_{h=1}^{n} U_{i_h a} I(X_{i_h k} \leq \sigma, X_{i_h j} \leq X_{i_s j})$ with respect to σ and s, which is exactly the minimum of $\sum_{i=1}^{n} U_{i_s a} I(X_{ik} \leq \sigma, X_{ij} \leq \tau)$ with respect to σ and τ. And the time complexity for finding both τ and σ is $O(n \log_2 n)$.

Finding the covariate(s) and treatment Up to now, we have discussed how to find the optimal thresholds when the covariates to use X_{ij}, X_{ik} and the treatment a are given. Certainly we need to explore all Rs defined using only one covariate, and all Rs defined using some pair of X_{ij} and X_{ik}. We also need to loop over all treatment options $a \in \mathcal{A}_t$.

Therefore, the overall time complexity is $O(n \log n d_t^2 m_t)$, where d_t is the dimension of $\boldsymbol{X}_i$ and $m_t = |\mathcal{A}_t|$ is the number of available treatment options.

B.3 Variables in Data Analysis

In the first stage, we have the following variables:

1. age: integer;

2. gender: 1 for male, 0 for female;
3. race: 1 for white, 0 for others;
4. education level: 1 for high school or below, 2 for some college, 3 for bachelor or up;
5. work status: 1 for full time, 0.5 for part time, 0 for no work;
6. bipolar type: 1 or 2;
7. status prior to the onset of the current episode: 1 for remission longer than 8 weeks;
8. status prior to the onset of the current episode: 1 for manic/hypomanic;
9. status prior to the onset of the current episode: 1 for mixed/cycling;
10. SUM-D at week 0;
11. SUM-ME at week 0.

In the second stage, we have the following variables:

1. binary indicator for adverse effect tremor;
2. binary indicator for adverse effect dry mouth;
3. binary indicator for adverse effect sedation;
4. binary indicator for adverse effect constipation;
5. binary indicator for adverse effect diarrhea;
6. binary indicator for adverse effect headache;

7. binary indicator for adverse effect poor memory;
8. binary indicator for adverse effect sexual dysfunction;
9. binary indicator for adverse effect increase appetite;
10. SUM-D at week 6;
11. SUM-ME at week 6.