Machine Learning Techniques for Optimal Treatment Regimes

Xin Zhou

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Michael R. Kosorok

Michael G. Hudgens

Yufeng Liu

Jennifer Lund

Donglin Zeng

ABSTRACT

Xin Zhou: Machine Learning Techniques for Optimal Treatment Regimes
(Under the direction of Michael R. Kosorok)

Personalized medicine has received increasing attention among statisticians, computer scientists and clinical practitioners. Patients often show significant heterogeneity in response to treatments. The estimation of optimal treatment regimes is of considerable interest to personalized medicine. In this dissertation, we develop methodology mainly using machine learning techniques to estimate optimal treatment regimes.

In the first part of the dissertation, we apply the k-nearest neighbor (kNN) rule, a simple nonparametric approach, to estimate the optimal treatment regime. We show that the kNN rule is universally consistent, and establish its convergence rate. Since kNN suffers from the curse of dimensionality, we develop an adaptive k-nearest neighbor (AkNN) rule, where the distance metric is adaptively determined from the data, to perform metric selection and variable selection simultaneously. The performance of the proposed methods is illustrated in simulation studies and in an analysis of the Nefazodone-CBASP clinical trial.

In the second part, we point out several weaknesses in outcome weighted learning (OWL), which was proposed recently to construct optimal treatment regimes by directly optimizing the clinical outcome. We then propose a general framework, called residual weighted learning (RWL), to alleviate these problems. RWL weights misclassification errors by residuals of the outcome from a regression fit on clinical covariates excluding treatment assignment. We also propose variable selection methods for linear and nonlinear rules, respectively, to further improve performance. We show that the resulting estimator is consistent, and obtain a rate of convergence. The performance is illustrated in simulation studies and in an analysis of cystic fibrosis clinical trial data.

In the third part, we develop a permutation test for qualitative treatment-covariates interactions. Qualitative interactions arise when the direction of the treatment effect changes among different subsets of subjects. In this work, we estimate the optimal treatment regime by a modified residual weighted learning method, called mirrored residual weighted learning (MRWL), and then apply a permutation test to make inference on the estimated regime. The performance of the proposed permutation test is illustrated in simulation studies and case studies.

ACKNOWLEDGEMENTS

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Michael R. Kosorok, for his patient guidance, constructive comments and strong support to my research work. I have been so fortunate to have an advisor who set an example of a world-class researcher. His encouragements during my PhD studies have helped me walk through the entire process. I will never forget his support when I was in hard time.

I would also like to give sincere thanks to my committee members Dr. Donglin Zeng, Dr. Michael G. Hudgens, Dr. Yufeng Liu and Dr. Jennifer Lund for their inspirational discussions and valuable suggestions for this research work. I am grateful to Dr. Joseph Ibrahim, Dr. Hongtu Zhu, Dr. Pei Fen Kuan, Dr. Bahjat Qaqish and Dr. Michael R. Kosorok for their financial supports during these years.

Special thanks to my friends and schoolmates for their encouragement and kind help. It is my pleasure of being studying and living with them. Their friendship will be an invaluable treasure throughout my life.

Last but not least, none of this would have been made possible without my family. I would like to thank my mother, Qiaoxia Wang, and parents-in-law, Yizhao Wu and Qizhen He, for their continuous support and selfless love. My kids, Jason, Sophie and Ryan, have cheered me along the way. My deepest appreciation goes to my wife, Xingye Wu, who always stands by me through the good times and bad. This dissertation is dedicated to her for her love, support and patience for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

Recently, personalized medicine has received much attention among statisticians, computer scientists, and clinicians. The purpose of personalized medicine is to tailor treatments to individual patient to maximize treatment benefit and safety in health care. There are already some marketed tailored therapies. For example, ceritinib, a newly approved drug by FDA for the treatment of lung cancer, is highly active in patients with advanced, ALK-rearranged non-small-cell lung cancer (Shaw et al. 2014). Patients often show significant heterogeneity in response to treatments. This inherent heterogeneity suggests a transition from the traditional "one size fits all" approach to modern personalized medicine.

The goal of personalized medicine is to provide meaningful improved health outcomes for patients by delivering the right treatment at the right dose at the right time. Personalized treatments can be seen as realizations of certain decision rules, which dictate what to do in a given state of an individual patient. The decision problems may be grouped into two categories: single-stage and multi-stage decision problems. Single-stage decision problems in personalized medicine are often called optimal treatment regimes, while multi-stage problems are usually referred to as dynamic treatment regimes. In single-stage problems, the clinician has to decide on the optimal treatment, from among a set of possible treatments, for an individual patient. The multi-stage decision problems arising in chronic diseases often involve complex choices with multiple stages, where treatments administered at one stage affect those to be administered at a later stage. A dynamic treatment regime is a sequence of decision rules, one per stage of intervention, for adapting a treatment plan over

time to an individual.

In this dissertation, we focus on the single-stage decision problem. It is the building block for the multi-stage decision problem. Although it is a special case of dynamic treatment regimes, it has wider applications. Dynamic treatment regimes are mainly applicable to the chronic diseases, while optimal treatment regimes arise in almost all clinical areas.

1.2 Framework and Assumptions

For each patient, we observe a triplet (X, A, R): individual baseline covariates $X \in \mathcal{X}$, treatment assignment $A \in \mathcal{A}$ and the clinical outcome $R \in \mathbb{R}$ with larger values of R being more desirable. We follow the convention that a random variable is indicated with a capital letter and its realization is represented by a lower-case letter.

A treatment regime d is a function of a patient's covariates X, and outputs a value in $\mathcal{A}$. The goal of optimal treatment regimes is to identify the best treatment regime d^{opt} after observing the corresponding baseline covariates x so that the resulting outcome is optimized.

We then introduce the potential outcomes framework. The notion of *potential outcomes* (also called *counterfactuals*) (Robins 1986) is defined as a patient's outcome if she had followed a particular treatment regime, possibly different from the regime that she was actually observed to follow. Mathematically, potential outcomes are

$$W^* = \{R^*(a) : a \in \mathcal{A}\}, \tag{1.1}$$

where $R^*(a)$ is the outcome if the patient were to have been administered the treatment a. For a treatment regime d, define the potential outcomes associated with d as $\{R^*(d)\}$. An optimal regime represents the "best" way to treat patients. Mathematically, an optimal

regime d^{opt} is the one satisfies

$$\mathbb{E}(R^*(d)|X = x) \leq \mathbb{E}(R^*(d^{opt})|X = x), \quad \forall d \text{ and all } x \in \mathcal{X}. \tag{1.2}$$

Since x is arbitrary, (1.2) is equivalent to finding the regime with $\mathbb{E}(R^*(d)) \leq \mathbb{E}(R^*(d^{opt}))$ for any d. The quantity $\mathcal{V}(d) := \mathbb{E}(R^*(d))$ is called the value function of d.

Potential outcomes for a given patient with a specified d are not observed. Thus we have to estimate d^{opt} in (1.2) using data from a observational study or a clinical trial. The following assumptions are required for estimation.

- Consistency assumption (Robins 1994): the potential outcomes under the observed treatment regime and the observed outcomes agree, *i.e.* $R = R^*(A)$.
- Stable unit treatment value assumption (SUTVA) (Rubin 1978): A patient's outcome is not influenced by other patients' treatment allocation.
- No unmeasured confounders (NUC) (Robins 1997): conditional on covariates, the treatment actually received is independent of potential outcomes. That is, for any regime $a \in \mathcal{A}$,

$$A \perp R^*(a)|X.$$

The first assumption is fundamental for the potential outcomes framework. This assumption requires that the outcome after a given treatment is the same, regardless of the manner in which treatments are 'assigned'. The second assumption is often reasonable when patients are independently drawn from a large population. The third assumption always holds under complete or sequential randomization, and hence is also called the sequential ignorability assumption. This assumption may also hold in observational studies where all relevant confounders have been measured.

There is an additional *positivity assumption* to describe the feasible treatment domain. We simply assume $p(X = x, A = a) > 0$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$. That is, the

participants with covariates x have an innegligible probability of receiving any available treatment. The positivity may be violated either theoretically or practically. A theoretical violation occurs if the study design excludes certain subjects from receiving a particular treatment. For instance, in a sequential multiple assignment randomized trial (SMART) described in (Murphy 2005), the treatment options at the second stage are different for responders and non-responders to the treatments at the first stage. A practical violation of the positivity assumption occurs when a particular group of the subjects has a very low probability of receiving some treatment.

1.3 Optimal treatment regimes

Assuming that the data generating mechanism is known, the best treatment regime is dictated by the Bayes rule. In particular, to fix idea, consider the case where there are two treatment choices, that is $\mathcal{A} = \{1, -1\}$. The potential outcomes are $\{R^*(1), R^*(-1)\}$. According to the consistency assumption, $R = R^*(1)\mathbb{I}(A = 1) + R^*(-1)\mathbb{I}(A = -1)$. The NUC assumption says $A \perp (R^*(1), R^*(-1))|X$. By those assumptions, we can derive the conditional expected potential outcome, under any treatment regime $d : \mathcal{X} \to \mathcal{A}$, as

$$\begin{aligned} & \mathbb{E}(R^*(d)|X = x) \\ = \; & \mathbb{E}\big(R^*(1)|X = x\big)\mathbb{I}\big(d(x) = 1\big) + \mathbb{E}\big(R^*(-1)|X = x\big)\mathbb{I}\big(d(x) = -1\big) \\ = \; & \mathbb{E}\big(R|X = x, A = 1\big)\mathbb{I}\big(d(x) = 1\big) + \mathbb{E}\big(R|X = x, A = -1\big)\mathbb{I}\big(d(x) = -1\big). \end{aligned}$$

Let

$$Q(x, a) := \mathbb{E}(R|X = x, A = a), \quad a \in \mathcal{A}, x \in \mathcal{X},$$

be the conditional expected outcome under a particular treatment. Thus based on the definition of optimal treatment regimes in (1.2), the optimal treatment is given by the following

Bayes rule:

$$d^{opt}(x) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(x, a) = \begin{cases} 1 & \text{if } Q(x,1) > Q(x,-1), \\ -1 & \text{if } Q(x,1) \leq Q(x,-1). \end{cases} \tag{1.3}$$

Naturally, there are three different paradigms to approach the single-stage decision problems. The first paradigm is based on the treatment effect: first estimate the expected outcomes $Q(x, 1)$ and $Q(x, -1)$, and plug the estimates into (1.3) to yield a treatment rule. There is a significant amount of literature developing optimal regimes based on predicting patients' outcomes (Cai et al. 2011, Zhao et al. 2009). Recently, Qian and Murphy (2011) proposed an ℓ_1 penalized least squares (ℓ_1-PLS) approach to estimate the optimal treatment regime. Using a sparse ℓ_1 penalty to model conditional expected outcomes, this method introduces parsimony and facilitates ease of interpretation.

The second paradigm is related to the treatment contrast function $C(x) := Q(x, 1) - Q(x, -1)$. First estimate the treatment contrast function, and assign treatment 1 if $C(x) > 0$, and -1 otherwise. For instance, Tian et al. (2012) developed a simple method using modified covariates to estimate the contrast function directly without the need for modelling the expected outcomes. Coupled with an efficiency augmentation procedure, this method yields clinically meaningful estimators in a variety of settings. The advantage learning (A-learning) also falls in this paradigm (Robins 2004, Murphy 2003, Moodie et al. 2007).

The first two paradigms are regression-based approaches. Their aim is to estimate the contrast function $C(x)$. Given a sample $D_n = \{(X_i, A_i, R_i)_{i=1}^n\}$. Let $C_n(x)$ be an estimate of $C(x)$ on D_n. Its associated treatment regime is $d_n(x) = \operatorname{sign}(C_n(x))$, where $sign(u) = 1$ for $u > 0$ and -1 otherwise. [The particular choice of the value of $sign(0)$ is not important. Here we fix $sign(0) = -1$]. The following theorem builds the foundation of regression-based approaches.

Theorem 1.3.1. *For a sequence $C_n(X)$ of estimates for $C(X)$, we have*

$$\mathcal{V}^{opt} - \mathcal{V}(d_n) \leq \sqrt{\mathbb{E}(C_n(X) - C(X))^2},$$

where $\mathcal{V}^{opt} := \mathcal{V}(d^{opt})$ is the optimal value function.

Proof. For any treatment regime d, we have

$$\begin{aligned}
\mathcal{V}(d) &= \mathbb{E}\left(\frac{R\mathbb{I}(A = d(X))}{P(A|X)}\right) \\
&= \mathbb{E}(\mathbb{E}(R|A = 1, X)\mathbb{I}(d(X) = 1) + \mathbb{E}(R|A = -1, X)\mathbb{I}(d(X) = -1)) \\
&= \mathbb{E}(Q(X, 1)\mathbb{I}(d(X) = 1) + Q(X, -1)\mathbb{I}(d(X) = -1)).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{V}^{opt} - \mathcal{V}(d_n) &= \mathbb{E}((Q(X, 1) - Q(X, -1))(\mathbb{I}(d^{opt}(X) = 1) - \mathbb{I}(d_n(X) = 1))) \\
&= \mathbb{E}(|C(X)|\mathbb{I}(d^{opt}(X) \neq d_n(X))) \\
&\leq \mathbb{E}(|C_n(X) - C(X)|\mathbb{I}(d^{opt}(X) \neq d_n(X))).
\end{aligned}$$

The last inequality is from the construction of d_n and d^{opt}. The desired result follows from the Cauchy-Schwarz inequality. □

Consistent regression estimation leads to consistent treatment selection. But as shown in Figure 1.1, a good treatment decision function $C_n(x)$ does not require to be a good regression estimate of $C(x)$. For the optimal treatment regime problem, it is sufficient for $C_n(x)$ to be close to $C(x)$ only near the zeros of $C(x)$, and elsewhere $C_n(x)$ does not need to be close to $C(x)$. In general, estimating the optimal treatment regime is easier than regression function estimation. Regression-based approaches are not the only way to solve the optimal treatment regime problem.

The third paradigm directly estimates the optimal treatment regime: estimate the set $G := \{x \in \mathcal{X} : C(x) > 0\}$, and select treatment $+1$ if $x \in G$, and -1 otherwise. An

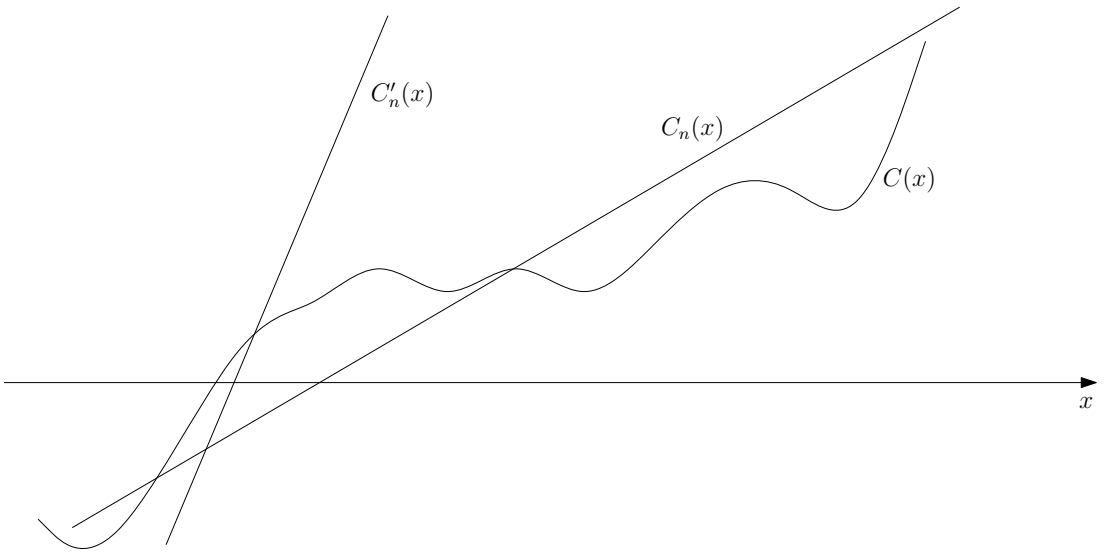


Figure 1.1: *$C_n(x)$ and $C_n'(x)$ are two linear estimates of the contrast function $C(x)$. Let $d_n(x) = \text{sign}(C_n(x))$ and $d_n'(x) = \text{sign}(C_n'(x))$ be their associated treatment selection rules. Though $C_n'(x)$ is a poor estimate of $C(x)$, its associated selection rule is better than that of $C_n(x)$.*

important quantity in this paradigm is the risk function,

$$\mathcal{R}(d) := \mathbb{E}(R|A \neq d(X)) = \mathbb{E}\left(\frac{R}{P(A|X)}\mathbb{I}(A \neq d(X))\right), \tag{1.4}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $P(A|X)$ is the probability of being assigned treatment A for subjects with covariates X. For clinical trial data, the probability $P(A|X)$ is predefined in the trial design. For observational studies, the probability has to be estimated from the observed data. Maximizing the value function is equivalent to minimizing the risk function. Zhao et al. (2012) viewed (1.4) as the weighted error, and proposed outcome weighted learning (OWL), which utilizes the weighted support vector machines to estimate the optimal treatment rule directly. Zhang et al. (2012a) also proposed a general framework to make use of classification methods to the treatment regime problem.

1.4 Outline of the dissertation

In this dissertation, we propose two machine learning methods to finding optimal treatment regimes. Chapter 2 apply the k-nearest neighbor (kNN) rule, a simple nonparametric approach. We show that the kNN rule is universally consistent, and establish its convergence rate. We also develop an adaptive k-nearest neighbor (AkNN) rule to perform metric selection and variable selection.

In Chapter 3, we demonstrate several weaknesses in outcome weighted learning (OWL), and then propose a general framework, called residual weighted learning (RWL), to alleviate these problems. RWL weights misclassification errors by residuals of the outcome from a regression fit on clinical covariates excluding treatment assignments. We also propose variable selection methods for linear and nonlinear rules, respectively. We show that the RWL estimator is consistent, and obtain the convergence rate.

In Chapter 4, we address an important practical issue, detecting qualitative interactions, in the language of optimal treatment regimes. Qualitative interactions arise when the direction of the treatment effect changes among different subsets. In this work, we estimate the optimal treatment regime by a modified residual weighted learning method, called mirrored residual weighted learning (MRWL), and then apply a permutation test to make inference on the estimated regime for We develop a permutation test for qualitative treatment-covariates interactions.

We conclude the dissertation in Chapter 5 with future research work.

CHAPTER 2: NEAREST NEIGHBOR REGIMES

2.1 Introduction

Recently, personalized medicine has received much attention among statisticians, computer scientists and clinicians. The purpose of personalized medicine is to tailor treatments to individual patients to maximize treatment benefit and safety in health care. There are already several marketed tailored therapies. For example, ceritinib, a recently FDA approved drug for the treatment of lung cancer, is highly active in patients with advanced, ALK-rearranged non-small-cell lung cancer (Shaw et al. 2014). Patients often show significant heterogeneity in response to treatments. This inherent heterogeneity suggests a transition from the traditional "one size fits all" approach to modern personalized medicine.

A major component of personalized medicine is the treatment selection rule, or optimal treatment regime. Formally, a treatment regime is a rule that assigns a treatment, from among a set of possible treatments, to a patient based on his or her clinical or genomic characteristics. There is a significant literature on optimal treatment strategies based on data from clinical trials or observational studies (Murphy 2003, Robins 2004, Zhang et al. 2012b, Zhao et al. 2009). Most of these methods model either the conditional mean outcomes or contrasts between mean outcomes. They obtain treatment regimes indirectly by inverting the regression estimates instead of directly optimizing the decision rule. For instance, Qian and Murphy (2011) proposed a two-step procedure that first estimates a conditional mean for the outcome and then determines the treatment regime by comparing conditional means across various treatments. The success of these indirect approaches depends highly on correct specification of posited models and on the precision of model

estimates.

Recently, several researchers have applied classification methods to optimal treatment regimes. For example, Zhao et al. (2012) proposed outcome weighted learning (OWL) to construct a treatment selection rule that directly optimizes the clinical outcome under this rule. They cast the treatment selection problem as a weighted classification problem, and apply state-of-the-art support vector machines to address it. Zhang et al. (2012a) also proposed a general framework to make use of classification methods to treatment regimes.

The k-nearest neighbor rule is a simple and intuitively appealing classification approach, where a subject is classified by a majority vote of its neighbors. Since its conception in 1951 (Fix and Hodges 1951), it has attracted many researchers, and retains its popularity today (Cover and Hart 1967, Stone 1977, Györfi 1981, Devroye and Györfi 1985, Hastie and Tibshirani 1996, Domeniconi et al. 2002, Lindenbaum et al. 2004, Atiya 2005, Hall et al. 2008, Biau et al. 2010, Samworth 2012). The rationale of nearest neighbor rules is that close covariate vectors share similar properties more often than not.

In this article, we propose a k-nearest neighbor (kNN) rule for optimal treatment regimes. Although the rule is simple, it possesses good theoretical properties. Firstly, we show that the kNN rule for optimal treatment regimes is universally consistent. The notion of universal consistency is borrowed from machine learning. In optimal treatment regimes, it requires for a rule that when the sample size approaches infinity the rule eventually learns the Bayes rule without knowing any specifics about the distribution of the data. Secondly, we establish its convergence rate. The convergence rate is as high as $n^{-1/2}$ with appropriately chosen k if the dimension of covariates is 1 or 2, and the rate is $n^{-2/(p+2)}$ for dimensionality $p \geq 3$.

Similar to the nearest neighbor rule for classification, our proposed method suffers from the curse of dimensionality, *i.e.*, performance deteriorates as dimensionality increases. To

alleviate this problem, we propose an adaptive k-nearest neighbor (AkNN) rule, where the distance metric is adaptively determined from the data. Through adaptive metric selection, AkNN performs variable selection implicitly. The superior performance of AkNN over kNN is illustrated in the simulation studies. Actually, the kNN rule is a special case of AkNN. In practical settings, we recommend AkNN.

The remainder of the article is organized as follows. In Section 2.2, we propose the kNN rule for optimal treatment regimes, establish its theoretical properties, and improve the rule through an adaptive procedure. We present simulation studies to evaluate performance of the proposed methods in Section 2.3. The method is then illustrated on the Nefazodone-CBASP clinical trial (Keller et al. 2000) in Section 2.4. We conclude the article with a discussion in Section 2.5. Theoretical proofs are given in the Appendix.

2.2 Methods

2.2.1 Nearest neighbor rules

Consider a randomized clinical trial with L treatment arms. We observe a triplet $(\boldsymbol{X}, A, R)$ from each patient, where $\boldsymbol{X} = (X_1, \cdots, X_p)^T \in \mathbb{R}^p$ denotes the patient's clinical covariates, $A \in \mathcal{A} = \{1, 2, \cdots, L\}$ denotes the treatment assignment, and $R \in \mathcal{R}$ is the observed clinical outcome. Assume without loss of generality that larger values of R are preferred. Let $\pi_\ell(\boldsymbol{x}) := P(A = \ell | \boldsymbol{X} = \boldsymbol{x})$ be the probability of being assigned treatment ℓ for a patient with clinical covariates $\boldsymbol{x}$. This probability is predefined in the trial design.

A treatment regime is a function from $\mathbb{R}^p$ to $\mathcal{A}$. Mathematically, the expected outcome under any regime d is given as $\mathcal{V}(d) := \mathbb{E}(R | A = d(\boldsymbol{X}))$. This expectation is called the value function associated with the regime d. In other words, the value function $\mathcal{V}(d)$ is the

expected value of R given that the regime d is applied to the given population of patients. There is a positivity assumption that $\pi_\ell(\boldsymbol{X}) > 0$ for any $\ell \in \mathcal{A}$. That is, any treatment option must be represented in the data in order to estimate an optimal regime. An optimal treatment regime d^* is a regime that maximizes $\mathcal{V}(d)$. The regime d^* is also called the Bayes rule. For simplicity, let $m_\ell(\boldsymbol{x}) := \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = \ell)$. Because of the positivity assumption, $m_\ell(\boldsymbol{x})$ is well defined. It is easy to obtain that

$$d^*(\boldsymbol{x}) = \underset{\ell \in \{1,\cdots,L\}}{\operatorname{argmax}} m_\ell(\boldsymbol{x}). \tag{2.1}$$

The k-nearest neighbor rule is a nonparametric method used for classification and regression (Fix and Hodges 1951). In this article, we apply the nearest neighbor rule to optimal treatment regimes. The idea is simple. We use the nearest neighbor algorithm to estimate the conditional outcome $m_\ell(\boldsymbol{x})$ for each arm, and plug into (2.1) to get the nearest neighbor estimate for the optimal treatment regime.

Assume that the observed data $D_n = \{(\boldsymbol{X}_i, A_i, R_i) : i = 1, \cdots, n\}$ are collected independently. We fix $\boldsymbol{x} \in \mathbb{R}^p$, and reorder the observed data D_n according to increasing values of $||\boldsymbol{X}_i - \boldsymbol{x}||$. The reordered data sequence is denoted by

$$\Big(\boldsymbol{X}_{(1,n)}(\boldsymbol{x}), A_{(1,n)}(\boldsymbol{x}), R_{(1,n)}(\boldsymbol{x})\Big), \cdots, \Big(\boldsymbol{X}_{(n,n)}(\boldsymbol{x}), A_{(n,n)}(\boldsymbol{x}), R_{(n,n)}(\boldsymbol{x})\Big).$$

Thus $\boldsymbol{X}_{(1,n)}(\boldsymbol{x}), \cdots, \boldsymbol{X}_{(k,n)}(\boldsymbol{x})$ are k nearest neighbors of $\boldsymbol{x}$. When the k-nearest neighborhood of $\boldsymbol{x}$ is small, its conditional outcome for arm ℓ can be estimated by

$$\widehat{m}_\ell(\boldsymbol{x}) = \frac{\sum_{i=1}^k R_{(i,n)}(\boldsymbol{x}) \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)}}{\sum_{i=1}^k \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)}},$$

where $\mathbb{I}(\cdot)$ is the indicator function, as suggested in Murphy (2005). Here we define $0/0 = 0$. For simplicity, denote

$$W_{n,i}^\ell(\boldsymbol{x}) = \begin{cases} \dfrac{\frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)}}{\sum_{j=1}^k \frac{\mathbb{I}(A_{(j,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\big(\boldsymbol{X}_{(j,n)}(\boldsymbol{x})\big)}}, & \text{if } i \le k, \\ 0 & \text{if } i > k. \end{cases}$$

The k-nearest neighbor estimate $\widehat{m}_\ell$ can now be rewritten as,

$$\widehat{m}_\ell(\boldsymbol{x}) = \sum_{i=1}^{n} W_{n,i}^{\ell}(\boldsymbol{x}) R_{(i,n)}(\boldsymbol{x}). \tag{2.2}$$

Thus the plug-in estimate of the Bayes rule in (2.1) is

$$d^{NN}(\boldsymbol{x}) = \underset{\ell \in \{1,\cdots,L\}}{\operatorname{argmax}} \widehat{m}_\ell(\boldsymbol{x}). \tag{2.3}$$

This is called the k-nearest neighbor regime, for short kNN, in this article.

We need to address the problem of distance ties, *i.e.*, when $||\boldsymbol{x} - \boldsymbol{X}_i|| = ||\boldsymbol{x} - \boldsymbol{X}_j||$ for some $i \neq j$. Devroye et al. (1996, Section 11.2) discussed several methods for breaking distance ties. In practical use, we adopt the tie-breaking method used in Stone (1977). Subjects who have the same distance from $\boldsymbol{x}$ as the k-th nearest neighbor are averaged on the outcome R. We denote the distance of the k-th nearest neighbor to $\boldsymbol{x}$ by $\rho_n(\boldsymbol{x})$, and define the sets $A_n(\boldsymbol{x}) := \{i : ||\boldsymbol{x} - \boldsymbol{X}_i|| < \rho_n(\boldsymbol{x})\}$ and $B_n(\boldsymbol{x}) := \{i : ||\boldsymbol{x} - \boldsymbol{X}_i|| = \rho_n(\boldsymbol{x})\}$. The revised rule of (2.2) is as follows:

$$\widetilde{m}_\ell(\boldsymbol{x}) = \frac{\sum_{i \in A_n(\boldsymbol{x})} R_{(i,n)}(\boldsymbol{x}) \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)} + \frac{k - |A_n(\boldsymbol{x})|}{|B_n(\boldsymbol{x})|} \sum_{i \in B_n(\boldsymbol{x})} R_{(i,n)}(\boldsymbol{x}) \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}}{\sum_{i \in A_n(\boldsymbol{x})} \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)} + \frac{k - |A_n(\boldsymbol{x})|}{|B_n(\boldsymbol{x})|} \sum_{i \in B_n(\boldsymbol{x})} \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}}. \tag{2.4}$$

The corresponding nearest neighbor regime is the regime in (2.3) obtained by replacing $\widehat{m}_\ell(\boldsymbol{x})$ with $\widetilde{m}_\ell(\boldsymbol{x})$. This is not a strictly k-nearest neighbor rule when there are distance ties on the k-th nearest neighbor, since the estimate uses more than k neighbors.

Nearest neighbor regimes are based on local averaging. Here, k is a tuning parameter. It is required that k is small enough so that local changes of the distribution can be detected. On the other hand, k needs to be large so that averaging over the arm is effective. We may tune this parameter by a cross validation procedure to balance the two requirements. The v-fold cross validation procedure is described as follows. The data are randomly partitioned into v roughly equal-sized parts. We use $v - 1$ parts of the data to predict the optimal treatments on the part left out. We repeat the procedure $v - 1$ more times to predict the

other parts, and obtain the predicted treatment for each patient. The cross-validated value function is given by $\mathbb{P}_n[R\mathbb{I}(A = Pred)/\pi_A(\boldsymbol{X})]/\mathbb{P}_n[\mathbb{I}(A = Pred)/\pi_A(\boldsymbol{X})]$, where $\mathbb{P}_n$ denotes the empirical average and $Pred$ is the predicted treatment in the cross validation procedure. In this article, we apply 10-fold cross validation to tune the parameter k.

2.2.2 Theoretical Properties

In the literature of machine learning, a classification rule is called universally consistent if its expected error probability converges to the Bayes error probability, in probability or almost surely, for all distributions underlying the data (Devroye et al. 1996). Although this generality in distribution may seem to be a very strong condition, it has been known since a seminal paper by Stone (1977) that there do exist such classification rules. The k-nearest neighbor classification is the first to be proved to possess such universal consistency (Stone 1977, Devroye et al. 1996). Here, we extend the concept of universal consistency to optimal treatment regimes.

Definition 2.2.1. *Given a sequence D_n of data, an optimal treatment regime d_n is universally (weakly) consistent if*

$$\lim_{n\to\infty} \mathcal{V}(d_n) = \mathcal{V}(d^*) \quad \textit{in probability}$$

for all probability measures P on $\mathbb{R}^p \times \mathcal{A} \times \mathcal{R}$, and universally strongly consistent if

$$\lim_{n\to\infty} \mathcal{V}(d_n) = \mathcal{V}(d^*) \quad \textit{almost surely}$$

for all probability measures P on $\mathbb{R}^p \times \mathcal{A} \times \mathcal{R}$.

A consistent rule guarantees that by increasing the amount of data, the probability that the achieved value function is within a very small distance of the Bayes value goes arbitrarily close to one. Strong consistency means that by using more data the achieved value gets arbitrarily close to the Bayes value for every sequence of possible data.

Before proceeding to the theoretical analysis, we need some definitions. Denote the probability measure for $\boldsymbol{X}$ by μ, and let $S_{\boldsymbol{x},\epsilon}$ be the closed ball centered at $\boldsymbol{x}$ of radius $\epsilon > 0$. The collection of all $\boldsymbol{x}$ with $\mu(S_{\boldsymbol{x},\epsilon}) > 0$ for all $\epsilon > 0$ is called the support of μ (Cover and Hart 1967). The set is denoted as $support(\mu)$.

The analysis of universal consistency requires some assumptions:

(A1) There exists a constants $\zeta > 0$ such that $\pi_\ell(\boldsymbol{x}) \geq \zeta$ for any $\boldsymbol{x} \in support(\mu)$ and $\ell \in \{1, \cdots, L\}$.

(A2) $\mathbb{E}|R| < \infty$.

(A3) Distance ties occur with probability zero in μ.

These assumptions are quite weak. Assumption (A1) is just the positivity assumption, and ζ can be obtained by design. Assumption (A2) is natural. This assumption is automatically satisfied for bounded outcomes, *i.e.* $|R| \leq M < \infty$ for some constant M. Assumption (A3) is to avoid the messy problem of distance ties. When (A3) does not hold, we may add a small uniform variable $U \sim uniform[0, \epsilon]$ independent of $(\boldsymbol{X}, A, R)$ to the vector $\boldsymbol{X}$. This causes the $(p+1)$-dimensional random vector $\boldsymbol{X}' = (\boldsymbol{X}, U)$ to satisfy Assumption (A3). We may perform the k-nearest neighbor rule on the modified data $D'_n = \{(\boldsymbol{X}'_i, A_i, R_i) : i = 1, \cdots, n\}$. Because of the independence of U, the corresponding conditional outcome $m'_\ell(\boldsymbol{x}') = \mathbb{E}(R|\boldsymbol{X}' = \boldsymbol{x}', A = \ell) = m_\ell(\boldsymbol{x})$. Hence Assumption (A3) is reasonable, but at the cost of potential compromise of performance by introducing an artificial variable to the rule. When ϵ is very small, we actually break ties randomly. The difference with Stone's tie-breaking method in the previous section is that Stone's method takes into account all subjects whose distance to $\boldsymbol{x}$ equals that of the k-th nearest neighbor, while the tie-breaking method here only picks one of them randomly. The remark following the proof of Theorem 2.2.1 in the Appendix demonstrates that Stone's tie-breaking estimate in (2.4) is asymptotically better than the random tie-breaking method here.

The following theorem shows universal consistency of the nearest neighbor regime. The proof is provided in the Appendix.

Theorem 2.2.1. *For any distribution P for $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1)∼(A3),*

(i) the kNN regime in (2.3) is universally weakly consistent if $k \to \infty$ and $k/n \to 0$;

(ii) the kNN regime in (2.3) is universally strongly consistent if $k/\log(n) \to \infty$ and $k/n \to 0$.

If Assumption (A2) is tightened to $|R| \leq M < \infty$ for some constant M, the kNN regime in (2.3) is universally strongly consistent if $k \to \infty$ and $k/n \to 0$.

The next natural question is whether the associated value of the kNN regime tends to the Bayes value at a specified rate. To establish the rate of convergence, we require stronger assumptions.

(A1′) $\sum_{i=1}^{n} W_{n,i}^{\ell}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathbb{R}^p$ and $\ell = 1, \cdots, L$; and there exists a constant c such that $W_{n,i}^{\ell}(\boldsymbol{x}) \leq c/k$ for all $\boldsymbol{x} \in \mathbb{R}^p$, $i = 1, \cdots, n$ and $\ell = 1, \cdots, L$.

(A2′) There exists a constant σ^2 such that $\sigma_\ell^2(\boldsymbol{x}) := Var(R|\boldsymbol{X} = \boldsymbol{x}, A = \ell) \leq \sigma^2$ for all $\boldsymbol{x} \in \mathbb{R}^p$ and $\ell = 1, \cdots, L$.

(A3′) Distance ties occur with probability zero in μ, and the support of μ is compact with diameter 2ρ.

(A4′) m_ℓ's are Lipschitz continuous, *i.e.*, there exists a constant $C > 0$ such that $|m_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x}')| \leq C||\boldsymbol{x} - \boldsymbol{x}'||$ for any $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\mathbb{R}^p$, and $\ell = 1, \cdots, L$.

Assumption (A1′) implies that randomization is not extremely skewed with respect to the covariates. Assumptions (A2′)∼(A4′) are standard in the literature of nearest neighbor rules (Györfi et al. 2002). The following theorem gives the convergence rate of kNN regimes. This theorem is proved in the Appendix.

Theorem 2.2.2. *For any distribution P for $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1′)∼(A4′), there exists a sequence k such that $k \to \infty$ and $k/n \to 0$, and*

$$\mathbb{E}\left\{\left(\mathcal{V}(d^*) - \mathcal{V}(d^{NN})\right)^2\right\} = O(n^{-\beta}).$$

When $p = 1$, $\beta = 1/2$; when $p = 2$, β can be arbitrarily close to $1/2$; when $p \geq 3$, $\beta = 2/(p+2)$.

The rate of convergence is as high as $n^{-1/2}$ if dimensionality p is 1 or 2. When p increases, the convergence rate decreases significantly. As with nearest neighbor rules in classification and regression, the kNN regime also suffers from the curse of dimensionality.

2.2.3 Adaptive rule

Determination of nearest neighbors highly depends on the distance metric. Thus the performance of the kNN regime is affected by metric selection. The nearest neighbor regime is consistent as shown previously. However, it is well known that the curse of dimensionality can severely hurt nearest neighbor rules in finite samples. The rate of convergence in Theorem 2.2.2 is slow when dimensionality is high. Hence appropriate variable selection may improve performance. In this section, we propose an adaptive k-nearest neighbor algorithm to estimate the optimal treatment regime, and to perform metric selection and variable selection simultaneously.

Let $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)$ be a diagonal matrix, and use the quadratic form $(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T\Sigma(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ to compute the (squared) distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. σ_j is the scaling factor for the j-th covariate. Setting $\sigma_j = 0$ is equivalent to discarding the j-th covariate. We intend to set a large σ_j^2 if the j-th covariate is important for treatment selection.

We apply the following method to evaluate the importance of an individual covariate. It is related to a test statistic comparing two treatment regimes (Murphy 2005). One regime

d^j only involves the j-th covariate, and the other d^0, called non-informative regime, assigns all patients to the treatment with the largest estimated conditional outcome $\widehat{\mathbb{E}}(R|A=\ell) := \sum_{i=1}^n \frac{R_i \mathbb{I}(A_i=\ell)}{\pi_\ell(\boldsymbol{X}_i)} / \sum_{i=1}^n \frac{\mathbb{I}(A_i=\ell)}{\pi_\ell(\boldsymbol{X}_i)}$. For a specific regime d, let d_i be the treatment assignment for the i-th subject according to d. The value function associated with d is estimated by

$$\widehat{\mathcal{V}}(d) = \frac{\sum_{i=1}^n R_i \frac{\mathbb{I}(A_i=d_i)}{\pi_{A_i}(\boldsymbol{X}_i)}}{\sum_{i=1}^n \frac{\mathbb{I}(A_i=d_i)}{\pi_{A_i}(\boldsymbol{X}_i)}}. \tag{2.5}$$

When we compare two treatment regimes, d^j and d^0, a consistent estimator of the variance of $\sqrt{n}(\widehat{\mathcal{V}}(d^j) - \widehat{\mathcal{V}}(d^0))$ is

$$\begin{aligned} & \widehat{Var}\left(\sqrt{n}(\widehat{\mathcal{V}}(d^j) - \widehat{\mathcal{V}}(d^0))\right) \\ = \; & \frac{1}{n}\sum_{i=1}^n \left\{ \left(\frac{\mathbb{I}(A_i = d_i^j)}{\pi_{A_i}(\boldsymbol{X}_i)}(R_i - \widehat{\mathcal{V}}(d^j)) \right)^2 + \left(\frac{\mathbb{I}(A_i = d_i^0)}{\pi_{A_i}(\boldsymbol{X}_i)}(R_i - \widehat{\mathcal{V}}(d^0)) \right)^2 \right\}. \end{aligned} \tag{2.6}$$

Note that even if d^j only depends on one single covariate, or d^0 is not related to any covariate, $\pi_{A_i}(\boldsymbol{X}_i)$ in (2.5) and (2.6) is still based on the whole covariate vector by design. The statistic

$$T_j = \frac{\sqrt{n}(\widehat{\mathcal{V}}(d^j) - \widehat{\mathcal{V}}(d^0))}{\sqrt{\widehat{Var}\left(\sqrt{n}(\widehat{\mathcal{V}}(d^j) - \widehat{\mathcal{V}}(d^0))\right)}} \tag{2.7}$$

asymptotically has a standard normal distribution under the null hypothesis that $\mathcal{V}(d^j) = \mathcal{V}(d^0)$. See Murphy (2005) for technical details. When the statistic is greater than zero, regime d^j is considered better than the non-informative regime d^0, otherwise d^0 is better. The statistic T_j reflects the importance of the j-th covariate on optimal treatment regimes. We estimate d^j by kNN only using the j-th covariate. The parameter k is tuned by 10-fold cross validation. The treatment assignments, d_i^j's, are also obtained by cross validation to avoid over-fitting.

We set $\sigma_j^2 = (T_j + \Delta)_+$ for each $j = 1, \cdots, p$, where $\Delta \in \mathbb{R}$ is a predefined parameter and $(\cdot)_+$ is the positive part. The adaptive k-nearest neighbor regime, denoted AkNN, follows the same procedure used in the kNN regime described above, except that AkNN

uses an adaptive distance metric $(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma (\boldsymbol{x}_1 - \boldsymbol{x}_2)$ to compute the (squared) distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. When Δ is small (for example, $\Delta \to -\infty$), all the σ_j^2's are zero, hence the AkNN regime degenerates to a non-informative regime. On the other hand, when Δ is large (for example, $\Delta \to \infty$), all the σ_j^2's are almost identical, and the AkNN regime is equivalent to the kNN one.

Here is a summary of the adaptive nearest neighbor procedure:

1) Normalize each covariate to a similar scale.

2) Calculate T_j in (2.7) for each covariate $j = 1, \cdots, p$.

3) Calculate $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)$, where $\sigma_j^2 = (T_j + \Delta)_+$.

4) Estimate a kNN regime using the distance metric,

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma (\boldsymbol{x}_1 - \boldsymbol{x}_2)}.$$

The scaling at the first step is to avoid covariates in greater numeric ranges dominating those in smaller numeric ranges. We recommend linearly scaling each covariate to the range $[-1, +1]$ or $[0, 1]$ (Hsu et al. 2003). For AkNN, there are two tuning parameters, k and Δ. We tune the parameters using 10-fold cross validation as done previously.

2.3 Simulation studies

We carried out extensive simulation studies to investigate finite sample performance of the proposed kNN and AkNN methods. In the simulations, we generated 10-dimensional vectors of clinical covariates $x_1, \cdots, x_{10}$, consisting of independent uniform random variables $U(-1, 1)$. The treatment A was generated from $\{-1, 1\}$ independently of X with $P(A = 1) = 0.5$. The response R was normally distributed with mean $Q_0 = \mu_0(\boldsymbol{x}) + \delta_0(\boldsymbol{x}) \cdot a$ and standard deviation 1, where $\mu_0(\boldsymbol{x})$ is the common effect for clinical covariates,

and $\delta_0(\boldsymbol{x}) \cdot a$ is the interaction between treatment and clinical covariates. We considered three scenarios with different choices of $\mu_0(\boldsymbol{x})$ and $\delta_0(\boldsymbol{x})$:

(1) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 1.8(0.3 - x_1 - x_2)$.

(2) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2$; $\delta_0(\boldsymbol{x}) = 10(1 - x_1^2 - x_2^2)(x_1^2 + x_2^2 - 0.2)$.

(3) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2$; $\delta_0(\boldsymbol{x}) = 5(1 - \sum_{i=1}^{10} \frac{x_i^2}{i})$.

The scenarios have different true decision boundaries. The decision boundaries in the first two scenarios are determined by x_1 and x_2 only. The decision boundary is a line in Scenario 1 and a ring in Scenario 2. The decision boundary in Scenario 3 is a sphere in $\mathbb{R}^{10}$, while the first several covariates play an important role in the regime.

We compared the performance of the following five methods: (1) The proposed kNN method; (2) The proposed AkNN method; (3) OWL proposed in Zhao et al. (2012) using the Gaussian RBF kernel (OWL-Gaussian); (4) OWL using the linear kernel (OWL-Linear); and (5) ℓ_1-PLS proposed by Qian and Murphy (2011).

OWL methods view the optimal treatment regime problem as a weighted classification problem, and treat the original outcomes as weights. OWL-Linear incorporates the linear kernel to estimate linear treatment regimes, while OWL-Gaussian has the ability to detect nonlinear regimes. Since OWL methods can only handle nonnegative outcomes, we subtracted $\min\{R_i\}$ from all outcome responses. This is the approach used in Zhao et al. (2012) [personal communication]. ℓ_1-PLS is a two-step method. In the simulation studies, ℓ_1-PLS approximated $\mathbb{E}(R|\boldsymbol{X}, A)$ using the basis function set $(1, \boldsymbol{X}, A, \boldsymbol{X}A)$, and applied LASSO for variable selection. The optimal treatment regime was determined by the sign of the difference between the estimated $\mathbb{E}(R|\boldsymbol{X}, A = 1)$ and $\mathbb{E}(R|\boldsymbol{X}, A = -1)$.

For each scenario, we varied sample sizes for training datasets from 50, 100, 200, 400, to 800, and repeated the simulation 500 times. All five methods had at least one

tuning parameter. In the simulations, we applied 10-fold cross validation to tune parameters. For comparison, the performances of the five methods were evaluated by the value function under the estimated optimal treatment regime when applied to an independent and large test dataset. Specifically, a test set with 10,000 observations was simulated to assess the performance. The estimated value function under any regime d is given by $\mathbb{P}_n^*[R\mathbb{I}(A = d(\boldsymbol{X}))/\pi_A(\boldsymbol{X})]/\mathbb{P}_n^*[\mathbb{I}(A = d(\boldsymbol{X}))/\pi_A(\boldsymbol{X})]$ (Murphy 2005), where $\mathbb{P}_n^*$ denotes the empirical average using the test data.

The simulation results are presented in Table 2.1. We report the mean and standard deviation of value functions over 500 replications. Scenario 1 was a linear example. The model specification in ℓ_1-PLS was correct, and this method performed very well. kNN showed better performance than both OWL methods. AkNN further improved the performance of kNN, and yielded similar performance to ℓ_1-PLS especially when the sample size is large. The conditional mean outcomes and decision boundaries in the remaining two scenarios were nonlinear. ℓ_1-PLS and OWL-Linear were misspecified, and hence they did not perform very well in either scenarios. We focused on the comparison among OWL-Gaussian, kNN and AkNN. In Scenario 2, either OWL-Gaussian or kNN did not perform better than ℓ_1-PLS or OWL-Linear, even though they can detect nonlinear decision boundary. The poor performance is probably due to noise covariates. It is well known that nearest neighbor rules deteriorate when there are irrelevant covariates presented in the data. Due to the variable selection mechanism with AkNN, this method outperformed other methods in terms of achieving much higher values. Scenario 3 was an example where all covariates contributed to the optimal treatment regime. However, their level of contribution varied. The first covariate was the most important, while the tenth was the least. In this scenario, kNN and OWL-Gaussian showed better performance than both OWL-Linear and ℓ_1-PLS. AkNN again produced the best performance. The adaptive selection on the distance metric enhances the nearest neighbor rule. From the simulations, AkNN is always better than kNN. As we explained before, when the tuning parameter Δ is very large, AkNN is almost

Table 2.1: *Mean (std) of empirical value function for three simulation scenarios evaluated on independent validation data. The best value function for each scenario and sample size combination is in bold.*

	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
Scenario 1 (Optimal value 2.25)					
l_1-PLS	**2.18(0.07)**	**2.21(0.04)**	**2.23(0.02)**	**2.24(0.02)**	**2.24(0.01)**
OWL-Linear	1.64(0.22)	1.77(0.20)	1.90(0.18)	2.00(0.13)	2.10(0.08)
OWL-Gaussian	1.62(0.23)	1.76(0.20)	1.88(0.18)	2.00(0.13)	2.10(0.08)
kNN	1.87 (0.14)	2.00 (0.10)	2.09 (0.06)	2.15 (0.04)	2.19 (0.02)
AkNN	1.98 (0.20)	2.13 (0.12)	2.20 (0.05)	2.23 (0.03)	**2.24 (0.02)**
Scenario 2 (Optimal value 3.87)					
l_1-PLS	2.28(0.12)	2.31(0.12)	2.34(0.12)	2.38(0.13)	2.44(0.12)
OWL-Linear	2.37(0.14)	2.42(0.14)	2.45(0.13)	2.49(0.11)	2.53(0.07)
OWL-Gaussian	2.37(0.13)	2.40(0.12)	2.43(0.12)	2.49(0.11)	2.54(0.09)
kNN	2.38 (0.13)	2.41 (0.10)	2.44 (0.07)	2.50 (0.06)	2.59 (0.05)
AkNN	**2.51 (0.26)**	**2.98 (0.42)**	**3.57 (0.30)**	**3.80 (0.05)**	**3.85 (0.02)**
Scenario 3 (Optimal value 3.85)					
l_1-PLS	2.39(0.09)	2.40(0.08)	2.41(0.09)	2.43(0.09)	2.45(0.10)
OWL-Linear	2.37(0.08)	2.38(0.08)	2.39(0.07)	2.41(0.07)	2.43(0.06)
OWL-Gaussian	2.50(0.16)	2.67(0.19)	2.90(0.17)	3.12(0.13)	3.35(0.11)
kNN	2.58 (0.07)	2.66 (0.07)	2.75 (0.06)	2.85 (0.06)	2.98 (0.05)
AkNN	**3.31 (0.31)**	**3.54 (0.11)**	**3.61 (0.06)**	**3.67 (0.03)**	**3.70 (0.02)**

equivalent to kNN. Considering the superior performance of AkNN over kNN, we suggest AkNN for general practical use.

2.4 Data analysis

We applied the proposed methods to analyze data from the Nefazodone-CBASP clinical trial (Keller et al. 2000). The Nefazodone-CBASP trial compared three different treatments for patients suffering chronic depression. Patients with non-psychotic chronic major depressive disorder (MDD) were randomized in a 1:1:1 ratio to either Nefazodone (NFZ), cognitive behavioral-analysis system of psychotherapy (CBASP), or the combination of

Nefazodone and CBASP (COMB). The primary outcome measurement in efficacy was the score on the 24-item Hamilton Rating Scale for Depression (HRSD). Lower HRSD is desirable. We considered 50 pre-treatment covariates as in Zhao et al. (2012). We excluded some patients with missing covariate values from the analyses. The data used here consisted of 647 patients. Among them, 216, 220, and 211 patients were assigned to NFZ, CBASP and COMB, respectively. Each clinical covariate was linearly scaled to the range $[-1, +1]$, as recommended in Hsu et al. (2003).

Since the trial had three treatment arms, we compared the performance of AkNN with l_1-PLS. The OWL methods can only deal with two treatments. From the simulation studies, AkNN always outperforms kNN, so we only considered AkNN in this section. Outcomes used in the analyses were opposites of the HRSD scores. l_1-PLS used the basis function set $(1, \boldsymbol{X}, A, \boldsymbol{X}A)$ in the regression model. We used nested 10-fold cross-validation for an unbiased comparison (Ambroise and McLachlan 2002). Specifically, the data were randomly partitioned into 10 roughly equal-sized parts. We used nine parts as training data to predict optimal treatments for patients in the part left out. The parameter tuning was based on inner 10-fold cross-validation on the training data. We repeated the procedure 10 times, and obtained the predicted treatment for each patient. We then computed the estimated value function as $\mathbb{P}_n[R\mathbb{I}(A = Pred)/\pi_A(\boldsymbol{X})]/\mathbb{P}_n[\mathbb{I}(A = Pred)/\pi_A(\boldsymbol{X})]$, where $\mathbb{P}_n$ denotes the empirical average over the data and $Pred$ is the predicted treatment in the cross validation procedure. To obtain reliable estimates, we repeated the nested cross-validation procedure 100 times with different fold partitions. The mean value functions over 100 repeats and the standard deviations are presented in Table 2.2. Note that the standard deviation here was calculated from 100 estimated value functions, and reflected variance across repetitions. It is slightly different from the standard deviation in traditional 10-fold cross validation procedure, which takes account of the variance not only across repetitions but also across folds. As evident from Table 2.2, AkNN achieved a similar performance to l_1-PLS. Both methods assigned the combination treatment to almost all

Table 2.2: *Mean (std) Hamilton rating scale for depression (HRSD) from the cross-validation procedure using different methods. Lower HRSD is better.*

	l_1-PLS	OWL-Linear	OWL-Gaussian	AkNN
NFZ vs. CBASP vs. COMB	11.19 (0.15)	–	–	11.18 (0.27)
NFZ vs. CBASP	16.30 (0.39)	15.91 (0.40)	16.15 (0.40)	15.70 (0.39)
NFZ vs. COMB	11.20 (0.16)	10.87 (0.00)	10.92 (0.06)	11.03 (0.27)
CBASP vs. COMB	10.95 (0.09)	10.87 (0.00)	10.88 (0.02)	11.41 (0.28)

patients. The original analysis in Keller et al. (2000) indicated that the combination of Nefazodone and CBASP (COMB) is significantly more efficacious than either treatment alone. Our analysis confirmed this is indeed true.

We also performed pairwise comparisons between two treatment arms. We included OWL-Linear and OWL-Gaussian in the analysis. The same nested cross validation procedure was used to evaluate performance. The analysis results are also presented in Table 2.2. As explained before, the standard deviations only reflect variance across repetitions, and hence are smaller. The zero standard deviations for OWL-Linear in Table 2.2 were due to the fact that OWL-Linear assigned all patients to the COMB arm in every repeat of the cross-validation procedure. For comparison between NFZ and CBASP, AkNN was slightly better than other methods. For comparison between NFZ and COMB, all methods produced similar performance. For comparison between CBASP and COMB, AkNN did not perform comparably to other methods. We carried out the significance test described in Section 2.2.3 to compare the regimes by AkNN and OWL-Linear (a non-informative regime) for CBASP vs. COMB. The test was based on the cross validation prediction. Since we repeated the cross validation 100 times, we obtained 100 p-values. The p-values range from 0.11 to 0.94 with a median at 0.56. Although the value from AkNN was not as good as that from OWL-Linear, the difference was not statistically significant.

2.5 Discussion

In this article, we have proposed a simple kNN rule for optimal treatment regimes, and developed an adaptive method (AkNN) to determine the distance metric. As shown in the simulation and data studies, the simple method can rival and improve upon sophisticated methods. The proposed methods are easy to implement. We have implemented them using an R package *FNN*, which utilizes fast k-nearest neighbor search algorithms, such as cover tree (Beygelzimer et al. 2006) and k-d tree (Friedman et al. 1977), to speed up the brute force search. Our nearest neighbor algorithms can be fast even for data with a large sample size.

The proposed methods are nonparametric. According to universal consistency, kNN will eventually learn the optimal treatment regime as the sample size increases. Two-step methods, such as ℓ_1-PLS, are generally parametric. Their performance mainly depends on how close the posited model is to the true model of the conditional mean outcome. When the model is correctly specified, two-step methods can be more efficient than nonparametric methods. However, the relationship among clinical covariates, treatment assignment, and outcome is complex in practice. Misspecified models may lead to biased estimates.

CHAPTER 3: RESIDUAL WEIGHTED LEARNING

3.1 Introduction

Personalized medicine is a medical paradigm that utilizes individual patient information to optimize patients health care. Recently, personalized medicine has received much attention among statisticians, computer scientists, and clinical practitioners. The primary motivation is the well-established fact that patients often show significant heterogeneity in response to treatments. For instance, in a recent study, it was demonstrated that the optimal timing for the initiation of antiretroviral therapy (ART) varies in patients co-infected with human immunodeficiency virus and tuberculosis. Patients with CD4+ T-cell counts of less than 50 per cubic millimeter benefited substantially from earlier ART with a lower rate of new AIDS-defining illnesses and mortality as compared with later ART, while those with larger CD4+ T-cell counts did not have such a benefit (Havlir et al. 2011). The inherent heterogeneity across patients suggests a transition from the traditional "one size fits all" approach to modern personalized medicine.

A major component of personalized medicine is the estimation of individualized treatment rules (ITRs). Formally, the goal is to seek a rule that assigns a treatment, from among a set of possible treatments, to a patient based on his or her clinical, prognostic or genomic characteristics. The individualized treatment rules are also called optimal treatment regimes. There is a significant literature on individualized treatment strategies based on data from clinical trials or observational studies (Murphy 2003, Robins 2004, Zhang et al. 2012b, Zhao et al. 2009). Much of the work relies on modeling either the conditional mean outcomes or contrasts between mean outcomes. These methods obtain ITRs indi-

rectly by inverting the regression estimates instead of directly optimizing the decision rule. For instance, Qian and Murphy (2011) applied a two-step procedure that first estimates a conditional mean for the outcome and then determines the treatment rule by comparing the conditional means across various treatments. The success of these indirect approaches highly depends on correct specification of posited models and on the precision of model estimates.

In contrast, Zhao et al. (2012) proposed outcome weighted learning (OWL), using data from a randomized clinical trial, to construct an ITR that directly optimizes the clinical outcome. They cast the treatment selection problem as a weighted classification problem, and apply state-of-the-art support vector machines for implementation. This approach opens the door to introducing machine learning techniques into this area. However, there is still significant room for improved performance. First, the estimated ITR of OWL is affected by a simple shift of the outcome. This behavior makes the estimate of OWL unstable. Second, since OWL requires the outcome to be nonnegative, OWL works similarly to weighted classification, in which misclassification errors, the differences between the estimated and true treatment assignments, are targeted to be reduced. Hence the ITR estimated by OWL tries to keep treatment assignments that subjects actually received. This is not always ideal since treatments are randomly assigned in the trial, and the probability is slim that the majority of subjects are assigned optimal treatments. Third, OWL does not have variable selection features. When there are many clinical covariates which are not related to the heterogeneous treatment effects, variable selection is critical to the performance.

To alleviate these problems, we propose a new method, called Residual Weighted Learning (RWL). Unlike OWL which weights misclassification errors by clinical outcomes, RWL weights these errors by residuals from a regression fit of the outcome. The predictors of the regression model include clinical covariates, but exclude treatment assignment. Thus the residuals better reflect the heterogeneity of treatment effects. Since

some residuals are negative, the hinge loss function used in OWL is not appropriate. We instead utilize the smoothed ramp loss function, and provide a difference of convex (d.c.) algorithm to solve the corresponding non-convex optimization problem. The smoothed ramp loss resembles the ramp loss (Collobert et al. 2006), but it is smooth everywhere. It is well known that ramp loss related methods are shown to be robust to outliers (Wu and Liu 2007). The robustness to outliers for the smoothed ramp loss is helpful for RWL, especially when residuals are poorly estimated. Moreover, through using residuals, RWL is able to deal relatively easily with almost all types of outcomes. For example, RWL can work with generalized linear models to construct ITRs for count/rate outcomes. We also propose variable selection approaches in RWL for linear and nonlinear rules, respectively, to further improve finite sample performance.

The theoretical analysis of RWL focuses on two aspects, universal consistency and convergence rate. The notion of universal consistency is borrowed from machine learning. It requires for a learning method that when the sample size approaches infinity the method eventually learns the Bayes rule without knowing any specifics of the distribution of the data. We show that RWL with a universal kernel (e.g. Gaussian RBF kernel) is universally consistent. In machine learning, there is a famous "no-free-lunch theorem", which states that the convergence rate of any particular learning rule may be arbitrarily slow (Devroye et al. 1996). In this work, we prove the "no-free-lunch theorem" for finding ITRs. Thus the rate of convergence studies for a particular rule must necessarily be accompanied by conditions on the distribution of the data. For RWL with Gaussian RBF kernel, we show that under the geometric noise condition (Steinwart and Scovel 2007) the convergence rate is as high as $n^{-1/3}$.

At first glance, one may think that there is not a large difference between OWL and RWL except that RWL uses residuals as alternative outcomes. Actually, RWL enjoys many benefits from this simple modification. First, by using residuals of outcomes, RWL is able

to reduce the variability introduced by the original outcomes. Second, since the numbers of subjects with positive and negative residuals are generally balanced, the ITR determined by RWL favors neither the treatment assignments that subjects actually received nor their opposites. Because of the above reasons, RWL improves finite sample performance. Third, RWL possesses location-scale invariance with respect to the original outcomes. Specifically, the estimated rule of RWL is invariant to a shift of the outcome; it is invariant to a scaling of the outcome with a positive number; the rule from RWL that maximizes the outcome is opposite to the rule that minimizes the outcome. These are intuitively sensible.

The contributions of this work are summarized as follows. (1) We propose the general framework of Residual Weighted Learning to estimate individualized treatment rules. By estimating residuals with linear or generalized linear models, RWL can effectively deal with different types of outcomes, such as continuous, binary and count outcomes. For censored survival outcomes, RWL could potentially utilize martingale residuals, although theoretical justification is still under development. (2) We develop variable selection techniques in RWL to further improve performance. (3) We present a comprehensive theoretical analysis of RWL on universal consistency and convergence rate. (4) As a by-product, we show the "no-free-lunch theorem" for ITRs that the convergence rate of any particular rule may be arbitrarily slow. This is a generic result, and it applies to any algorithm for optimal treatment regimes.

The remainder of the chapter is organized as follows. In Section 3.2, we discuss Outcome Weighted Learning (OWL) and propose Residual Weighted Learning (RWL) to improve finite sample performance for continuous outcomes. Then we develop a general framework for RWL to handle other types of outcomes, and use binary and count/rate outcomes as examples. In Section 3.3, we establish consistency and convergence rate results for the estimated rules. The variable selection techniques for RWL are discussed in Section 3.4. We present simulation studies to evaluate performance of the proposed methods in

Section 3.5. The method is then illustrated on the EPIC cystic fibrosis randomized clinical trial (Treggiari et al. 2009; 2011) in Section 3.6. We conclude with a discussion in Section 3.7. All proofs are given in Appendix.

3.2 Methodology

3.2.1 Outcome Weighted Learning

Consider a two-arm randomized trial. We observe a triplet $(\boldsymbol{X}, A, R)$ from each patient, where $\boldsymbol{X} = (X_1, \cdots, X_p)^T \in \mathcal{X}$ denotes the patient's clinical covariates, $A \in \mathcal{A} = \{1, -1\}$ denotes the treatment assignment, and R is the observed clinical outcome, also called the "reward" in the literature on reinforcement learning. We assume that R is bounded, and larger values of R are more desirable. An individualized treatment rule (ITR) is a function from $\mathcal{X}$ to $\mathcal{A}$. Let $\pi(a, \boldsymbol{x}) := P(A = a|\boldsymbol{X} = \boldsymbol{x})$ be the probability of being assigned treatment a for patients with clinical covariates $\boldsymbol{x}$. It is predefined in the trial design. Here we consider a general situation. In most clinical trials, the treatment assignment is independent of $\boldsymbol{X}$, but in some designs, such as stratified designs, A may depend on $\boldsymbol{X}$. We assume $\pi(a, \boldsymbol{x}) > 0$ for all $a \in \mathcal{A}$ and $\boldsymbol{x} \in \mathcal{X}$.

An optimal ITR is a rule that maximizes the expected outcome under this rule. Mathematically, the expected outcome under any ITR d is given as

$$\mathbb{E}(R|A = d(\boldsymbol{X})) = \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}\Big(A = d(\boldsymbol{X})\Big)\right), \tag{3.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Interested readers may refer to Qian and Murphy (2011) and Zhao et al. (2012) regarding the derivation of (3.1). This expectation is called the value function associated with the rule d, and is denoted $V(d)$. In other words, the value function $V(d)$ is the expected value of R given that the rule $d(\boldsymbol{X})$ is applied to the given population of patients. An optimal ITR d^* is a rule that maximizes $V(d)$. Finding d^*

is equivalent to the following minimization problem:

$$d^* \in \arg\min_d \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}\Big(A \neq d(\boldsymbol{X})\Big)\right). \tag{3.2}$$

Zhao et al. (2012) viewed this as a weighted classification problem, in which one wants to classify A using $\boldsymbol{X}$ but also weights each misclassification error by $R/\pi(A, \boldsymbol{X})$.

Assume that the observed data $\{(\boldsymbol{x}_i, a_i, r_i) : i = 1, \cdots, n\}$ are collected independently. For any decision function $f(\boldsymbol{x})$, let $d_f(\boldsymbol{x}) = sign\big(f(\boldsymbol{x})\big)$ be the associated rule, where $sign(u) = 1$ for $u > 0$ and -1 otherwise. The particular choice of the value of $sign(0)$ is not important. In this work, we fix $sign(0) = -1$. Using the observed data, the weighted classification error in (3.2) can be approximated by the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi(a_i, \boldsymbol{x}_i)}\mathbb{I}\Big(d_f(\boldsymbol{x}_i) \neq a_i\Big). \tag{3.3}$$

The optimal decision function $\widehat{f}(\boldsymbol{x})$ is the one that minimizes the outcome weighted error (3.3).

It is well known that empirical risk minimization for a classification problem with the 0-1 loss function is an NP-hard problem. To alleviate this difficulty, one often finds a surrogate loss to replace the 0-1 loss. Outcome weighted learning (OWL) proposed by Zhao et al. (2012) uses the hinge loss function, and also applies the regularization technique used in the support vector machine (SVM) (Vapnik 1998). In other words, instead of minimizing (3.3), OWL aims to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi(a_i, \boldsymbol{x}_i)}\Big(1 - a_i f(\boldsymbol{x}_i)\Big)_+ + \lambda||f||^2, \tag{3.4}$$

where $(u)_+ = \max(u, 0)$ is the positive part of u, $||f||$ is some norm for f, and λ is a tuning parameter controlling the trade-off between empirical risk and complexity of the decision function f.

OWL opens the door to the application of statistical learning techniques to personalized medicine. However, this approach is not perfect. First, a simple shift on the outcome R

should not affect the optimal decision rule, as seen from (3.1) and (3.2). That is, d^* in (3.2) does not change if R is replaced by $R + c$, for any constant c. Unfortunately this nice invariance property does not hold for the decision function $\widehat{f}(\boldsymbol{x})$ of OWL in (3.4) when r_i is replaced by $r_i + c$. Consider an extreme case where c is very large and $\pi(1, \boldsymbol{x}) = \pi(-1, \boldsymbol{x}) = 0.5$ for all $\boldsymbol{x} \in \mathcal{X}$, the weights $(r_i + c)/\pi(a_i, \boldsymbol{x}_i)$ are almost identical, and the weighted problem is approximately transformed to an unweighted one. Hence the performance of OWL can be affected by a simple shift of R. Second, OWL further assumes that R is nonnegative to gain computational efficiency from convex programming. A direct consequence of this assumption, as seen from (3.3), is that the treatment regime $d_f(\boldsymbol{x}_i)$ tends to match the treatment a_i that was actually assigned to the patient, especially when the decision function f is chosen from a rich class of functions. This property is not ideal for data from a randomized clinical trial, since treatments are actually randomly assigned to patients.

3.2.2 Residual Weighted Learning

In this section, we only consider the case that the outcome R is continuous, and extend the framework to other types of outcomes in Section 3.2.4.

As demonstrated previously, the decision rule in (3.2) is invariant to a shift of outcome R by any constant. Moreover, Lemma 3.2.1 shows that d^* in (3.2) is invariant under a shift of R by a function of $\boldsymbol{X}$. That is, d^* remains unchanged if R is replaced by $R - g(\boldsymbol{X})$ for any function g, as long as g is not related to d. So there is an optimal solution $\widehat{f}$ in (3.4) corresponding to a particular function g, when r_i is replaced by $r_i - g(\boldsymbol{x}_i)$, and the associated rule $\widehat{d}_f = sign(\widehat{f})$ can be seen as an estimated optimal rule in (3.2). As g varies, we obtain a collection of $\widehat{d}_f$'s. When the sample size is very large, these $\widehat{d}_f$'s perform similarly by the infinite sample property of OWL shown in Zhao et al. (2012). However, when the sample size is limited, as in clinical settings, the choice of g is critical

to the performance of OWL. Zhao et al. (2012) did not delve deeply into this problem. In this work, we will provide a solution to this, and improve finite sample performance.

Lemma 3.2.1. *For any measurable $g : \mathcal{X} \to \mathbb{R}$ and any probability distribution for $(\boldsymbol{X}, A, R)$,*

$$\mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) = \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) - \mathbb{E}\left(g(\boldsymbol{X})\right).$$

Intuitively, the function g with the smallest variance of $\dfrac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)$ is a good choice. However, such a function depends on the decision rule d, as shown in the following theorem. The proof of the theorem is provided in Appendix.

Theorem 3.2.1. *Among all measurable $g : \mathcal{X} \to \mathbb{R}$, $\widetilde{g}(\boldsymbol{X}) = \mathbb{E}\left(\dfrac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right)$ is the function g that minimizes the variance of $\dfrac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)$.*

Our purpose is to find a function g, which is not related to d, to reduce the variance of $\dfrac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)$ as much as possible. As shown in the proof, the minimizer $\widetilde{g}$ can be written as,

$$\widetilde{g}(\boldsymbol{X}) = \mathbb{E}(R|\boldsymbol{X}, A = 1)\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big) + \mathbb{E}(R|\boldsymbol{X}, A = -1)\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big).$$

That is, $\widetilde{g}(\boldsymbol{X})$ jumps between $\mathbb{E}(R|\boldsymbol{X}, A = 1)$ and $\mathbb{E}(R|\boldsymbol{X}, A = -1)$ as $d(\boldsymbol{X})$ varies. Hence when d is unknown, a reasonable choice of g is

$$g^*(\boldsymbol{X}) = \frac{\mathbb{E}(R|\boldsymbol{X}, A = 1) + \mathbb{E}(R|\boldsymbol{X}, A = -1)}{2} = \mathbb{E}\left(\frac{R}{2\pi(A, \boldsymbol{X})}|\boldsymbol{X}\right). \tag{3.5}$$

We propose to minimize the following empirical risk, rather than the original one in (3.3):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i - \widehat{g}^*(\boldsymbol{x}_i)}{\pi(a_i, \boldsymbol{x}_i)}\mathbb{I}\Big(d_f(\boldsymbol{x}_i) \neq a_i\Big), \tag{3.6}$$

where $\widehat{g}^*$ is an estimate of g^*. For simplicity, let $\widehat{r}_i = r_i - \widehat{g}^*(\boldsymbol{x}_i)$ be the estimated residual. Here, we do not weight misclassification errors by clinical outcomes as OWL does,

and instead we weight them by residuals from a regression fit of outcomes. Thus the optimal decision function is invariant under any translation of clinical outcomes. We call the method Residual Weighted Learning (RWL).

RWL also has the following interpretation. The outcome R can be characterized as

$$R = \mu(\boldsymbol{X}) + \delta(\boldsymbol{X}) \cdot A + \epsilon,$$

where ϵ is the mean zero random error. $\mu(\boldsymbol{X})$ reflects common effects of clinical covariates $\boldsymbol{X}$ for both treatment arms. It is easy to see that $\mu(\boldsymbol{X}) = (\mathbb{E}(R|\boldsymbol{X}, A = 1) + \mathbb{E}(R|\boldsymbol{X}, A = -1))/2$, so the residual, $R - g^*(\boldsymbol{X})$, in RWL is just $\delta(\boldsymbol{X}) \cdot A + \epsilon$, which captures all sources of heterogeneous treatment effects. The rationale of optimal treatment regimes is to keep treatment assignments that subjects have actually received if those subjects are observed to have large outcomes, and to switch assignments if outcomes are small. However, the largeness and smallness for outcomes are relative. A rather large outcome may still be considered as small when compared with subjects having similar clinical covariates, as shown in Figure 3.1. Outcomes are not comparable among subjects with different clinical covariates, while the residual, by removing common covariates effects, is a better measurement. Figure 3.1 illustrates how residuals work using an example with a single covariate X. The raw data are shown on the left, and residuals are shown on the right. Residuals are comparable among subjects, and larger residuals represent better outcomes.

Another benefit from using residuals is a clear cut-off, *i.e.* 0, to distinguish between subjects with good and poor clinical outcomes. To minimize the empirical risk in (3.6), for subjects with positive residuals, RWL is apt to recommend the same treatment assignments that subjects have actually received; for subjects with negative residuals, RWL is more likely to give the opposite treatment assignments to what they have received. The optimal ITR is the one maximizing the conditional expected outcome given in (3.1). An empirical

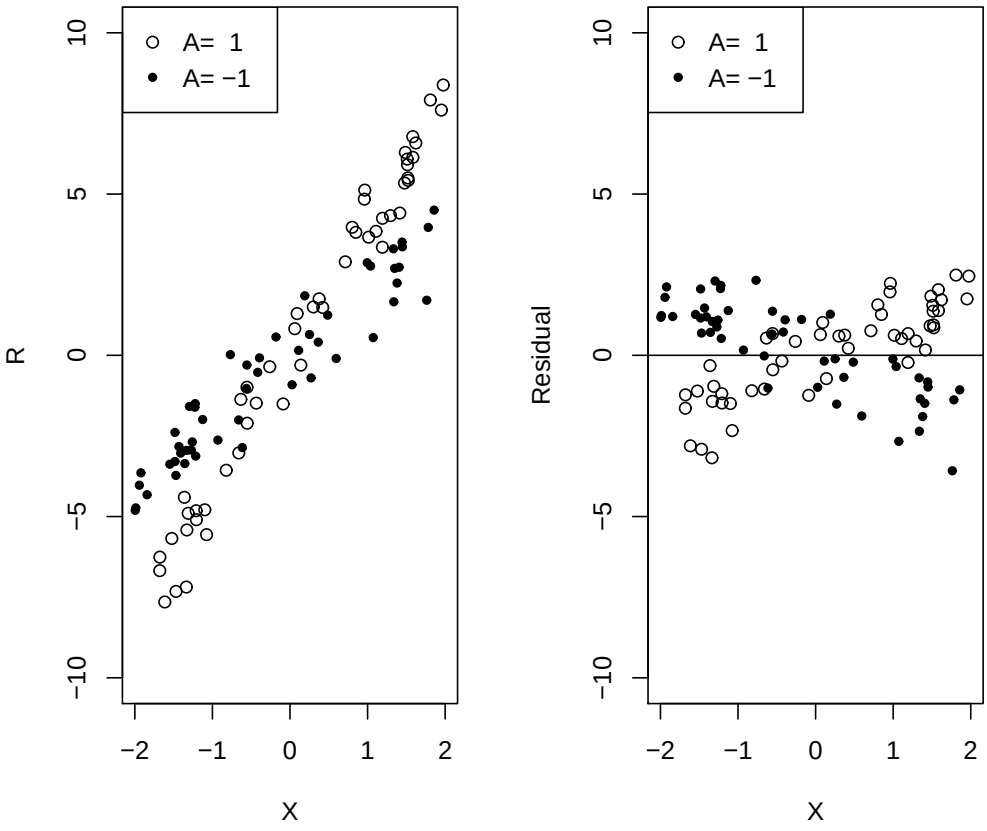


Figure 3.1: *Example of residual weighted learning. The raw data are shown on the left, consisting of a single covariate X, treatment assignment $A = 1$ or -1, and continuous outcome R with $E(R|X, A) = 3X + XA$. The allocation ratio is 1:1. The residuals, $R - 3X$, are shown on the right.*

estimator of (3.1) is,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi(a_i,\boldsymbol{x}_i)}\mathbb{I}\Big(d(\boldsymbol{x}_i)=a_i\Big). \tag{3.7}$$

Though it is unbiased, it may give an estimate outside the range of R. While for finite samples a better estimator of (3.1) as shown in Murphy (2005) is,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{r_i\mathbb{I}\left(d(\boldsymbol{x}_i)=a_i\right)}{\pi(a_i,\boldsymbol{x}_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}\left(d(\boldsymbol{x}_i)=a_i\right)}{\pi(a_i,\boldsymbol{x}_i)}}. \tag{3.8}$$

The denominator is an estimator of $\mathbb{E}\Big(\frac{\mathbb{I}\left(A=d(\boldsymbol{X})\right)}{\pi(A,\boldsymbol{X})}\Big)$. Similar reasoning as that used in the

proof of Lemma 3.2.1 yields $\mathbb{E}\Big(\frac{\mathbb{I}\Big(A=d(\boldsymbol{X})\Big)}{\pi(A,\boldsymbol{X})}\Big) = 1$. The estimator $\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}\Big(d(\boldsymbol{x}_i)=a_i\Big)}{\pi(a_i,\boldsymbol{x}_i)}$ is called the treatment matching factor in this work. The factor varies between 0 and 2. If a rule favors treatment assignments that subjects have actually received, its treatment matching factor is greater than 1, while in contrast if a rule prefers the opposite treatment assignments, its treatment matching factor is less than 1. For a randomized clinical trial, we expect that the estimated rule is associated with a treatment matching factor close to 1. For OWL, since all the weights are nonnegative, the estimated rule of OWL tends to keep, if possible, the treatments that subjects actually received. Thus the associated treatment matching factor would be greater than 1, especially when the sample size is small, or when a complicated rule is applied. Hence the estimator in (3.8) might not be large even though (3.7) is maximized. RWL alleviates this problem by using residuals to make an initial guess on the optimal rule. Owing to the balance between subjects with positive and negative residuals, RWL implicitly finds a rule with its treatment matching factor close to 1.

There are many ways to estimate g^*. In this work, we consider two models. The first one is the main effects model. Assume that $\mathbb{E}\left(\frac{R}{2\pi(A,\boldsymbol{X})}|\boldsymbol{X}\right) = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$. The estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\beta}_0$ can be obtained by minimizing the sum of weighted squares,

$$\sum_{i=1}^{n}\frac{1}{2\pi(a_i,\boldsymbol{x}_i)}(r_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

It can be solved easily by almost any statistical software. Then the estimate is $\widehat{g}^*(\boldsymbol{x}) = \widehat{\beta}_0 + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}$. The second is the null model. Assume that $\mathbb{E}\left(\frac{R}{2\pi(A,\boldsymbol{X})}|\boldsymbol{X}\right) = \beta_0$. It is easy to obtain $\widehat{g}^*(\boldsymbol{x}) = \widehat{\beta}_0 = \sum_{i=1}^{n}\frac{r_i}{\pi(a_i,\boldsymbol{x}_i)}/\sum_{i=1}^{n}\frac{1}{\pi(a_i,\boldsymbol{x}_i)}$.

In summary, we propose a method called Residual Weighted Learning to identify the optimal ITR by minimizing the residual weighted classification error (3.6). The impact of using residuals is two-fold: it stabilizes the variance of the value function and controls the treatment matching factor. Both improve finite sample performance.

3.2.3 Implementation of RWL

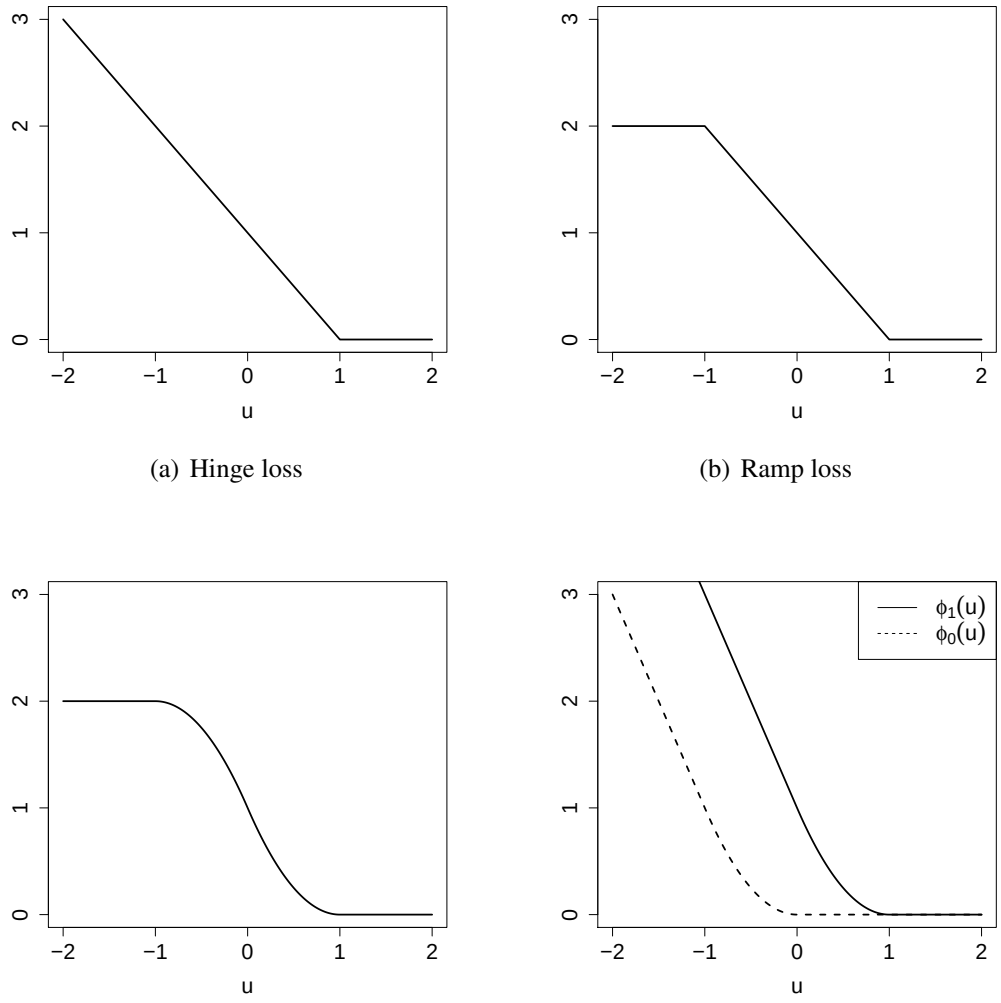
As in OWL, we intend to use a surrogate loss function to replace the 0-1 loss in (3.6). Since some residuals are negative, convex surrogate functions do not work here. We consider a non-convex loss

$$T(u) = \begin{cases} 0 & \text{if } u \geq 1, \\ (1-u)^2 & \text{if } 0 \leq u < 1, \\ 2-(1+u)^2 & \text{if } -1 \leq u < 0, \\ 2 & \text{if } u < -1. \end{cases}$$

It is called the smoothed ramp loss in this work. Figure 3.2 shows the hinge loss, ramp loss and smoothed ramp loss functions. The hinge loss is the loss function used in support vector machines (Vapnik 1998). The ramp loss (Collobert et al. 2006) is also called the truncated hinge loss function (Wu and Liu 2007). It is well known that by truncating the unbounded hinge loss, ramp loss related methods are shown to be robust to outliers in the training data for the classification problem (Wu and Liu 2007). Compared with the ramp loss, the smoothed ramp loss is smooth everywhere. Hence it has computational advantages in optimization. Moreover, the smoothed ramp loss, which resembles the ramp loss, is robust to outliers too. In the framework of RWL, subjects who did not receive optimal treatment assignments but had large positive residuals, or those who did receive optimal assignments but had large negative residuals may be considered as outliers. The robustness to outliers for the smoothed ramp loss is helpful in dealing with outliers in the setting of optimal treatment regimes, especially when residuals are poorly estimated, or when the outcome R has a large variance.

We incorporate RWL into the regularization framework, and aim to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{r}_i}{\pi(a_i,\boldsymbol{x}_i)}T(a_i f(\boldsymbol{x}_i)) + \frac{\lambda}{2}||f||^2, \tag{3.9}$$



(a) Hinge loss

(b) Ramp loss

(c) Smoothed ramp loss

(d) Decomposition

Figure 3.2: *Hinge loss (a), ramp loss (b), and smoothed ramp loss (c) functions. (d) shows the difference of convex decomposition of the smoothed ramp loss,* $T(u) = \phi_1(u) - \phi_0(u)$.

where $||f||$ is some norm for f, and λ is a tuning parameter. Recall that $\widehat{r}_i$ is the estimated residual. The smoothed ramp loss function $T(u)$ is symmetric about the point $(0, 1)$ as shown in Figure 3.2(c). A nice property that comes from the symmetry is that the rule that minimizes the outcome R (*i.e.* maximizes $-R$) is just opposite to the rule that maximizes R. This is intuitively sensible. However, OWL does not possess this property.

We derive an algorithm for linear RWL in Section 3.2.3, and then generalize it to the

case of nonlinear learning through kernel mapping in Section 3.2.3.

Linear Decision Rule for Optimal ITR

Suppose that the decision function $f(\boldsymbol{x})$ that minimizes (3.9) is a linear function of $\boldsymbol{x}$, *i.e.* $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$. Then the associated ITR will assign a subject with clinical covariates $\boldsymbol{x}$ into treatment 1 if $\boldsymbol{w}^T\boldsymbol{x} + b > 0$ and -1 otherwise. In (3.9), we define $||f||$ as the Euclidean norm of $\boldsymbol{w}$. Then minimizing (3.9) can be rewritten as

$$\min_{\boldsymbol{w},b} \quad \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{r}_i}{\pi(a_i,\boldsymbol{x}_i)}T\big(a_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\big). \tag{3.10}$$

The smoothed ramp loss is a non-convex loss, and as a result, the optimization problem in (3.10) involves non-convex minimization. This optimization problem is difficult since there are many local minima or stationary points. For instance, any $(\boldsymbol{w}, b)$ with $\boldsymbol{w} = \boldsymbol{0}$ and $|b| \geq 1$ is a stationary point. Similar with the robust truncated hinge loss support vector machine (Wu and Liu 2007), we apply the d.c. (Difference of Convex) algorithm (An and Tao 1997) to solve this non-convex minimization problem. The d.c. algorithm is also known as the Concave-Convex Procedure (CCCP) in the machine learning community (Yuille and Rangarajan 2003). Assume that an objective function can be rewritten as the sum of a convex part $Q_{vex}(\Theta)$ and a concave part $Q_{cav}(\Theta)$. The d.c. algorithm as shown in Algorithm 1 solves the non-convex optimization problem by minimizing a sequence of convex subproblems. One can easily see that the d.c. algorithm is a special case of the Majorize-Minimization (MM) algorithm.

Algorithm 1: The d.c. algorithm for minimizing $Q(\Theta) = Q_{vex}(\Theta) + Q_{cav}(\Theta)$.

Initialize $\Theta^{(0)}$
repeat
  $\Theta^{(t+1)} = \text{argmin}_{\Theta} Q_{vex}(\Theta) + < Q'_{cav}(\Theta^{(t)}), \Theta >$
until *convergence of* $\Theta^{(t)}$

Let

$$\phi_s(u) = \begin{cases} 0 & \text{if } u \geq s \\ (s-u)^2 & \text{if } s-1 \leq u < s \\ 2s-2u-1 & \text{if } u < s-1 \end{cases} .$$

Note that ϕ_s is smooth. We have a difference-of-convex decomposition of the smoothed ramp loss,

$$T(u) = \phi_1(u) - \phi_0(u), \tag{3.11}$$

as shown in Figure 3.2(d). Denote Θ as $(\boldsymbol{w}, b)$. Applying (3.11), the objective function in (3.10) can be decomposed as

$$\begin{aligned} Q^s(\Theta) \;=\; & \underbrace{\frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i > 0) + \phi_0(u_i)\mathbb{I}(\widehat{r}_i < 0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i, \boldsymbol{x}_i)}}_{Q_{vex}(\Theta)} \\ & \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i < 0) + \phi_0(u_i)\mathbb{I}(\widehat{r}_i > 0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i, \boldsymbol{x}_i)}}_{Q_{cav}(\Theta)}, \end{aligned}$$

where $u_i = a_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$. For simplicity, we introduce the notation,

$$\beta_i = \frac{\partial Q_{cav}}{\partial u_i} = -\frac{1}{n}\Big[\frac{d\phi_1(u_i)}{du_i}\mathbb{I}(\widehat{r}_i < 0) + \frac{d\phi_0(u_i)}{du_i}\mathbb{I}(\widehat{r}_i > 0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i, \boldsymbol{x}_i)}, \tag{3.12}$$

for $i = 1, \cdots, n$. Thus the convex subproblem at the $(t+1)$'th iteration of the d.c. algorithm is

$$\min_{\boldsymbol{w}, b} \quad \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i > 0) + \phi_0(u_i)\mathbb{I}(\widehat{r}_i < 0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i, \boldsymbol{x}_i)} + \sum_{i=1}^{n}\beta_i^{(t)}u_i. \tag{3.13}$$

There are many efficient methods available for solving smooth unconstrained optimization problems. In this work, we use the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Nocedal 1980). L-BFGS is a quasi-Newton method that approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using a limited amount of computer memory. For more information on the quasi-Newton method and L-BFGS, see Nocedal and Wright (2006).

The procedure is summarized in the following algorithm:

Algorithm 2: The d.c. algorithm for linear RWL with the smoothed ramp loss

Set ϵ to a small quantity, say, 10^{-8}

Initialize $\boldsymbol{\beta}_i^{(0)} = \frac{2|\widehat{r}_i|}{n\pi(a_i,\boldsymbol{x}_i)}\mathbb{I}(\widehat{r}_i < 0)$

repeat

- Compute $(\widehat{\boldsymbol{w}}, \widehat{b})$ by solving (3.13),
- Update $u_i = a_i(\widehat{\boldsymbol{w}}^T\boldsymbol{x}_i + \widehat{b})$
- Update $\beta_i^{(t+1)}$ by (3.12)

until $||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}||_\infty < \epsilon$

Nonlinear Decision rule for Optimal ITR

For a nonlinear decision rule, the decision function $f(\boldsymbol{x})$ is represented by $h(\boldsymbol{x}) + b$ with $h(\boldsymbol{x}) \in \mathcal{H}_K$ and $b \in \mathbb{R}$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel function K. The kernel function $K(\cdot,\cdot)$ is a positive definite function mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. The norm in $\mathcal{H}_K$, denoted by $||\cdot||_K$, is induced by the following inner product:

$$< f, g >_K = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i\beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

for $f(\cdot) = \sum_{i=1}^n \alpha_i K(\cdot, \boldsymbol{x}_i)$ and $g(\cdot) = \sum_{j=1}^m \beta_j K(\cdot, \boldsymbol{x}_j)$. Then minimizing (3.9) can be rewritten as

$$\min_{h,b} \quad \frac{\lambda}{2}||h||_K^2 + \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{r}_i}{\pi(a_i,\boldsymbol{x}_i)} T\Big(a_i\big(h(\boldsymbol{x}_i) + b\big)\Big). \tag{3.14}$$

Due to the representer theorem (Kimeldorf and Wahba 1971), the nonlinear problem can be reduced to finding finite-dimensional coefficients v_i, and $h(\boldsymbol{x})$ can be represented as $\sum_{j=1}^n v_j K(\boldsymbol{x}, \boldsymbol{x}_j)$. Thus the problem (3.14) is transformed to

$$\min_{\boldsymbol{v},b} \quad \frac{\lambda}{2}\sum_{i,j=1}^{n} v_i v_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{r}_i}{\pi(a_i,\boldsymbol{x}_i)} T\Big(a_i\big(\sum_{j=1}^{n} v_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b\big)\Big). \tag{3.15}$$

Following a similar derivation to that used in the previous section, the convex subproblem at the $(t+1)$'th iteration of the d.c. algorithm is as follows,

$$\begin{aligned}\min_{\boldsymbol{v},b} \quad & \frac{\lambda}{2}\sum_{i,j=1}^{n} v_i v_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ & + \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i>0)+\phi_0(u_i)\mathbb{I}(\widehat{r}_i<0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i,\boldsymbol{x}_i)} + \sum_{i=1}^{n}\beta_i^{(t)}u_i,\end{aligned}$$

where $u_i = a_i\big(\sum_{j=1}^{n} v_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b\big)$. After solving the subproblem by L-BFGS, we update $\beta^{(t+1)}$ by (3.12). The procedure is repeated until $\boldsymbol{\beta}$ converges. When we obtain the solution $(\widehat{\boldsymbol{v}},\widehat{b})$, the decision function is $\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{n}\widehat{v}_j K(\boldsymbol{x},\boldsymbol{x}_j)+\widehat{b}$. Note that if we choose a linear kernel $K(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}^T\boldsymbol{z}$, the obtained rule reduces to the previous linear rule. The most widely used nonlinear kernel in practice is the Gaussian Radial Basis Function (RBF) kernel, that is,

$$K_\sigma(\boldsymbol{x},\boldsymbol{z}) = \exp\Big(-\sigma^2||\boldsymbol{x}-\boldsymbol{z}||^2\Big),$$

where $\sigma>0$ is a free parameter whose inverse $1/\sigma$ is called the width of K_σ.

3.2.4 A general framework for Residual Weighted Learning

We have proposed a method to identify the optimal ITR for continuous outcomes. However, in clinical practice, the endpoint outcome could also be binary, count, or rate. In this section, we provide a general framework to deal with all these types of outcomes.

The procedure is described as follows. First, estimate the conditional expected outcomes given clinical covariates $\boldsymbol{X}$, $\widehat{\mathbb{E}}\left(\frac{R}{2\pi(A,\boldsymbol{X})}\big|\boldsymbol{X}\right)$, by an appropriate regression model. It is equivalent to fitting a weighted regression model, in which each subject is weighted by $\frac{1}{2\pi(A,\boldsymbol{X})}$. Second, we calculate the estimated residual $\widehat{r}_i$ by comparing the observed outcome and expected outcome estimated in the first step. If the observed outcome is better than the expected one, the estimated residual $\widehat{r}_i$ is positive, otherwise it is negative. Specifically, if larger values of R are more desirable, as with our assumption for continu-

ous outcomes, $\widehat{r}_i = r_i - \widehat{\mathbb{E}}\left(\frac{R}{2\pi(A,\boldsymbol{X})}\big|\boldsymbol{X}=\boldsymbol{x}_i\right)$; if smaller values of R are preferred, *e.g.* the number of adverse events, then $\widehat{r}_i = \widehat{\mathbb{E}}\left(\frac{R}{2\pi(A,\boldsymbol{X})}\big|\boldsymbol{X}=\boldsymbol{x}_i\right) - r_i$. Third, identify the decision function $\widehat{f}(\boldsymbol{x})$ by using the estimated $\widehat{r}_i$ in RWL.

The underlying idea is simple. Generally, the outcome R is a random variable depending on $\boldsymbol{X}$ and A. We estimate common effects of $\boldsymbol{X}$ by ignoring the treatment assignment A, and then all information on the heterogeneous treatment effects is contained in the residuals. As demonstrated previously, the benefits from using residuals include stabilizing the variance and controlling the treatment matching factor, both of which have a positive impact on finite sample performance. In previous sections, we discussed RWL for continuous outcomes in detail. We will provide two additional examples to illustrate the general framework.

The first example is for binary outcomes, *i.e.* $R \in \{0,1\}$. Assume that $R = 1$ is desirable. We may fit a weighted main effects logistic regression model,

$$\mathbb{E}(R|\boldsymbol{X}) = \frac{\exp(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta})}{1+\exp(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta})}.$$

The estimates $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ can be obtained numerically by statistical software, for example, the glm function in R. The residual $\widehat{r}_i$ is $r_i - \frac{\exp(\widehat{\beta}_0+\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})}{1+\exp(\widehat{\beta}_0+\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})}$. Then we estimate the optimal ITR by RWL.

The second is for count/rate outcomes. We use the pulmonary exacerbation (PE) outcome of the cystic fibrosis data in Section 3.6 as an example. The outcome is the number of PEs during the study, and it is a rate variable. The observed data are $D_n = \{\boldsymbol{x}_i, a_i, r_i, t_i\}_{i=1}^n$, where r_i is the number of PEs in the duration t_i. We treat the count r_i as the outcome, and $(\boldsymbol{x}_i, t_i)$ as clinical covariates. We may fit a weighted main effects Poisson regression model,

$$\log\left(\mathbb{E}(R|\boldsymbol{X},T)\right) = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta} + \log(T).$$

Since fewer PEs are desirable, we compute $\widehat{r}_i$ as $\exp\left(\widehat{\beta}_0 + \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} + \log(t_i)\right) - r_i$, which is the opposite of the raw residual in generalized linear models.

3.3 Theoretical Properties

In this section, we establish theoretical results for RWL. To the end, we assume that a sample $D_n = \{\boldsymbol{x}_i, a_i, r_i\}_{i=1}^n$, is independently drawn from a probability measure P on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$, where $\mathcal{X} \subset \mathbb{R}^p$ is compact, and $\mathcal{M} = [-M, M] \subset \mathbb{R}$. For any ITR $d : \mathcal{X} \to \mathcal{A}$, the risk is defined as

$$\mathcal{R}(d) = \mathbb{E}\left[\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}(A \neq d(\boldsymbol{X}))\right].$$

The ITR that minimizes the risk is the Bayes rule $d^* = \arg\min_d \mathcal{R}(d)$, and the corresponding risk $\mathcal{R}^* = \mathcal{R}(d^*)$ is the Bayes risk. Recall that the Bayes rule is $d^*(\boldsymbol{x}) = sign(\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1))$.

In RWL, we substitute the 0-1 loss $\mathbb{I}(A \neq sign(f(\boldsymbol{X})))$ by the smoothed ramp loss, $T(Af(\boldsymbol{X}))$. Accordingly, we define the T-risk as

$$\mathcal{R}_{T,g}(f) = \mathbb{E}\left[\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})}T(Af(\boldsymbol{X}))\right],$$

and, similarly, the minimal T-risk as $\mathcal{R}^*_{T,g} = \inf_f \mathcal{R}_{T,g}(f)$ and $f^*_{T,g} = \arg\min_f \mathcal{R}_{T,g}(f)$. In the theoretical analysis, we do not require g to be a regression fit of R. g can be any arbitrary measurable function. We may further assume that $(R - g(\boldsymbol{X}))/\pi(A, \boldsymbol{X})$ is bounded almost surely.

The performance of an ITR associated with a real-valued function f is measured by the excess risk $\mathcal{R}(sign(f)) - \mathcal{R}^*$. In terms of the value function, we note that the excess risk is just $\mathcal{V}(d^*) - \mathcal{V}(sign(f))$. Similarly, we define the excess T-risk as $\mathcal{R}_{T,g}(f) - \mathcal{R}^*_{T,g}$.

Let $f_{D_n,\lambda_n} \in \mathcal{H}_K + \{1\}$, *i.e.* $f_{D_n,\lambda_n} = h_{D_n,\lambda_n} + b_{D_n,\lambda_n}$ where $h_{D_n,\lambda_n} \in \mathcal{H}_K$ and $b_{D_n,\lambda_n} \in \mathbb{R}$, be a global minimizer of the following optimization problem:

$$\min_{f=h+b\in\mathcal{H}_K+\{1\}} \quad \frac{\lambda_n}{2}||h||_K^2 + \frac{1}{n}\sum_{i=1}^{n} \frac{r_i - g(\boldsymbol{x}_i)}{\pi(a_i, \boldsymbol{x}_i)} T\Big(a_i f(\boldsymbol{x}_i))\Big).$$

Here we suppress g from the notation of f_{D_n,λ_n}, h_{D_n,λ_n} and b_{D_n,λ_n}.

The purpose of the theoretical analysis is to estimate the excess risk $\mathcal{R}(sign(f_{D_n,\lambda_n})) - \mathcal{R}^*$ as the sample size n tends to infinity. Convergence rates will be derived under the choice of the parameter λ_n and conditions on the distribution P.

3.3.1 Fisher Consistency

We establish Fisher consistency of the decision function based on the smoothed ramp loss. Specifically, the following result holds:

Theorem 3.3.1. *For any measurable function f, if $f^*_{T,g}$ minimizes T-risk $\mathcal{R}_{T,g}(f)$, then the Bayes rule $d^*(\boldsymbol{x}) = sign(f^*_{T,g}(\boldsymbol{x}))$ for all $\boldsymbol{x}$ such that $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) \neq \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$. Furthermore, $\mathcal{R}_{T,g}(d^*) = \mathcal{R}^*_{T,g}$.*

The proof is provided in Appendix. This theorem shows the validity of using the smoothed ramp loss as a surrogate loss in RWL.

3.3.2 Excess Risk

The following result establishes the relationship between the excess risk and excess T-risk. The proof can be found in Appendix.

Theorem 3.3.2. *For any measurable $f : \mathcal{X} \to \mathbb{R}$ and any probability distribution for $(\boldsymbol{X}, A, R)$,*

$$\mathcal{R}(sign(f)) - \mathcal{R}^* \leq \mathcal{R}_{T,g}(f) - \mathcal{R}^*_{T,g}.$$

This theorem shows that for any decision function f, the excess risk of f under the 0-1 loss is no larger than the excess risk of f under the smoothed ramp loss. It implies that if we obtain f by approximately minimizing $\mathcal{R}_{T,g}(f)$ so that the excess T-risk is small, then the risk of f is close to the Bayes risk. In the next two sections, we investigate properties of f_{D_n,λ_n} through the excess T-risk.

3.3.3 Universal Consistency

In the literature of machine learning, a classification rule is called universally consistent if its expected error probability converges in probability to the Bayes error probability for all distributions underlying the data (Devroye et al. 1996). We extend this concept to ITRs. Precisely, given a sequence D_n of training data, an ITR d_n is said to be *universally consistent* if

$$\lim_{n\to\infty} \mathcal{R}(d_n) = \mathcal{R}^*$$

holds in probability for all probability measures P on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$. In this section, we will establish universal consistency of the rule $d_{D_n,\lambda_n} = sign(f_{D_n,\lambda_n})$. Before proceeding to the main result, we introduce some preliminary background on RKHSs, which is essential for our work.

Let K be a kernel, and $\mathcal{H}_K$ be its associated RKHS. We often use the quantity

$$C_K = \sup_{\boldsymbol{x}\in\mathcal{X}} \sqrt{K(\boldsymbol{x},\boldsymbol{x})}.$$

In this work, we assume C_K is finite. For the Gaussian RBF kernel, $C_K = 1$. The assumption also holds for the linear kernel when $\mathcal{X} \subset \mathbb{R}^p$ is compact. By the reproducing property, it is easy to see that $||f||_\infty \leq C_K||f||_K$ for any $f \in \mathcal{H}_K$. A continuous kernel K on a compact metric space $\mathcal{X}$ is called *universal* if its associated RKHS $\mathcal{H}_K$ is dense in $C(\mathcal{X})$, *i.e.*, for every function $g \in C(\mathcal{X})$ and all $\epsilon > 0$ there exists an $f \in \mathcal{H}_K$ such that $||f - g||_\infty < \epsilon$ (Steinwart and Christmann 2008, Definition 4.52). $C(\mathcal{X})$ is the space

of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ on the compact metric space $\mathcal{X}$ endowed with the usual supremum norm. The widely used Gaussian RBF kernel is an example of a universal kernel.

We are ready to present the main result of this section. The following theorem shows the convergence of the T-risk on the sample dependent function f_{D_n,λ_n}. We apply empirical process techniques to show consistency. The proof is provided in Appendix.

Theorem 3.3.3. *Assume that we choose a sequence $\lambda_n > 0$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$. For a measurable function g, assume that $\frac{|R-g(\boldsymbol{X})|}{\pi(A,\boldsymbol{X})} \leq M_0$ almost surely. Then for any distribution P for $(\boldsymbol{X}, A, R)$, we have that in probability,*

$$\lim_{n\to\infty} \mathcal{R}_{T,g}(f_{D_n,\lambda_n}) = \inf_{f\in\mathcal{H}_K+\{1\}} \mathcal{R}_{T,g}(f).$$

By Theorem 3.3.1, the Bayes rule d^* minimizes the T-risk $\mathcal{R}_{T,g}(f)$. Universal consistency follows if $\inf_{f\in\mathcal{H}_K+\{1\}} \mathcal{R}_{T,g}(f) = \mathcal{R}^*_{T,g}$ by Theorem 3.3.2. That is, d^* needs to be approximated by functions in the space $H_K + \{1\}$. Clearly, this is not always true. A counterexample is when K is linear.

Let μ be the marginal distribution of $\boldsymbol{X}$. Clearly, the rule d^* is measurable with respect to μ. Recall that a probability measure is regular if it is defined on the Borel sets. By Lusin's theorem, a measurable function can be approximated by a continuous function when μ is regular (Rudin 1987, Theorem 2.24). Thus the RKHS $\mathcal{H}_K$ of a universal kernel K is rich enough to provide arbitrarily accurate decision rules for all distributions. The finding is summarized in the following universal approximation lemma. The rigorous proof is provided in Appendix.

Lemma 3.3.1. *Let K be a universal kernel, and $\mathcal{H}_K$ be the associated RKHS. For all distributions P with regular marginal distribution μ on $\boldsymbol{X}$, we have*

$$\inf_{f\in\mathcal{H}_K+\{1\}} \mathcal{R}_{T,g}(f) = \mathcal{R}^*_{T,g}.$$

This immediately leads to universal consistency of RWL with a universal kernel.

Proposition 3.3.1. *Let K be a universal kernel, and $\mathcal{H}_K$ be the associated RKHS. Assume that we choose a sequence $\lambda_n > 0$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$. For a measurable function g, assume that $\frac{|R-g(\boldsymbol{X})|}{\pi(A,\boldsymbol{X})} \leq M_0$ almost surely. Then for any distribution P for $(\boldsymbol{X}, A, R)$ with regular marginal distribution on $\boldsymbol{X}$, we have that in probability,*

$$\lim_{n\to\infty} \mathcal{R}(sign(f_{D_n,\lambda_n})) = \mathcal{R}^*.$$

3.3.4 Convergence Rate

A consistent rule guarantees that by increasing the amount of data the rule can eventually learn the optimal decision with high probability. The next natural question is whether there are rules d_n with $\mathcal{R}(d_n)$ tending to the Bayes risk $\mathcal{R}^*$ at a specified rate for all distributions. Unfortunately, this is impossible as shown in the following theorem. The proof follows the arguments in Devroye et al. (1996, Theorem 7.2), which states in the classification problem the rate of convergence of any particular rule to the Bayes risk may be arbitrarily slow. We provide the outline of the proof in Appendix.

Theorem 3.3.4. *Assume there are infinite distinct points in $\mathcal{X}$. Let $\{c_n\}$ be a sequence of positive numbers converging to zero with $\frac{1}{16} \geq c_1 \geq c_2 \geq \cdots$. For every sequence d_n of ITRs, there exists a distribution of $(\boldsymbol{X}, A, R)$ on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$ such that*

$$\mathcal{R}(d_n) - \mathcal{R}^* \geq 2Mc_n.$$

This is a generic result, and applies to any algorithm for ITRs. It implies that rate of convergence studies for particular rules must necessarily be accompanied by conditions on $(\boldsymbol{X}, A, R)$. Before moving back to RWL, we introduce the following quantity:

$$\mathcal{A}(\lambda) = \inf_{h\in\mathcal{H}_K, b\in\mathbb{R}} \frac{\lambda}{2}||h||_K^2 + \mathcal{R}_{T,g}(h+b) - \mathcal{R}_{T,g}^*.$$

The term $\mathcal{A}(\lambda)$ describes how well the infinite-sample regularized T-risk $\frac{\lambda}{2}||h||_K^2+\mathcal{R}_{T,g}(h+b)$ approximates the optimal T-risk $\mathcal{R}^*_{T,g}$. A similar quantity is called the *approximation error function* in the literature of learning theory (Steinwart and Christmann 2008, Definition 5.14). Since $\mathcal{A}(\lambda)$ is an infimum of affine linear functions, $\mathcal{A}(\lambda)$ is continuous. Thus

$$\lim_{\lambda \to 0} \mathcal{A}(\lambda) = \inf_{h \in \mathcal{H}_K, b \in \mathbb{R}} \mathcal{R}_{T,g}(h+b) - \mathcal{R}^*_{T,g}.$$

By Lemma 3.3.1, for a universal kernel K, $\lim_{\lambda \to 0} \mathcal{A}(\lambda) = 0$.

In order to establish the convergence rate, let us additionally assume that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that

$$\mathcal{A}(\lambda) \leq c\lambda^\beta, \quad \forall \lambda > 0. \tag{3.16}$$

The assumption is standard in the literature of learning theory (Steinwart and Christmann 2008, p. 229). The following theorem is the main result in this section. This theorem is proved in Appendix using techniques of concentration inequalities developed in Bartlett and Mendelson (2002).

Theorem 3.3.5. *For any distribution P for $(\boldsymbol{X}, A, R)$ that satisfies the approximation error assumption (3.16), take $\lambda_n = n^{-\frac{1}{2\beta+1}}$. For a measurable function g, assume that $\frac{|R-g(\boldsymbol{X})|}{\pi(A,\boldsymbol{X})} \leq M_0$ almost surely. Then with probability at least $1-\delta$,*

$$\mathcal{R}(sign(f_{D_n,\lambda_n})) - \mathcal{R}^* \leq \widetilde{c}\sqrt{\log(4/\delta)} n^{-\frac{\beta}{2\beta+1}},$$

where the constant $\widetilde{c}$ is independent of δ and n, and $\widetilde{c}$ decreases as M_0 decreases.

The resulting rate depends on the polynomial decay rate β of the approximation error, and the optimal rate is about $n^{-1/3}$ when β approaches 1. Theorem 3.3.5 also implies that even when the sample size n is fixed, an appropriately chosen g with a small bound M_0 would improve performance. It is easy to verify that g^* in (3.5) is the function g that minimizes $\mathbb{E}\left(\frac{(R-g(\boldsymbol{X}))^2}{\pi(A,\boldsymbol{X})}\right)$. So when we choose g^* as in RWL, we may possibly obtain a smaller M_0, as shown in Figure 3.1.

We are particularly interested in the Gaussian RBF kernel since it is widely used and is universal. It is important to understand the approximation error assumption (3.16) for the Gaussian RBF kernel. We first introduce the following "geometric noise" assumption for P on $(\boldsymbol{X}, A, R)$ (Steinwart and Scovel 2007). Let

$$\delta(\boldsymbol{x}) = \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1).$$

Define $\mathcal{X}^+ = \{\boldsymbol{x} \in \mathcal{X} : \delta(\boldsymbol{x}) > 0\}$, $\mathcal{X}^- = \{\boldsymbol{x} \in \mathcal{X} : \delta(\boldsymbol{x}) < 0\}$, and $\mathcal{X}^0 = \{\boldsymbol{x} \in \mathcal{X} : \delta(\boldsymbol{x}) = 0\}$. Now we define a distance function $\boldsymbol{x} \mapsto \tau_{\boldsymbol{x}}$ by

$$\tau_{\boldsymbol{x}} = \begin{cases} \widetilde{d}(\boldsymbol{x}, \mathcal{X}^0 \cup \mathcal{X}^+) & \text{if} \quad \boldsymbol{x} \in \mathcal{X}^-, \\ \widetilde{d}(\boldsymbol{x}, \mathcal{X}^0 \cup \mathcal{X}^-) & \text{if} \quad \boldsymbol{x} \in \mathcal{X}^+, \\ 0 & \text{if} \quad \boldsymbol{x} \in \mathcal{X}^0, \end{cases}$$

where $\widetilde{d}(\boldsymbol{x}, A)$ denotes the distance of $\boldsymbol{x}$ to a set A with respect to the Euclidean norm. Now we present the geometric noise condition for distributions.

Definition 3.3.1. *Let $\mathcal{X} \subset \mathbb{R}^p$ be compact and P be a probability measure of $(\boldsymbol{X}, A, R)$ on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$. We say that P has geometric noise exponent $q > 0$ if there exists a constant $C > 0$ such that*

$$\int_{\mathcal{X}} exp\left(-\frac{\tau_{\boldsymbol{x}}^2}{t}\right) |\delta(\boldsymbol{x})| \mu(d\boldsymbol{x}) \leq C t^{qp/2}, \quad t > 0, \tag{3.17}$$

where μ is the marginal measure on $\boldsymbol{X}$.

The integral condition (3.17) describes the concentration of the measure $|\delta(\boldsymbol{x})| d\mu$ near the decision boundary in the sense that the less the measure is concentrated in this region the larger the geometric noise exponent can be chosen. The following lemma shows that the geometric noise condition can be used to guarantee the approximation error assumption (3.16) when the parameter σ of the Gaussian RBF kernel K_σ is appropriately chosen.

Lemma 3.3.2. *Let $\mathcal{X}$ be the closed unit ball of the Euclidean space $\mathbb{R}^p$, and P be a distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$ that has geometric noise exponent $0 < q < \infty$ with constant C in (3.17). Let $\sigma = \lambda^{-\frac{1}{(q+1)p}}$. Then for the Gaussian RBF kernel K_σ, there is a constant $c > 0$ depending only on the dimension p, the geometric noise exponent q and constant C, such that for all $\lambda > 0$ we have*

$$\mathcal{A}(\lambda) \leq c\lambda^{q/(q+1)}.$$

The proof provided in Appendix follows the idea from Theorem 2.7 in Steinwart and Scovel (2007) by replacing their $2\eta(\boldsymbol{x}) - 1$ with $\delta(\boldsymbol{x})$. The key point is the following inequality, for all measurable f,

$$\mathcal{R}_{T,g}(f) - \mathcal{R}^*_{T,g} \leq 2\mathbb{E}(|\delta(\boldsymbol{x})| \cdot |f - d^*|),$$

which is the counterpart of Equation (24) in Steinwart and Scovel (2007).

Now we are ready to present the convergence rate of RWL with Gaussian RBF kernel by combining Theorem 3.3.5 and Lemma 3.3.2:

Proposition 3.3.2. *Let $\mathcal{X}$ be the closed unit ball of the Euclidean space $\mathbb{R}^p$, and P be a distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$ that has geometric noise exponent $0 < q < \infty$ with constant C in (3.17). For a measurable function g, assume that $\frac{|R-g(\boldsymbol{X})|}{\pi(A,\boldsymbol{X})} \leq M_0$ almost surely. Consider a Gaussian RBF kernel K_{σ_n}. Take $\lambda_n = n^{-\frac{q}{3q+1}}$ and $\sigma_n = \lambda_n^{-\frac{1}{(q+1)p}}$. Then with probability at least $1 - \delta$,*

$$\mathcal{R}(sign(f_{D_n,\lambda_n})) - \mathcal{R}^* \leq \widetilde{c}\sqrt{\log(4/\delta)}n^{-\frac{q}{3q+1}},$$

where the constant $\widetilde{c}$ is independent of δ and n, and $\widetilde{c}$ decreases as M_0 decreases.

The optimal rate for the risk is about $n^{-1/3}$ when the geometric noise exponent q is sufficiently large. A better rate may be achieved by using techniques in Steinwart and Scovel (2007), but it is beyond the scope of this work.

3.4 Variable selection for RWL

Variable selection is an important area in modern statistical research. Gunter et al. (2011) distinguished prescriptive covariates from predictive covariates. The latter refers to covariates which are related to the prediction of outcomes, and the former refers to covariates which help to prescribe optimal ITRs. Clearly, prescriptive covariates are predictive too. In practice, a large number of clinical covariates are often available for estimating optimal ITRs. However, many of them might not be prescriptive. Thus careful variable selection could lead to better performance.

3.4.1 Variable selection for linear RWL

There is a vast body of literature on variable selection for linear classification in statistical learning. For instance, Zhu et al. (2003) and Fung and Mangasarian (2004) investigated the support vector machine (SVM) using the ℓ_1-norm penalty (Tibshirani 1994). Wang et al. (2008) applied the elastic-net penalty (Zou and Hastie 2005) for variable selection with an SVM-like method. In this work, we replace the ℓ_2-norm penalty in RWL with the elastic-net penalty,

$$\lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w},$$

where $||\boldsymbol{w}||_1 = \sum_{j=1}^{p} |w_j|$ is the ℓ_1-norm. As a hybrid of the ℓ_1-norm and ℓ_2-norm penalties, the elastic-net penalty retains the variable selection features of the ℓ_1-norm penalty, and tends to provide similar estimated coefficients for highly correlated variables, *i.e.* the grouping effect, as the ℓ_2-norm penalty does. Hence, highly correlated variables are selected or removed together.

The elastic-net penalized linear RWL aims to minimize

$$\lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{r}_i}{\pi(a_i,\boldsymbol{x}_i)}T\big(a_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\big),$$

where $\lambda_1(>0)$ and $\lambda_2(\geq 0)$ are regularization parameters. We still apply the d.c. algorithm to solve the above optimization problem. Following a similar decomposition, the convex subproblem at the $(t+1)$'th iteration of the d.c. algorithm is as follows,

$$\begin{aligned}\min_{\boldsymbol{w},b} \quad & \lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{n}\beta_i^{(t)}u_i \\ & + \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i > 0) + \phi_0(u_i)\mathbb{I}(\widehat{r}_i < 0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i,\boldsymbol{x}_i)}, \end{aligned} \tag{3.18}$$

where $u_i = a_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$, and $\beta_i^{(t)}$ is computed by (3.12). The objective function is a sum of a smooth function and a non-smooth ℓ_1-norm penalty. Since L-BFGS can only deal with smooth objective functions, it may not work here. Instead we use projected scaled sub-gradient (PSS) algorithms (Schmidt 2010), which are extensions of L-BFGS to the case of optimizing a smooth function with an ℓ_1-norm penalty. Several PSS algorithms are proposed in Schmidt (2010). Among them, the Gafni-Bertseka variant is particularly effective for our purpose. After solving the subproblem, we update $\beta_i^{(t+1)}$ by (3.12). The procedure is repeated until $\boldsymbol{\beta}$ converges. The obtained decision function is $\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{w}}^T\boldsymbol{x} + \widehat{b}$, and thus the estimated optimal ITR is the sign of $\widehat{f}(\boldsymbol{x})$.

3.4.2 Variable selection for RWL with nonlinear kernels

Many researchers have noticed that nonlinear kernel methods may perform poorly when there are irrelevant covariates presented in the data (Weston et al. 2000, Grandvalet and Canu 2002, Lin and Zhang 2006, Lafferty and Wasserman 2008). For optimal treatment regimes, we may suffer a similar problem when using a nonlinear kernel. For the Gaussian RBF kernel, we assume that the geometric noise condition in Definition 3.3.1 is satisfied with exponent q. When several additional non-descriptive covariates are included in the data, the geometric noise exponent is decreased according to (3.17), and hence so is the convergence rate in Proposition 3.3.2. The presence of non-descriptive covariates thus can deteriorate the performance of RWL.

The variable selection approach proposed in this section is inspired by the KNIFE algorithm (Allen 2013), where a set of scaling factors on the covariates are employed to form a regularized loss function. The idea of scaling covariates within the kernel has appeared in several methods from the machine learning community (Weston et al. 2000, Grandvalet and Canu 2002, Rakotomamonjy 2003, Argyriou et al. 2006).

We take the Gaussian RBF kernel as an example. Define the covariates-scaled Gaussian RBF kernel,

$$K_{\boldsymbol{\eta}}(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\sum_{j=1}^{p} \eta_j (x_j - z_j)^2\right),$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_p)^T \geq \mathbf{0}$. Here each covariate x_j is scaled by $\sqrt{\eta_j}$. Setting $\eta_j = 0$ is equivalent to discarding the j'th covariate. The hyperparameter σ in the original Gaussian RBF kernel is absorbed to the scaling factors. Similar with KNIFE, we seek $(\widehat{\boldsymbol{v}}, \widehat{b}, \widehat{\boldsymbol{\eta}})$ to minimize the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{v}, b, \boldsymbol{\eta}} \quad & \lambda_1 ||\boldsymbol{\eta}||_1 + \frac{\lambda_2}{2} \sum_{i,j=1}^{n} v_i v_j K_{\boldsymbol{\eta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ & + \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{r}_i}{\pi(a_i, \boldsymbol{x}_i)} T\Big(a_i\big(\sum_{j=1}^{n} v_j K_{\boldsymbol{\eta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + b\big)\Big), \qquad (3.19) \\ \text{subject to} \quad & \boldsymbol{\eta} \geq \mathbf{0}, \end{aligned}$$

where $\lambda_1 (> 0)$ and $\lambda_2 (> 0)$ are regularization parameters. Compared with previous nonlinear RWL optimization problem in (3.15), (3.19) has an ℓ_1-norm penalty on scaling factors. The objective function is singular at $\boldsymbol{\eta} = \mathbf{0}$ due to nonnegativity of scaling factors. So (3.19) may produce zero solutions for some of the $\boldsymbol{\eta}$, and hence performs variable selection.

We adopt a similar decomposition trick that has been used several times before. The

subproblem at the $(t+1)$'th iteration is

$$\begin{aligned} \min_{\boldsymbol{v},b,\boldsymbol{\eta}} \quad & \lambda_1||\boldsymbol{\eta}||_1 + \frac{\lambda_2}{2}\sum_{i,j=1}^{n} v_i v_j K_{\boldsymbol{\eta}}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{i=1}^{n}\beta_i^{(t)}u_i \\ & + \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_1(u_i)\mathbb{I}(\widehat{r}_i>0) + \phi_0(u_i)\mathbb{I}(\widehat{r}_i<0)\Big]\frac{|\widehat{r}_i|}{\pi(a_i,\boldsymbol{x}_i)}, \qquad (3.20) \\ \text{subject to} \quad & \boldsymbol{\eta}\geq 0, \end{aligned}$$

where $u_i = a_i\big(\sum_{j=1}^{n} v_j K_{\boldsymbol{\eta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)+b\big)$, and $\beta_i^{(t)}$ is computed by (3.12). Recall that the d.c. algorithm is a special case of the Majorize-Minimization (MM) algorithm. Although with the covariate-scaled kernel the subproblem is nonconvex, we still apply a similar iterative procedure as in the d.c. algorithm, *i.e.* solving a sequence of subproblems (3.20), but based on the MM algorithm. The subproblem (3.20) is a smooth optimization problem with box constraints. In this work, we use L-BFGS-B (Byrd et al. 1995, Morales and Nocedal 2011), which extends L-BFGS to handle simple box constraints on variables. After solving the subproblem (3.20), we update u_i and $\boldsymbol{\beta}$. The procedure is repeated until convergence. Then the obtained decision function is $\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n}\widehat{v}_i K_{\widehat{\boldsymbol{\eta}}}(\boldsymbol{x}_i,\boldsymbol{x}) + \widehat{b}$, and thus the estimated optimal ITR is the sign of $\widehat{f}(\boldsymbol{x})$. The covariates with nonzero scaling factors $\widehat{\eta}$ are identified to be important in estimating ITRs.

This variable selection technique can be applied to other nonlinear kernels, such as the polynomial kernel, and even to the linear kernel. However for the linear kernel, we recommend the approach in the prior section, where the subproblem (3.18) is a convex optimization problem with $p+1$ variables. Note that the subproblem (3.20) in this section is nonconvex even for the linear kernel, and there are $n+p+1$ variables for the optimization problem. This is much more challenging.

3.5 Simulation studies

We carried out extensive simulation studies to investigate finite sample performance of the proposed RWL methods.

We first evaluated the performance of RWL methods with low-dimensional covariates. In the simulations, we generated 5-dimensional vectors of clinical covariates $x_1, \cdots, x_5$, consisting of independent uniform random variables $U(-1, 1)$. The treatment A was generated from $\{-1, 1\}$ independently of X with $P(A = 1) = 0.5$. That is, $\pi(a, \boldsymbol{x}) = 0.5$ for all a and $\boldsymbol{x}$. The response R was normally distributed with mean $Q_0 = \mu_0(\boldsymbol{x}) + \delta_0(\boldsymbol{x}) \cdot a$ and standard deviation 1, where $\mu_0(\boldsymbol{x})$ is the common effect for clinical covariates, and $\delta_0(\boldsymbol{x}) \cdot a$ is the interaction between treatment and clinical covariates. We considered five scenarios with different choices of $\mu_0(\boldsymbol{x})$ and $\delta_0(\boldsymbol{x})$:

(0) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 0.4(x_2 - 0.25x_1^2 - 1)$.

(1) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 1.8(0.3 - x_1 - x_2)$.

(2) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 1.3(x_2 - 2x_1^2 + 0.3)$.

(3) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2 + 0.5x_4^2$; $\delta_0(\boldsymbol{x}) = 3.8(0.8 - x_1^2 - x_2^2)$.

(4) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2$; $\delta_0(\boldsymbol{x}) = 10(1 - x_1^2 - x_2^2)(x_1^2 + x_2^2 - 0.2)$.

Scenario 0 is similar with the second scenario in Zhao et al. (2012). When x_2 is restricted to $[-1, 1]$, the treatment arm -1 is always better than the treatment arm 1 on average. Thus the true optimal ITR in Scenario 0 is to assign all subjects to arm -1. The decision boundaries for the remaining four scenarios are determined by x_1 and x_2 only. The scenarios have different decision boundaries in truth. The decision boundary is a line in Scenario 1, a parabola in Scenario 2, a circle in Scenario 3, and a ring in Scenario 4. Their true optimal ITRs are illustrated in Figure 3.3. It is unclear how often a circle or ring boundary structure

will occur in practice, although general nonlinear boundaries are likely to arise frequently. The purpose of including these in the simulations is to verify that the proposed approach can handle even the most difficult boundary structures. The coefficients in Scenarios $0 \sim 3$ were chosen to reflect a medium effect size according to Cohen's d index (Cohen 1988); and the coefficients in Scenario 4 reflect a small effect size. The Cohen's d index is defined as the standardized difference in mean responses between two treatment arms, that is,

$$es = \frac{|\mathbb{E}(R|A=1) - \mathbb{E}(R|A=-1)|}{\sqrt{[Var(R|A=1) + Var(R|A=-1)]/2}}.$$

We compared the performance of the following five methods:

(1) The proposed RWL using the Gaussian RBF kernel (RWL-Gaussian). Residuals were estimated by a linear main effects model.

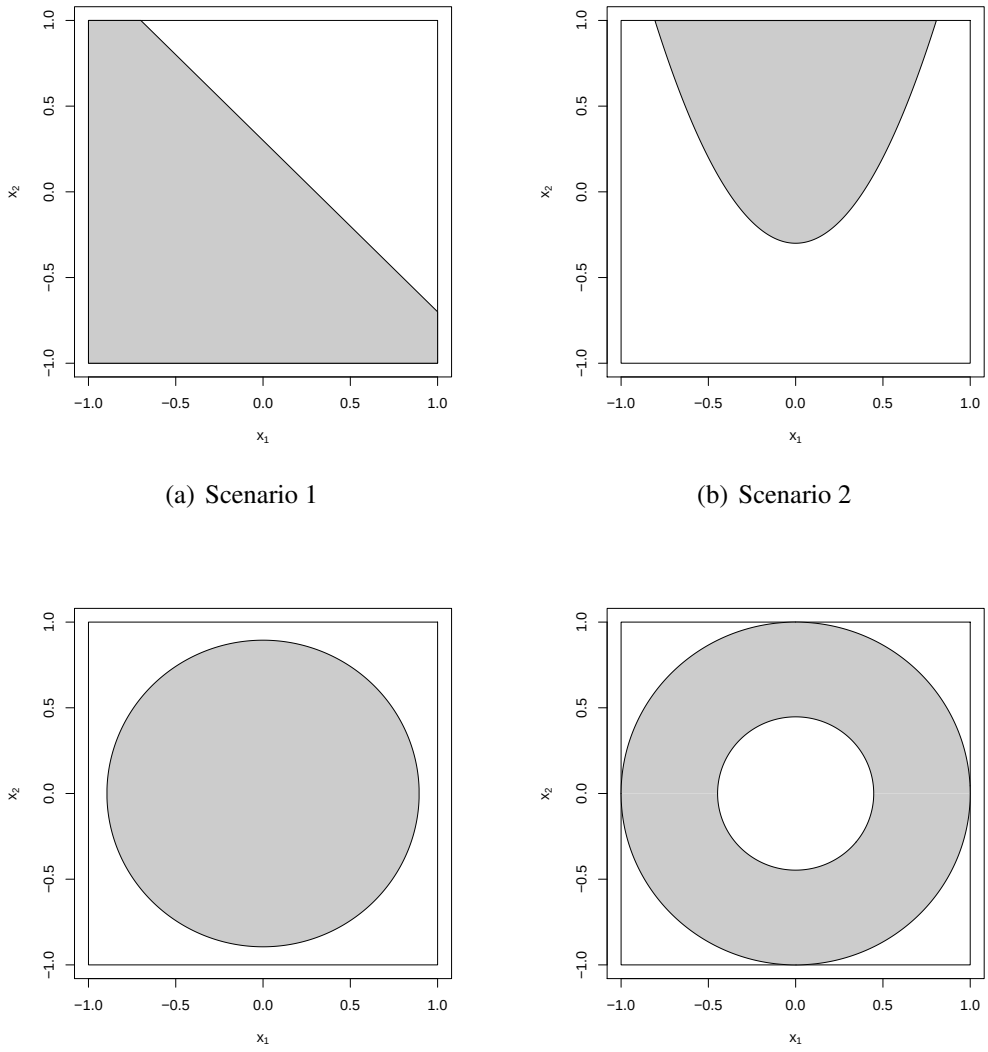
(2) The proposed RWL using the linear kernel (RWL-Linear). Residuals were estimated by a linear main effects model.

(3) OWL proposed in Zhao et al. (2012) using the Gaussian RBF kernel (OWL-Gaussian).

(4) OWL using the linear kernel (OWL-Linear).

(5) ℓ_1-PLS proposed by Qian and Murphy (2011).

The OWL methods were reviewed in the method section. Since the OWL methods can only handle nonnegative outcomes, we subtracted $\min\{r_i\}$ from all outcome responses. This is essentially the approach used in Zhao et al. (2012) [personal communication]. In the simulation studies, ℓ_1-PLS approximated $\mathbb{E}(R|\boldsymbol{X}, A)$ using the basis function set $(1, \boldsymbol{X}, A, \boldsymbol{X}A)$, and applied LASSO (Tibshirani 1994) for variable selection. The optimal ITR was determined by the sign of the difference between the estimated $\mathbb{E}(R|\boldsymbol{X}, A=1)$ and $\mathbb{E}(R|\boldsymbol{X}, A=-1)$.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 3.3: *True optimal ITRs in simulation studies. For subjects in the shade area, the best treatment is 1; for subjects in the white area, the best treatment is* -1.

For each scenario, we considered two sample sizes for training datasets: $n = 100$ and $n = 400$, and repeated the simulation 500 times. All five methods had at least one tuning parameter. For example, there was one tuning parameter, λ, in RWL-linear; and two tuning parameters, λ and σ in RWL-Gaussian. In the simulations, we applied a 10-fold cross-validation procedure to tune parameters. For RWL-Linear, we searched over a pre-specified finite set of λ's to select the best one maximizing the average of estimated

values from validation data; while for RWL-Gaussian, we searched over a pre-specified finite set of (λ, σ)'s to select the best pair. The selection of tuning parameters in ℓ_1-PLS and OWL followed similarly. In the comparisons, the performances of five methods were evaluated by two criteria: the first criterion is the value function under the estimated optimal ITR when we applied to an independent and large test dataset; the second criterion is the misclassification error rate of the estimated optimal ITR from the true optimal ITR on the large test dataset. Specifically, a test set with 10,000 observations was simulated to assess the performance. The estimated value function under any ITR d is given by $\mathbb{P}_n^*[R\mathbb{I}(A = d(\boldsymbol{X}))/\pi(A, \boldsymbol{X})]/\mathbb{P}_n^*[\mathbb{I}(A = d(\boldsymbol{X}))/\pi(A, \boldsymbol{X})]$ (Murphy 2005), where $\mathbb{P}_n^*$ denotes the empirical average using the test data; the misclassification rate under any ITR d is given by $\mathbb{P}_n^*[\mathbb{I}(d(\boldsymbol{X}) \neq d^*(\boldsymbol{X}))]$, where d^* is the true optimal ITR which is known when generating the simulated data.

The simulation results are presented in Table 3.1. We report the mean and standard deviation of value functions and misclassification rates over 500 replications. In Scenario 0, the true optimal ITR would assign all subjects to treatment arm -1. When the sample size was small $(n = 100)$, ℓ_1-PLS, OWL-Linear and RWL-Linear showed similar performance, while OWL-Gaussian and RWL-Gaussian were slightly worse. Note that when x_2 can go beyond 1, the heterogeneous treatment effect does exist with non-linear (parabola) decision boundary. The deteriorated performance of OWL-Gaussian and RWL-Gaussian may be due to unexpected extrapolation. When the sample size was large $(n = 400)$, all methods performed similarly, and assigned almost all subjects to arm -1. Scenario 1 was a linear example. The model specification in ℓ_1-PLS was correct, and this method performed very well. The proposed RWL methods, with either the linear kernel or the Gaussian RBF kernel, yielded similar performance with ℓ_1-PLS, and were much better than OWL methods. Another advantage of using RWL versus OWL was also reflected by a smaller variance in value functions. The idea of using residuals instead of the original outcomes is able to stabilize the variance, as we discussed in the method section.

The conditional mean outcomes and decision boundaries in the remaining three scenarios were nonlinear. ℓ_1-PLS, RWL-Linear, and OWL-Linear were misspecified, and hence they did not perform very well in these three scenarios. We focus on the comparison between RWL-Gaussian and OWL-Gaussian. These two methods equipped with the Gaussian RBF kernel should have the ability to capture the nonlinear structure of decision boundaries. In Scenarios 2 and 3, RWL-Gaussian outperformed OWL-Gaussian in terms of achieving higher values and smaller variances. It was challenging to find the optimal ITR in Scenario 4 since the decision boundary was complicated and the effective size was small. When the sample size was small ($n = 100$), the performance of RWL-Gaussian was only slightly better than that of OWL-Gaussian, but with larger variance. It might be too hard to learn the complicated decision boundary from the limited sample. However, when the sample size increased, RWL-Gaussian showed significantly better performance than other methods. The excellent performance of RWL-Gaussian confirms its power in finding the optimal ITR. Although using the Gaussian RBF kernel, the performance of OWL-Gaussian was not comparable with that of RWL-Gaussian. In Scenario 2, OWL-Gaussian performed even worse than ℓ_1-PLS and RWL-Linear, both of which can only detect linear boundaries. While in Scenario 4, OWL-Gaussian gave similar performance as OWL-Linear. Zhao et al. (2012) also reported similar finding in their simulation studies that there are no large differences in the performance between OWL-Linear and OWL-Gaussian.

As we demonstrated in Section 3.2.2, OWL methods tend to favor the treatment assignments that subjects actually received, and this behavior deteriorates finite sample performance. We calculated treatment matching factors of these five methods on the training data. They are shown in Table 3.2. We expect the treatment matching factor to be close to 1. However, the treatment matching factors for OWL methods were greater than 1. OWL-Gaussian achieved the largest treatment matching factors, especially when the sample size was small. This may partially explain why OWL-Gaussian may not outperform OWL-Linear even when the decision boundary is nonlinear. Whereas for RWL methods,

the treatment matching factors were close to 1. By appropriately controlling the treatment matching factors, the aim of RWL methods is to maximize the estimate in (3.8), and hence to improve finite sample performance.

We applied a linear main effects model to estimate residuals. The model was correctly specified for Scenarios 0, 1 and 2, and misspecified for Scenarios 3 and 4. However, the superior performance of RWL methods in Scenarios 3 and 4 demonstrates the robustness of our methods to residual estimates. We also considered a null model, which was misspecified for all scenarios, to estimate residuals. The results were only slightly worse than those shown in Table 3.1 for Scenarios 1 and 2 when $n = 100$, and other scenarios were similar especially when the sample size increased [results not shown].

We then evaluated the performance of RWL methods with moderate-dimensional clinical covariates. We adopted the same data generating procedure as in the above five scenarios except that the dimension of clinical covariates was increased from 5 to 50 by adding 45 independent uniform random variables $U(-1, 1)$. Thus among all these 50 clinical covariates, only x_1 and x_2 are attributed to the decision boundaries for Scenarios $1 \sim 4$. We repeated the simulations, and the results are shown in Table 3.3. In Scenario 0, all methods performed similar, and assigned almost all subjects to the treatment arm -1 when the sample size was large $(n = 400)$. In Scenario 1, ℓ_1-PLS performed very well because of correct model specification and inside variable selection techniques. Although RWL methods showed better performance than OWL methods, they were not comparable to ℓ_1-PLS. The decision boundaries in Scenarios $2 \sim 4$ are nonlinear. Both RWL and OWL failed to detect the decision boundaries in these scenarios. They produced similar performance with either the linear kernel or Gaussian RBF kernel. RWL and OWL methods with Gaussian RBF kernel may perform particularly poorly when many non-descriptive covariates are present.

We tested the performance of variable selection approaches described in Section 3.4 for the linear kernel and Gaussian RBF kernel. They are called RWL-VS-Linear and RWL-VS-Gaussian, respectively, in this work. There were two tuning parameters, λ_1 and λ_2, for both approaches. We applied a 10-fold cross-validation procedure to tune parameters from a pre-specified finite set of (λ_1, λ_2)'s. The simulation results are also shown in Table 3.3. In Scenario 0, both RWL-VS methods showed similar performance as RWL methods. In Scenario 1, both RWL-VS-Linear and RWL-VS-Gaussian improved the performance of RWL methods. When the sample size was 400, the performance was as good as that of ℓ_1-PLS. For Scenarios $2 \sim 4$, we only present results from RWL-VS-Gaussian. When the sample size was small ($n = 100$), the performance was slightly better (in Scenarios 2 and 3) than, or similar (in scenario 4) as that of RWL-Gaussian. It is well known that large sample sizes are needed to find interactions in models. So for similar reasons, we may expect large samples to be necessary to find and confirm the existence of heterogeneity of treatment effects, and accurately define the decision boundary. In the simulations, when the sample size increased, the performance of RWL-VS-Gaussian improved significantly. Variable selection is necessary for RWL when there are many non-descriptive clinical covariates.

3.6 Data analysis

We applied the proposed methods to analyze data from the EPIC randomized clinical trial (Treggiari et al. 2011). The trial was designed to determine the optimal anti-pseudomonal treatment strategy in children with cystic fibrosis (CF) who recently acquired Pseudomonas aeruginosa (*Pa*). Newly identified *Pa* was defined as the first lifetime documented *Pa* positive culture within 6 months of baseline or a *Pa* positive culture within 6 months of baseline after a two-year absence of *Pa* culture-positivity. Other eligibility criteria have been previously reported (Treggiari et al. 2009). A total of 304 patients ages 1$\sim$12

were randomized in a 1:1 ratio to one of two maintenance treatment strategies: cycled therapy (treatment with anti-pseudomonal therapy provided in quarterly cycles regardless of *Pa* positivity) and culture-based therapy (treatment only in response to identification of *Pa* positive cultures from quarterly cultures). All patients regardless of randomization strategy received an initial course of anti-pseudomonal therapy in response to their *Pa* positive culture at eligibility and proceeded with treatment according to their randomization strategy. In this work, we considered two endpoints over the course of the 18-month study. One is the number of *Pa* positive cultures from scheduled follow-up quarterly cultures, and the other is the number of pulmonary exacerbations (PE) requiring any (intravenous, inhaled, or oral) antibiotic use or hospitalization during the study. In the original clinical trial, there was no significant difference between the two maintenance treatment strategies at the population level for either the *Pa* outcome (p-value 0.222) or PE outcome (p-value 0.280).

We considered 10 baseline clinical covariates, including age, gender, F508del genotype, weight, height, BMI, first documented lifetime *Pa* positive culture at eligibility versus positive after a two-year history of negative cultures (neverpapos), use of antibiotics within 6 months prior to baseline (anyabxhist), *Pa* positive at the baseline visit versus *Pa* positive within 6 months prior to baseline only (pabl), and positive versus negative baseline S. aureus culture status (sabl). There were three levels (homozygous, heterozygous and other) for F508del genotype. We coded genotype as two dichotomous covariates, homozygous and heterozygous, each referenced in relation to the "othe" genotype category. Four patients did not give consent to have their data put into the databank. We also excluded 17 patients with missing covariate values from the analyses. The data used here consisted of 283 patients: 141 in the cycled therapy, and 142 in the culture-based therapy. Each clinical covariate was linearly scaled to the range $[-1, +1]$, as recommended in Hsu et al. (2003).

For the *Pa*-related endpoint, we used the ratio of the number of *Pa* negative cultures

to the number of *Pa* cultures over the follow up period as the outcome. We were seeking an ITR to reduce the number of *Pa* positive cultures, so larger outcomes were preferred. We compared seven methods: ℓ_1-PLS, OWL-Linear, OWL-Gaussian, RWL-Linear, RWL-Gaussian, RWL-VS-Linear, and RWL-VS-Gaussian. ℓ_1-PLS used the basis function set $(1, \boldsymbol{X}, A, \boldsymbol{X}A)$ in the regression model. We treated the outcome as continuous, and used a linear main effects model to estimate residuals for RWL methods. We used a 10-fold cross-validation procedure to tune parameters. The estimated rule was then used to predict the optimal treatment for each patient. The predicted treatment allocation is shown in Table 3.4. To evaluate the performance of estimated rules, we again carried out a 10-fold cross validation procedure. The data were randomly partitioned into 10 roughly equal-sized parts. We estimated the ITR on nine parts of the data using the tuned parameters, and predicted optimal treatments on the part left out. We repeated the procedure nine more times to predict the other parts, and obtained the predicted treatment for each patient. The first evaluation criterion was the value function, which is given by $\mathbb{P}_n[R\mathbb{I}(A = Pred)/\pi(A, \boldsymbol{X})]/\mathbb{P}_n[\mathbb{I}(A = Pred)/\pi(A, \boldsymbol{X})]$, where $\mathbb{P}_n$ denotes the empirical average over a fold in cross validation and $Pred$ is the predicted treatment in the cross validation procedure. The second evaluation criterion was related to significance tests. The test was performed when a 10-fold cross validation procedure is finished, that is, when we had the predicted treatments for all patients. By comparing the predicted treatments with the treatments that were actually assigned to patients, we divided patients into two groups: those who followed the estimated optimal rule and those who did not follow the rule. We tested the difference between the two groups using the two-sample t-test to see whether the group that followed the estimated rule was better than the other group at the significance level $\alpha = 0.05$. The whole procedure was repeated 100 times with different fold partitions in the cross validation. We obtained 1000 value functions and 100 p-values. The mean and standard deviation of these value functions, the proportion of significant p-values and median of these p-values are also presented in Table 3.4. As reference rules,

we also considered two fixed treatment rules: assign all patients to (1) cycled therapy arm or (2) culture-based therapy arm. We show in Table 3.4 their value functions from repeated 10-fold cross validation procedures and one-sided p-values from two-sample t-tests. Note that the p-values for fixed rules were not changed during the cross validation procedure. ℓ_1-PLS showed the best performance, and identified that one covariate, baseline *Pa* status, is important in estimating ITRs. The significance tests also confirmed its superior performance. Both OWL methods failed to detect the heterogeneity of treatment effects. They assigned all patients to the cycled therapy arm. Although the performances of RWL methods (without variable selection) were only slightly better, on average, than those of OWL methods, they did detect some heterogeneity across treatments. Variable selection further improved the performance in terms of achieving higher values and better significance tests. Moreover, both RWL-VS-Linear and RWL-VS-Gaussian selected the same covariate as ℓ_1-PLS.

Our results are consistent with results from the trial suggesting that the subgroup of patients who were *Pa* negative at the baseline visit (albeit positive within 6 months of baseline to meet eligibility) may have more greatly benefited from cycled therapy to suppress *Pa* positivity during the trial. However, it is important to note clinically that further analyses of the trial data demonstrated the comparable effectiveness using culture-based therapy as compared to cycled therapy in reducing the overall prevalence of *Pa* positive cultures at the end of the trial.

The number of PEs during the study was a rate variable. We were seeking an ITR to lower the number of PEs, so fewer PEs were desirable. We compared several methods: PoissonReg, ℓ_1-PoissonReg, OWL-Linear, OWL-Gaussian, RWL-Linear, RWL-Gaussian, RWL-VS-Linear, and RWL-VS-Gaussian. PoissonReg fitted a Poisson regression model using the basis function set $(1, \boldsymbol{X}, A, \boldsymbol{X}A)$, computed the predicted number of PEs for a particular patient at each arm, and assigned the treatment with smaller prediction to this

patient. ℓ_1-PoissonReg was similar with PoissonReg, but used an ℓ_1 penalized term to perform variable selection. For OWL methods, we used the opposite of individual annual PE rate as the continuous outcome. The annual PE rate for the i-th patient is defined as $\frac{r_i}{t_i} * 365.25$, where r_i is the number of PEs for the i-th patient, and t_i is the duration (in days) when the i-th patient stayed in the study. For RWL methods, we used a main effects Poisson regression model to estimate residuals, as described in Section 3.2.4. A 10-fold cross-validation was used to tune parameters. The predicted treatment allocation is shown in Table 3.5. A similar evaluation procedure as that for the *Pa* endpoint was also performed, where we used instead the annual PE rate for the group that followed the rule as the first evaluation criterion. The group-wise annual PE rate is given as $\frac{\mathbb{P}_n[R\mathbb{I}(A=Pred)]}{\mathbb{P}_n[T\mathbb{I}(A=Pred)]} * 365.25$, where R is the number of PEs, T is the duration (in days) in the study, $\mathbb{P}_n$ denotes the empirical average over a fold in cross validation and $Pred$ is the predicted treatment assignment in the cross validation procedure. For the second criterion, a Poisson regression model was used to test whether the group that followed the estimated rule had fewer PEs than the other group that did not follow the estimated rule. The results are presented in Table 3.5. We also considered two fixed rules as references. All RWL methods outperformed OWL and Poisson regression methods in terms of achieving better PE rates and better significance test results. ℓ_1-PoissonReg identified three important covariates, baseline *Pa* status, baseline S. aureus staus and height, for treatment assignment.

For this case, RWL-VS-Linear and RWL-VS-Gaussian did not improve performances comparing with their counterparts without variable selection. RWL-VS-Linear did not discard any clinical covariates. As its tuned parameter λ_1 was very small, RWL-VS-Linear gave a similar performance as RWL-Linear. The estimated rule from RWL-Linear was

determined by the following decision function F,

$$\begin{aligned} F &= 1.43 * \mathbb{I}(gender = female) + 0.07 * \text{age} - 0.58 * \mathbb{I}(neverpapos = yes) \\ &\quad -1.01 * \mathbb{I}(pabl =' +') - 0.17 * \mathbb{I}(sabl =' +') + 0.02 * \text{weight} + 0.01 * \text{height} \\ &\quad -0.17 * \mathbb{I}(anyabxhist = yes) + 0.11 * \text{bmi} - 0.90 * \mathbb{I}(genotype = homozygous) \\ &\quad -0.83 * \mathbb{I}(genotype = heterozygous) - 2.90. \end{aligned}$$

The corresponding ITR is that if $F > 0$, assign cycled therapy to this patient, otherwise assign the culture-based therapy. The proposed rule suggests that female gender and increasing age are key determinants for the potential of a patient to benefit from cycled therapy, which is not surprising since these characteristics are known risk factors for exacerbations. Counterintuitively, this rule suggests that Pa positivity at baseline leads against the recommendation for cycled therapy unlike with the prior outcome of increased Pa frequency. The rule also suggests that patients with genotypes other than delF508 heterozygous and homozygous may benefit more from cycled therapy as compared to patient with the delF508 mutation. Overall, these results indicate the importance of evaluating multiple endpoints to evaluate consistency in recommendations for treatment based on these proposed rules.

For RWL-VS-Gaussian, three clinical covariates (age, weight, baseline S. aureus status) were identified to be unimportant for estimating ITRs. It is well known that the shrinkage by the ℓ_1 penalty causes the estimates of the non-zero coefficients to be biased towards zero (Hastie et al. 2001). For similar reasons, the estimates from RWL-VS-Gaussian might be biased. One approach for reducing this bias is to run the ℓ_1 penalty method to identify important covariates, and then fit a model without the ℓ_1 term to the selected set of covariates. We applied RWL-Gaussian to the identified set of covariates. The results are also presented in Table 3.5. The obtained annual PE rate was better than that from RWL-Gaussian alone. Thus variable selection is still helpful to improve the performance.

In summary, the proposed RWL methods can identify potentially useful ITRs. For ex-

ample, the estimated ITR for the *Pa* endpoint in the EPIC trial is that if the baseline *Pa* status is positive, assign culture-based therapy to the patient; if the baseline *Pa* status is negative, assign the cycled therapy. As the *Pa* endpoint was a secondary endpoint in the trial, the clinical utility of these findings must be weighed in context of all study findings and safety. But nonetheless identifying such a strategy is important to improve personalized clinical practice if the strategy is confirmed by future comparative studies.

3.7 Discussion

In this work, we have proposed a general framework called Residual Weighted Learning (RWL) to use outcome residuals to estimate optimal ITRs. The residuals may be obtained from linear models or generalized linear models, and hence RWL can handle various types of outcomes, including continuous, count and binary outcomes.

The proposed RWL methods appear to achieve better performance compared to other existing methods as shown in the simulation studies and real data analyses. The goal of RWL is to improve finite sample performance of OWL methods. It is not surprising that RWL outperforms OWL in terms of achieving higher value function and smaller variance. Two-step methods, such as ℓ_1-PLS and ℓ_1-PoissonReg, are generally parametric. Their performance largely depends on how close the posited model is to the true model of the conditional mean outcome. When the model is correctly specified, two-step methods can be more efficient than nonparametric RWL methods. However, the relationship among clinical covariates, treatment assignment, and outcome is complex in practice. Misspecified models may lead to biased estimates.

Both two-step methods and RWL require correct model specification to guarantee their performance. We consider two levels of model specification. The first one is the model for

the decision boundary; the other is the model for the outcome. Two-step methods, such as ℓ_1-PLS, only specify models for conditional outcomes, and the decision boundary is implied from these models. Thus the performance of two-step methods are only related to these models. However, for RWL, the two levels are separated. RWL methods target the decision boundary directly. The model for decision boundary is more critical to performance. RWL with Gaussian RBF kernel can approximate any form of decision boundary when the sample size is sufficiently large. Even when using linear kernels, RWL is more robust to model misspecification on the decision boundary than the two-step methods. If the conditional mean outcome is assumed to be linear in the parameters, the decision boundary will be also; the converse does not hold. For RWL with linear kernel, the decision boundary remains a simple linear form, while the conditional mean outcome is allowed to be nonlinear.

Another source of robustness of RWL comes from the model for the outcome, in which we estimate residuals. By Theorem 3.3.3, no matter how bad the residual estimate is, RWL is still consistent. In the simulation studies, we have investigated the robustness of residual estimates on finite samples. We applied the null model, as an example of incorrect model specification, to estimate residuals, and achieved satisfactory performance. The simulation studies confirm the robustness of RWL. In this work, we used simple models to obtain the residuals. However, RWL is quite flexible in the methods used to estimate residuals. When the sample size is small or the heterogeneous treatment effect is weak, we may consider a complicated method to accurately estimate residuals to improve the performance of RWL. Alternatively, nonparametric regression methods, such as support vector regression (Vapnik 1998) or random forests (Breiman 2001), are an option to eliminate model misspecification.

RWL is closely related to a robust augmented inverse probability weighted estimator (AIPWE) of the value function. AIPWE was first introduced into optimal treatment

regimes by Zhang et al. (2012b). Under the setup in this article, *i.e.*, $\mathcal{A} = \{1, -1\}$ and $\pi(a, \boldsymbol{x})$ is known, for any rule d, its associated value function can be estimated by an AIPWE,

$$\text{AIPWE}(d) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}(d(\boldsymbol{x}_i) = a_i)}{\pi(a_i, \boldsymbol{x}_i)}R_i - \frac{\mathbb{I}(d(\boldsymbol{x}_i) = a_i) - \pi(a_i, \boldsymbol{x}_i)}{\pi(a_i, \boldsymbol{x}_i)}\widehat{m}(\boldsymbol{x}_i)\right\},$$

where $\widehat{m}(\boldsymbol{x}) = \widehat{\mu}_1(\boldsymbol{x})\mathbb{I}(d(\boldsymbol{x}) = 1) + \widehat{\mu}_{-1}(\boldsymbol{x})\mathbb{I}(d(\boldsymbol{x}) = -1)$, $\widehat{\mu}_1(\boldsymbol{x})$ is an estimate of $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1)$, and $\widehat{\mu}_{-1}(\boldsymbol{x})$ is an estimate of $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$. If we take $\widehat{g}^*(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x})$, for a fixed $d(\boldsymbol{x})$, then maximizing AIPWE is equivalent to minimizing the risk in (3.6). Hence AIPWE is essentially a "contrast" weighted learning method, where the contrast is a modelled form that links closely to the residual. Conversely, if we take $\widehat{\mu}_1(\boldsymbol{x}) = \widehat{\mu}_{-1}(\boldsymbol{x}) = \widehat{g}^*(\boldsymbol{x})$, where $\widehat{g}^*(\boldsymbol{x})$ estimates $g^*(\boldsymbol{x})$ in (3.5), then maximizing AIPWE is again equivalent to minimizing (3.6). However, RWL can now be obtained by replacing the 0-1 loss with the smoothed ramp loss, and hence RWL can be viewed as a modification of AIPWE with alternative useful connections to classification methods.

Variable selection is critical for good performance of RWL, as demonstrated in the simulation studies and data analyses. We suggest that practical use of RWL should necessarily be accompanied by variable selection. In Section 3.4, we introduce the concepts of predictive and prescriptive covariates. It has been noticed in practical settings that the main effects, rather than treatment-covariate interactions, tend to explain most of the variability in the outcome, and thus they are important for good prediction (Gunter et al. 2011). It is not uncommon that interactions between prescriptive covariates and treatment are overlooked due to their small predictive ability. Hence variable selection techniques designed for prediction applications might not work perfectly in finding the prescriptive covariates. In contrast, RWL focuses on treatment-covariate interactions, and variable selection in RWL only targets the prescriptive covariates.

RWL methods with linear and Gaussian RBF kernels were discussed in this work. The decision rule from linear RWL is easy to interpret, but at the risk of misspecification. As

a universal kernel, the Gausian RBF kernel is free of model misspecification. An obvious downside is that the final decision rule may be difficult to interpret, and clinical practitioners may not be comfortable using "black box" decision rules. One compromise is to apply linear RWL with a rich set of bases, for example, including two-way and three-way covariate interactions.

When the endpoint outcome is the survival time, observations are commonly subject to right censoring because of subject dropout or censoring due to the end of study. Currently, RWL does not cover this type of outcomes in theory due to the censorship. We recommend using martingale residuals (Fleming and Harrington 1991) as alternative outcomes. Specifically, residuals in RWL could be opposites of martingale residuals that are estimated by a Cox proportional hazards model on clinical covariates but excluding treatment assignment. Actually, the idea is not new. Therneau et al. (1990) suggested that martingale residuals from a null Cox model could be used as outcomes for Classification and Regression Trees (CART). The rigourous theoretical justification is under development. Here we provide an intuitive explanation. Let T denote the survival time, and C denote the censoring time. The observed data quadruplet is $(\boldsymbol{X}, A, R = T \wedge C, \Delta = \mathbb{I}(T \leq C))$. Firstly, for general survival analysis, (R, Δ) is seen as the censored outcome. We view the survival data from a different perspective. We treat Δ as the outcome, and $(\boldsymbol{X}, R)$ as clinical covariates. In the framework of counting processes, the martingale residual can be explained as the difference between the observed number of failures (0 or 1) during the time in the study for a subject and the expected numbers based on the fitted model. Thus it is almost in line with our derivation of RWL, although R may not be appropriate as a clinical covariate. Secondly, martingale residuals from a Cox model fitted excluding a covariate of interest can be used to determine the appropriate functional form of the covariate (Therneau et al. 1990). If we estimate martingale residuals by ignoring treatment assignment, they contain information related to the heterogeneity of treatment effects. Thirdly, martingale residuals have some properties reminiscent of linear models, for example, the residuals are asymp-

totically uncorrelated. Moreover, the weighted sum of martingale residuals is zero, which is essential for RWL to control the treatment matching factor.

Several extensions are worthy of further investigation. In this work, we only considered two-arm trials. In practice, some trials involve multiple arms, *i.e.* have three or more treatment arms. A recent review by Baron et al. (2013) reported that 17.6% of published randomized clinical trials in 2009 were multiple-arm. Because of the number of possible comparisons, it is more complex to find ITRs in such trials compared with two-arm trials. It would thus be worthwhile to extend RWL to multiple-arm trials. In the machine learning literature, two schools of approaches are generally suggested for multi-category classification. One is to directly take all classes into consideration. Multicategory SVMs (Lee et al. 2004) and penalized logistic regression (Zhu and Hastie 2004) are two examples. The other school is to construct and combine several binary classifiers. One-versus-one, one-versus-all, and more general error-correcting output codes (Dietterich and Bakiri 1995, ECOC) are in this school. Similar generalizations may be possible for finding ITRs for multiple-arm trials.

For chronic diseases, multi-stage dynamic treatment regimes are more useful than single-stage optimal treatment regimes. In the context of multi-stage decision problems, a dynamic treatment regime is a sequence of decision rules, one per stage of intervention, for adapting a treatment plan over time to an individual. Zhao et al. (2015) has extended OWL to dynamic treatment regimes. Considering the improved performance of RWL over OWL, it is of interest to extend RWL to dynamic treatment regimes.

The intended use of the estimated optimal ITRs discovered using the proposed approach is for treating new patients. The theoretical, simulation and data analysis results demonstrate that such estimated ITRs are likely to lead to improved clinical outcomes for these patients. However, since the data used for estimating the ITRs should not also be used for confirmation, it is expected that an additional randomized clinical trial compar-

ing the estimated ITR with standard of care, or some other suitable treatment comparison, will be used for confirmation as is typically done for new candidate treatments (Zhao et al. 2011).

Table 3.1: *Mean (std) of empirical value functions and misclassification rates evaluated on independent test data for 5 simulation scenarios with 5 covariates. The best value function and minimal misclassification rate for each scenario and sample size combination are in bold.*

	$n = 100$		$n = 400$	
	Value	Misclassification	Value	Misclassification
Scenario 0 (Optimal value 1.43)				
ℓ_1-PLS	**1.41 (0.07)**	**0.05 (0.10)**	1.42 (0.01)	0.02 (0.04)
OWL-Linear	1.39 (0.11)	0.07 (0.14)	**1.43 (0.02)**	**0.01 (0.04)**
OWL-Gaussian	1.33 (0.15)	0.13 (0.18)	1.41 (0.05)	0.03 (0.08)
RWL-Linear	1.40 (0.05)	0.06 (0.09)	1.42 (0.02)	0.03 (0.05)
RWL-Gaussian	1.34 (0.10)	0.13 (0.13)	1.40 (0.04)	0.07 (0.07)
Scenario 1 (Optimal value 2.25)				
ℓ_1-PLS	**2.22 (0.03)**	**0.06 (0.04)**	**2.24 (0.02)**	**0.03 (0.03)**
OWL-Linear	1.91 (0.22)	0.23 (0.09)	2.08 (0.14)	0.16 (0.06)
OWL-Gaussian	1.88 (0.24)	0.24 (0.09)	2.08 (0.12)	0.16 (0.06)
RWL-Linear	2.19 (0.04)	0.09 (0.04)	2.23 (0.01)	0.05 (0.02)
RWL-Gaussian	2.17 (0.06)	0.11 (0.04)	2.22 (0.02)	0.05 (0.02)
Scenario 2 (Optimal value 1.96)				
ℓ_1-PLS	1.71 (0.07)	0.24 (0.04)	1.75 (0.01)	0.22 (0.01)
OWL-Linear	1.51 (0.12)	0.32 (0.05)	1.59 (0.10)	0.29 (0.06)
OWL-Gaussian	1.49 (0.15)	0.33 (0.06)	1.63 (0.11)	0.26 (0.05)
RWL-Linear	1.66 (0.08)	0.26 (0.04)	1.74 (0.03)	0.23 (0.02)
RWL-Gaussian	**1.75 (0.09)**	**0.20 (0.05)**	**1.90 (0.03)**	**0.10 (0.03)**
Scenario 3 (Optimal value 3.88)				
ℓ_1-PLS	3.00 (0.11)	0.38 (0.03)	3.03 (0.04)	0.37 (0.01)
OWL-Linear	2.98 (0.10)	0.38 (0.03)	3.01 (0.02)	0.38 (0.01)
OWL-Gaussian	3.21 (0.18)	0.32 (0.05)	3.56 (0.12)	0.22 (0.04)
RWL-Linear	3.15 (0.13)	0.34 (0.03)	3.28 (0.04)	0.31 (0.01)
RWL-Gaussian	**3.62 (0.12)**	**0.19 (0.04)**	**3.82 (0.04)**	**0.10 (0.02)**
Scenario 4 (Optimal value 3.87)				
ℓ_1-PLS	2.29 (0.14)	0.49 (0.08)	2.38 (0.15)	0.55 (0.08)
OWL-Linear	2.42 (0.14)	0.55 (0.08)	2.49 (0.11)	0.59 (0.06)
OWL-Gaussian	2.43 (0.15)	0.52 (0.07)	2.59 (0.15)	0.50 (0.07)
RWL-Linear	2.42 (0.14)	0.54 (0.09)	2.49 (0.10)	0.58 (0.08)
RWL-Gaussian	**2.68 (0.28)**	**0.43 (0.08)**	**3.49 (0.08)**	**0.22 (0.03)**

Table 3.2: *Mean (std) of treatment matching factors evaluated on the training data for 5 simulation scenarios with 5 covariates.*

	$n = 100$	$n = 400$
Scenario 0		
ℓ_1-PLS	1.00 (0.04)	1.00 (0.01)
OWL-Linear	1.03 (0.07)	1.00 (0.01)
OWL-Gaussian	1.14 (0.24)	1.02 (0.06)
RWL-Linear	1.00 (0.04)	1.00 (0.01)
RWL-Gaussian	1.00 (0.06)	1.00 (0.03)
Scenario 1		
ℓ_1-PLS	0.99 (0.10)	0.99 (0.05)
OWL-Linear	1.10 (0.08)	1.04 (0.04)
OWL-Gaussian	1.15 (0.13)	1.05 (0.05)
RWL-Linear	0.99 (0.09)	0.99 (0.04)
RWL-Gaussian	0.99 (0.08)	0.99 (0.04)
Scenario 2		
ℓ_1-PLS	1.00 (0.09)	1.00 (0.04)
OWL-Linear	1.06 (0.09)	1.03 (0.04)
OWL-Gaussian	1.18 (0.20)	1.10 (0.07)
RWL-Linear	1.00 (0.08)	1.00 (0.04)
RWL-Gaussian	1.01 (0.07)	1.00 (0.04)
Scenario 3		
ℓ_1-PLS	1.00 (0.05)	1.00 (0.01)
OWL-Linear	1.02 (0.06)	1.00 (0.02)
OWL-Gaussian	1.32 (0.17)	1.14 (0.05)
RWL-Linear	1.01 (0.06)	1.01 (0.04)
RWL-Gaussian	1.04 (0.06)	1.02 (0.04)
Scenario 4		
ℓ_1-PLS	1.00 (0.08)	1.00 (0.03)
OWL-Linear	1.07 (0.10)	1.02 (0.04)
OWL-Gaussian	1.37 (0.34)	1.27 (0.24)
RWL-Linear	1.00 (0.06)	1.00 (0.03)
RWL-Gaussian	1.00 (0.08)	1.02 (0.04)

Table 3.3: *Mean (std) of empirical value functions and misclassification rates evaluated on independent test data for 5 simulation scenarios with 50 covariates. The best value function and minimal misclassification rate for each scenario and sample size combination are in bold.*

	$n = 100$		$n = 400$	
	Value	Misclassification	Value	Misclassification
Scenario 0 (Optimal value 1.43)				
ℓ_1-PLS	1.34 (0.15)	0.11 (0.17)	1.42 (0.03)	0.03 (0.06)
OWL-Linear	1.36 (0.15)	0.08 (0.17)	**1.43 (0.03)**	**0.01 (0.03)**
OWL-Gaussian	1.35 (0.14)	0.10 (0.16)	1.42 (0.04)	**0.01 (0.05)**
RWL-Linear	**1.38 (0.11)**	**0.07 (0.13)**	1.42 (0.04)	0.03 (0.06)
RWL-Gaussian	1.34 (0.13)	0.11 (0.15)	1.40 (0.05)	0.04 (0.07)
RWL-VS-Linear	1.34 (0.11)	0.11 (0.14)	1.40 (0.04)	0.05 (0.06)
RWL-VS-Gaussian	1.33 (0.12)	0.13 (0.14)	1.41 (0.03)	0.05 (0.06)
Scenario 1 (Optimal value 2.25)				
ℓ_1-PLS	**2.17 (0.11)**	**0.09 (0.06)**	**2.23 (0.02)**	**0.04 (0.03)**
OWL-Linear	1.51 (0.12)	0.37 (0.03)	1.68 (0.10)	0.32 (0.03)
OWL-Gaussian	1.49 (0.13)	0.38 (0.04)	1.65 (0.11)	0.32 (0.04)
RWL-Linear	1.75 (0.14)	0.29 (0.05)	2.14 (0.03)	0.13 (0.02)
RWL-Gaussian	1.77 (0.12)	0.29 (0.04)	2.14 (0.03)	0.13 (0.02)
RWL-VS-Linear	2.05 (0.14)	0.17 (0.07)	2.22 (0.02)	0.06 (0.03)
RWL-VS-Gaussian	2.02 (0.18)	0.18 (0.08)	2.21 (0.08)	0.06 (0.04)
Scenario 2 (Optimal value 1.96)				
ℓ_1-PLS	**1.56 (0.19)**	**0.30 (0.07)**	1.73 (0.03)	0.23 (0.01)
OWL-Linear	1.42 (0.14)	0.37 (0.04)	1.47 (0.06)	0.35 (0.01)
OWL-Gaussian	1.40 (0.14)	0.38 (0.04)	1.47 (0.07)	0.35 (0.02)
RWL-Linear	1.42 (0.12)	0.36 (0.04)	1.60 (0.05)	0.28 (0.02)
RWL-Gaussian	1.40 (0.14)	0.36 (0.04)	1.62 (0.05)	0.28 (0.02)
RWL-VS-Gaussian	**1.56 (0.15)**	**0.30 (0.07)**	**1.92 (0.06)**	**0.09 (0.05)**
Scenario 3 (Optimal value 3.88)				
ℓ_1-PLS	2.90 (0.19)	0.40 (0.05)	3.00 (0.04)	0.38 (0.01)
OWL-Linear	2.96 (0.14)	0.39 (0.04)	3.00 (0.02)	0.38 (0.01)
OWL-Gaussian	2.97 (0.11)	0.39 (0.03)	3.00 (0.03)	0.38 (0.01)
RWL-Linear	2.90 (0.18)	0.40 (0.05)	3.00 (0.05)	0.38 (0.01)
RWL-Gaussian	2.88 (0.18)	0.41 (0.05)	2.96 (0.09)	0.39 (0.02)
RWL-VS-Gaussian	**3.17 (0.33)**	**0.33 (0.09)**	**3.82 (0.13)**	**0.09 (0.05)**
Scenario 4 (Optimal value 3.87)				
ℓ_1-PLS	2.32 (0.09)	**0.49 (0.05)**	2.36 (0.09)	0.52 (0.05)
OWL-Linear	2.41 (0.14)	0.55 (0.08)	2.48 (0.11)	0.59 (0.06)
OWL-Gaussian	**2.43 (0.13)**	0.55 (0.07)	2.47 (0.11)	0.58 (0.06)
RWL-Linear	2.38 (0.13)	0.53 (0.07)	2.44 (0.12)	0.56 (0.07)
RWL-Gaussian	2.37 (0.13)	0.53 (0.07)	2.44 (0.12)	0.56 (0.07)
RWL-VS-Gaussian	2.37 (0.12)	0.51 (0.05)	**3.47 (0.47)**	**0.21 (0.15)**

Table 3.4: *Comparison of methods for estimating ITRs on the EPIC data with the Pa endpoint. Higher values are better.*

	Predicted treatment #cycled vs #culture	Mean value (std)	Prop. of sig. p-values	Median p-value
ℓ_1-PLS	168 : 115	0.913 (0.040)	1	0.004
OWL-Linear	283 : 0	0.894 (0.054)	0	0.111
OWL-Gaussian	283 : 0	0.894 (0.054)	0	0.111
RWL-Linear	178 : 105	0.896 (0.046)	0.37	0.072
RWL-Gaussian	169 : 114	0.900 (0.045)	0.64	0.0404
RWL-VS-Linear	168 : 115	0.909 (0.044)	0.90	0.002
RWL-VS-Gaussian	168 : 115	0.907 (0.045)	0.85	0.007
Fixed rule (cycled)	283 : 0	0.894 (0.054)	0	0.111
Fixed rule (culture-based)	0 : 283	0.864 (0.054)	0	0.889

Table 3.5: *Comparison of methods for estimating ITR on the EPIC data with the PE endpoint. Lower annual PE rates are better.*

	Predicted treatment #cycled vs #culture	Mean annual PE rate (std)	Prop. of sig. p-values	Median p-value
PoissonReg	150 : 133	0.742 (0.224)	0.21	0.245
ℓ_1-PoissonReg	120 : 163	0.707 (0.233)	0.55	0.042
OWL-Linear	141 : 142	0.722 (0.236)	0.27	0.123
OWL-Gaussian	140 : 143	0.714 (0.247)	0.33	0.079
RWL-Linear	105 : 178	0.673 (0.226)	0.82	0.013
RWL-Gaussian	108 : 175	0.667 (0.220)	0.90	0.008
RWL-VS-Linear	105 : 178	0.676 (0.227)	0.75	0.013
RWL-VS-Gaussian	108 : 175	0.688 (0.218)	0.69	0.024
RWL-VS-Gaussian + RWL-Gaussian	102 : 181	0.639 (0.220)	0.97	0.001
Fixed rule (cycled)	283 : 0	0.725 (0.245)	0	0.140
Fixed rule (culture-based)	0 : 283	0.823 (0.253)	0	0.860

CHAPTER 4: QUALITATIVE TREATMENT-COVARIATE INTERACTIONS

4.1 Introduction

Medical researchers typically examine response to treatment for heterogeneous samples of subjects in a clinical trial. For example, they may want to know whether a treatment effect is different in older subjects versus younger subjects, or in men versus women. If for a group of subjects, a treatment is beneficial, but it is harmful for others, this should be discovered. Examining only the overall treatment effect may obscure an important effect of treatment in some specific subgroups.

Interaction is one of the fundamental concepts of statistical analysis. For personalized medicine, it is important to understand the treatment interaction, which is the non-random variability in the direction or magnitude of the treatment effect. Peto (1982) has distinguished quantitative interactions from qualitative interactions. A qualitative interaction arises when the direction of the treatment effect changes among different subsets of patients, while for the quantitative interaction, the treatment effect varies only in the magnitude, but not the direction.

Qualitative interactions are relatively rare (Yusuf et al. 1991), but when they do occur they are usually quite useful for personalized medicine. Most existing approaches to test qualitative interactions focus on predefined subgroups (Gail and Simon 1985, Zelterman 1990, Li and Chan 2006). Optimal treatment regimes have been recently received increasing attention in the statistical community. The target of optimal treatment regimes is just the qualitative interaction. In this work, we formulate hypotheses for qualitative interactions in the language of optimal treatment regimes, and propose a permutation test

for qualitative interactions without any prior information on the subgroups.

The remainder of the chapter is organized as follows. In Section 4.2, we review the residual weighted learning (RWL), and propose a modified version, called mirrored residual weighted learning (MRWL), to reduce the computational cost. Parameter tuning is an important issue in statistical learning. We extend the .632+ bootstrap to optimal treatment regime for parameter tuning. Then we propose a permutation test for qualitative interactions. In Section 4.3, we present simulation studies to evaluate performance of the proposed permutation test. The method is then illustrated on the two randomized clinical trials, the Nefazodone-CBASP clinical trial (Keller et al. 2000) and the EPIC cystic fibrosis randomized clinical trial (Treggiari et al. 2009; 2011), in Section 4.4. We conclude the article in Section 4.5.

4.2 Method

4.2.1 Context and notations

Consider a two-arm randomized trial. We observe a triplet $(\boldsymbol{X}, A, R)$ from each patient, where $\boldsymbol{X} = (X_1, \cdots, X_p)^T \in \mathcal{X}$ denotes the patient's clinical covariates, $A \in \mathcal{A} = \{1, -1\}$ denotes the treatment assignment, and R is the observed clinical outcome. Assume without loss of generality that larger values of R are more desirable. Let $\pi(a, \boldsymbol{x}) := P(A = a|\boldsymbol{X} = \boldsymbol{x})$ be the probability of being assigned treatment a for patients with clinical covariates $\boldsymbol{x}$. It is predefined in the trial design. We assume $\pi(a, \boldsymbol{x}) > 0$ for all $a \in \mathcal{A}$ and $\boldsymbol{x} \in \mathcal{X}$.

A treatment regime is a function from $\mathcal{X}$ to $\mathcal{A}$. An optimal treatment regime is a regime that maximizes the expected outcome under this regime, $\mathbb{E}(R|A = d(\boldsymbol{X}))$. This expectation, denoted by $\mathcal{V}(d)$, is called the value function associated with the regime d.

Assuming that the data generating mechanism is known, the optimal treatment regime is dictated by the following Bayes rule,

$$d^*(\boldsymbol{x}) = \text{sign}\Big(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)\Big),$$

where $\text{sign}(u) = 1$ for $u > 0$ and -1 otherwise. We are particularly interested in a non-informative rule d_0, *i.e.*, all subjects are assigned to a same treatment regardless of their clinical covariates. The optimal non-informative rule is defined as,

$$d_0^*(\boldsymbol{x}) = \text{sign}\Big(\mathbb{E}(R|A=1) - \mathbb{E}(R|A=-1)\Big), \quad \text{for any } \boldsymbol{x} \in \mathcal{X}.$$

4.2.2 Residual weighted learning

Mathematically, the value function associated with a regime d is given as

$$\mathcal{V}(d) = \mathbb{E}(R|A=d(\boldsymbol{X})) = \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A=d(\boldsymbol{X})\big)\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function (Qian and Murphy 2011). Thus finding d^* is equivalent to the following minimization problem:

$$d^* \in \arg\min_d \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A \neq d(\boldsymbol{X})\Big)\right). \tag{4.1}$$

Zhao et al. (2012) viewed this as a weighted classification problem, and proposed outcome weighted learning (OWL) to apply state-of-the-art support vector machines for implementation. This approach opens the door to application of statistical learning techniques to optimal treatment regimes. However, as pointed out by Zhou et al. (2015), this approach is not perfect. First, the estimated regime of OWL is affected by a simple shift of the outcome R. Hence estimates from OWL are unstable especially when the sample size is small. Second, since OWL needs the outcome to be nonnegative to gain computational efficiency from convex programming, OWL works similarly to weighted classification, whose intention is to reduce misclassification errors, the differences between the estimated and true

treatments. Thus the regime by OWL tends to keep the treatments that subjects actually received. This behavior is not ideal for data from a randomized clinical trial, since treatments are actually randomly assigned to patients.

To alleviate these problems, Zhou et al. (2015) proposed residual weighted learning (RWL), in which the misclassification errors are weighted by residuals of the outcome R from a regression fit on clinical covariates $\boldsymbol{X}$. The residuals are calculated as

$$\widehat{R} = R - \widehat{g}(\boldsymbol{X}), \tag{4.2}$$

where $\widehat{g}$ is an estimate of $\mathbb{E}(\frac{R}{2\pi(A,\boldsymbol{X})}|\boldsymbol{X})$. $\widehat{g}$ can be obtained by using weighted linear models, weighted generalized linear models or other nonparametric regression methods, such as support vector regression (Vapnik 1998) and random forests (Breiman 2001). Unlike OWL in (4.1), RWL targets the following optimization problem,

$$d^* \in \arg\min_d \mathbb{E}\left(\frac{\widehat{R}}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A \neq d(\boldsymbol{X})\Big)\right). \tag{4.3}$$

The use of residuals in optimal treatment regimes is justified as follows, for any measurable function g,

$$\mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A \neq d(\boldsymbol{X})\Big)\right) = \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A \neq d(\boldsymbol{X})\Big)\right) - \mathbb{E}\left(g(\boldsymbol{X})\right).$$

Assume that the observed data $\{(\boldsymbol{X}_i, A_i, R_i) : i = 1, \cdots, n\}$ are collected independently. For any decision function $f(\boldsymbol{X})$, let $d_f(\boldsymbol{X}) = \text{sign}\big(f(\boldsymbol{X})\big)$ be the associated regime. RWL aims to minimize the following regularized empirical risk,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{R}_i}{\pi(A_i,\boldsymbol{X}_i)}T\Big(A_i f(\boldsymbol{X}_i)\Big) + \lambda||f||^2, \tag{4.4}$$

where $T(\cdot)$ is a continuous surrogate loss function, $||f||$ is some norm for f, and λ is a tuning parameter.

Since some residuals are negative, convex surrogate loss functions are not appropriate in (4.4). Zhou et al. (2015) considered a non-convex loss, the smoothed ramp loss function. The non-convex loss is shown to be robust to outliers (Wu and Liu 2007), which

is helpful for RWL, especially when residuals are poorly estimated. On the other hand, the non-convexity presents significant challenges for solving the optimization problem (4.4). Unlike convex functions, non-convex functions may possess local optima that are not global optima, and most of efficient optimization algorithms such as gradient descent and coordinate descent are only guaranteed to converge to a local optimum. Zhou et al. (2015) applied a difference of convex (d.c.) algorithm to address the non-convex optimization problem by solving a sequence of convex subproblems. However, the d.c. algorithm is still computationally intensive.

4.2.3 Mirrored residual weighted learning (MRWL)

As a desire to reduce the computational cost of RWL, we develop a modified RWL method, which still takes the advantage of efficient convex optimization. The modified method is based on a finding in Liu et al. (2015),

$$\mathbb{E}\left(\frac{|R|}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A\text{sign}(R)\neq d(\boldsymbol{X})\Big)\right)=\mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A\neq d(\boldsymbol{X})\Big)\right)+\mathbb{E}\left(\frac{R^{-}}{\pi(A,\boldsymbol{X})}\right),$$

where $R^{-}=\max(-R,0)$. Therefore finding d^{*} in (4.3) is equivalent to the following optimization problem,

$$d^{*}\in\arg\min_{d}\mathbb{E}\left(\frac{|\widehat{R}|}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A\text{sign}(\widehat{R})\neq d(\boldsymbol{X})\Big)\right), \tag{4.5}$$

where negative residuals are reflected to positive, and accordingly their treatments are switched to the opposites. We propose to minimize the following empirical risk to find a decision function f,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\mathbb{I}\Big(A_i\text{sign}(\widehat{R}_i)\neq d_f(\boldsymbol{X}_i)\Big).$$

Similarly with OWL and RWL, we instead seek the decision function f by minimizing a regularized surrogate risk,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)f(\boldsymbol{X}_i)\Big)+\lambda||f||^2, \tag{4.6}$$

where $T(\cdot)$ is a continuous surrogate loss function, $||f||$ is some norm for f, and λ is a tuning parameter controlling the trade-off between empirical risk and complexity of the decision function f. This method is called mirrored residual weighted learning (MRWL). As the weights $\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}$ are nonnegative, convex surrogate can be employed for efficient computation. In this article, we apply the Huberized hinge loss function (Wang et al. 2008),

$$\phi(u) = \begin{cases} 0 & \text{if } u \geq 1, \\ \frac{1}{4}(1-u)^2 & \text{if } -1 \leq u < 1, \\ -u & \text{if } u < -1. \end{cases}$$

Other loss functions, such as the hinge loss, can be also applied in MRWL. The Huberized hinge loss is smooth everywhere. Hence it has computational advantages in optimization.

MRWL possesses almost all desirable properties of RWL. First, by using residuals, MRWL stabilizes the variability introduced from the original outcome. Second, to minimize the empirical risk in (4.6), similarly with RWL, for subjects with positive residuals, MRWL tends to recommend the same treatments that subjects have actually received; for subjects with negative residuals, MRWL is apt to give the opposite treatments to what they have received. Thus due to the balance between positive and negative residuals, MRWL favors neither the treatments that subjects actually received nor their opposites. Third, MRWL is location-scale invariant with respect to the original outcomes. Specifically, the estimated regime from MRWL is invariant to a shift of the outcome; it is invariant to a scaling of the outcome with a positive number; the regime from RWL that maximizes the outcome is opposite to the one that minimizes the outcome. These are intuitively sensible. The only nice property of RWL that is not inherited by MRWL is the robustness to outliers because of the unbounded convex loss in MRWL. However, we may apply an appropriate method or model to estimate the residuals to reduce the probability of outliers. The computational advantage for MRWL is a big plus for our later proposed permutation test.

We are particular interested in a linear regime, which is simple and easy for interpretation. Xu et al. (2015) argued that for many popular models, it is sufficient to find an optimal linear rule. Consider a linear decision function $f(\boldsymbol{X}) = \boldsymbol{w}^T\boldsymbol{X} + b$. The associated regime d_f will assign a subject with clinical covariates $\boldsymbol{X}$ into treatment 1 if $\boldsymbol{w}^T\boldsymbol{X} + b > 0$ and -1 otherwise. In (4.6), we define $||f||$ as the Euclidean norm of $\boldsymbol{w}$. Then the minimization problem (4.6) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)\big(\boldsymbol{w}^T\boldsymbol{X}_i + b\big)\Big) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{4.7}$$

There are many efficient numerical methods for solving this smooth unconstrained optimization problems. One example is the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Nocedal 1980). L-BFGS is a quasi-Newton method that approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using a limited amount of computer memory.

In Appendix C.1, we provided more details about MRWL, including the nonlinear extension of MRWL. We compared finite sample performance of MRWL and RWL, and presented the simulation results in Appendix C.2. As evident from Table C.1, MRWL showed similar performance as RWL when the decision boundary was not very complicated.

Variable selection is an important topic in modern statistical research. Zhou et al. (2015) applied the elastic-net penalty for linear RWL. In this article, we also replace the ℓ_2-norm penalty in linear MRWL with the elastic-net penalty,

$$\lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w},$$

where $||\boldsymbol{w}||_1 = \sum_{j=1}^{p}|w_j|$ is the ℓ_1-norm, and $\lambda_1(>0)$ and $\lambda_2(>0)$ are regularization parameters.

The elastic-net penalized linear MRWL aims to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)\big(\boldsymbol{w}^T\boldsymbol{X}_i + b\big)\Big) + \lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{4.8}$$

It is popular to use coordinate descent methods to solve such an optimization problem (Friedman et al. 2010, Yang and Zou 2013). However, under some circumstance, coordinate descent methods may suffer from slow convergence. In this article, we apply the general iterative shrinkage and thresholding (GIST) algorithm (Gong et al. 2013). GIST updates all coordinates simultaneously through a gradient descent and a thresholding rule, and it is easy to implement. We do not delve into the details of GIST. Interested readers may refer to Gong et al. (2013). The obtained decision function is $\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{w}}^T\boldsymbol{x} + \widehat{b}$, and thus the estimated optimal treatment regime is the sign of $\widehat{f}(x)$.

4.2.4 Theoretical properties of elastic-net penalized linear MRWL

In this section, we establish theoretical properties for the elastic-net penalized linear MRWL. To the end, suppose that a sample $D_n = \{\boldsymbol{X}_i, A_i, R_i\}_{i=1}^n$, is independently drawn from a probability measure P on $\mathcal{X} \times \mathcal{A} \times \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^p$. For any treatment regime $d : \mathcal{X} \to \mathcal{A}$, the risk is defined as

$$\mathcal{R}(d) = \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}(A \neq d(\boldsymbol{X}))\right).$$

The regime that minimizes the risk is the Bayes regime $d^* = \arg\min_d \mathcal{R}(d)$, and the corresponding risk $\mathcal{R}^* = \mathcal{R}(d^*)$ is the Bayes risk. Recall that the Bayes regime is $d^*(\boldsymbol{x}) = \text{sign}(\mathbb{E}(R|A = 1, \boldsymbol{X} = \boldsymbol{x}) - \mathbb{E}(R|A = 1, \boldsymbol{X} = \boldsymbol{x})$.

For any measurable function g, let $\widehat{R}_g = R - g(\boldsymbol{x})$. We define the ϕ-risk as

$$\mathcal{R}_{\phi,g}(f) = \mathbb{E}\left(\frac{|\widehat{R}_g|}{\pi(A, \boldsymbol{X})}\phi\Big(A\text{sign}(\widehat{R}_g)f(\boldsymbol{X})\Big)\right).$$

In the theoretical analysis, we do not require g to be a regression fit of R. g can be any arbitrary measurable function.

Let $(\widetilde{\boldsymbol{w}}_{\lambda_1,\lambda_2}, \widetilde{b}_{\lambda_1,\lambda_2})$ be a solution of the following optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p, b\in\mathbb{R}} \lambda_1||\boldsymbol{w}||_1 + \frac{\lambda_2}{2}\boldsymbol{w}^T\boldsymbol{w} + \mathcal{R}_{\phi,g}(\boldsymbol{w}^T\boldsymbol{X} + b). \tag{4.9}$$

Let $\boldsymbol{w}_n, b_n$ be a minimizer of the following empirical optimization problem,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_{gi}|}{\pi(A_i,\boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)\big(\boldsymbol{w}^T\boldsymbol{X}_i+b\big)\Big)+\lambda_{1n}||\boldsymbol{w}||_1+\frac{\lambda_{2n}}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{4.10}$$

For simplicity, we denote $\widetilde{f}_{\lambda_1,\lambda_2}(\boldsymbol{X})=\widetilde{\boldsymbol{w}}^T_{\lambda_1,\lambda_2}\boldsymbol{X}+\widetilde{b}_{\lambda_1,\lambda_2}$ and $f_n(\boldsymbol{X})=\boldsymbol{w}_n^T\boldsymbol{X}+b_n$.

We require some assumptions for consistency analysis.

(A1) $\mathcal{X}$ is compact.

(A2) $\mathbb{E}\left(|\widehat{R}_g|/\pi(A,\boldsymbol{X})\right)<\infty$.

(A3) There exist $\lambda_1^*\geq 0$ and $\lambda_2^*\geq 0$ such that $\mathcal{R}(\text{sign}(\widetilde{f}_{\lambda_1,\lambda_2}))=\mathcal{R}^*$.

(A4) There is a unique solution $(\widetilde{\boldsymbol{w}}_{\lambda_1^*,\lambda_2^*},\widetilde{b}_{\lambda_1^*,\lambda_2^*})$ for (4.9) when $\lambda_1=\lambda_1^*$ and $\lambda_2=\lambda_2^*$.

Assumption (A1) is common in theoretical analysis. Assumption (A2) is weaker than the bounded assumption in the theoretical analysis of RWL. It is plausible in real data applications. Assumption (A3) is a key assumption. It implies that the Bayes rule can be approximated by the sign of a linear solution in (4.9) when the tuning parameters λ_1 and λ_2 are appropriately set. Assumption (A4) is related to identification. It generally holds when the dimensionality p is not high. The following theorem establishes consistency of the elastic-net penalized linear MRWL.

Theorem 4.2.1. *For any distribution P for $(\boldsymbol{X},A,R)$ satisfying assumptions (A1)∼(A4) and sequences of $\lambda_{1n}>0$ and $\lambda_{2n}>0$ such that $\lambda_{1n}\to\lambda_1^*$ and $\lambda_{2n}\to\lambda_2^*$, we have that in probability $\boldsymbol{w}_n\to\widetilde{\boldsymbol{w}}_{\lambda_1^*,\lambda_2^*}$, $b_n\to\widetilde{b}_{\lambda_1^*,\lambda_2^*}$, and $\mathcal{R}(\text{sign}(f_n))\to\mathcal{R}^*$.*

The consistency of $\boldsymbol{w}_n$ and b_n is a direct application of Theorem 2.7 in Newey and McFadden (1986). The consistency of the risk $\mathcal{R}(\text{sign}(f_n))$ is a simple exercise of dominated convergence theorem. Here we skip the proof.

4.2.5 Parameter tuning

There are two parameters in the elastic-net penalized MRWL. Note that the parameter λ_1 controls variable selection, and hence it is sensitive to the performance. Cross-validation is a commonly used method for tuning parameters in optimal treatment regimes (Zhao et al. 2012, Zhou et al. 2015). However, when the parameters are sensitive to performance, cross-validation may not be appropriate. First, cross-validation does not take into consideration the performance when applying the parameters on the whole data. Consider a case where there is only one covariate related to the optimal treatment regime. In the cross-validation procedure, we may obtain the tuned parameters which only identify the correct covariate in the cross-validation procedure, but it is possible that the tuned parameters may result in a non-informative regime on the whole data, in which none of covariates are identified. Second, consider two optimization problems for MRWL. One is (4.8), the other is as follows,

$$\sum_{i=1}^{n} \frac{|\widehat{R}_i|}{\pi(A_i, \boldsymbol{X}_i)} \phi\Big(A_i \text{sign}(\widehat{R}_i)\big(\boldsymbol{w}^T \boldsymbol{X}_i + b\big)\Big) + \lambda_1' ||\boldsymbol{w}||_1 + \frac{\lambda_2'}{2} \boldsymbol{w}^T \boldsymbol{w}. \tag{4.11}$$

They are equivalent with respect to $\lambda_1' = n\lambda_1$ and $\lambda_2' = n\lambda_2$. However, when we perform k-fold cross-validation on them, the tuned parameters are not equivalent any more. Specifically, suppose the tuned parameters in (4.8) are λ_1^* and λ_2^*. Then the tuned parameters in (4.11) must be $\frac{k-1}{k} n\lambda_1^*$ and $\frac{k-1}{k} n\lambda_2^*$. The additional scale $\frac{k-1}{k}$ may have a large impact when the parameters are sensitive to performance.

In the literature on error rate estimation for patter classification, .632+ bootstrap (Efron and Tibshirani 1997) is an alternative resampling method to cross-validation. It has several nice properties that may be helpful to alleviate the above-mentioned problems of cross-validation. For instance, .632+ bootstrap estimator takes into account the performance on the whole data, and the bootstrap resampling keeps the sample size unchanged. In this section, we develop a .632+ bootstrap estimator of the value function in optimal treatment

regimes.

Assume that the observed data $\boldsymbol{Z} = \{(\boldsymbol{X}_i, A_i, R_i) : i = 1, \cdots, n\}$ are collected independently. For simplicity, denote $\boldsymbol{Z}_i = (\boldsymbol{X}_i, A_i, R_i)$. For a certain optimal treatment regime algorithm, let $d_{\boldsymbol{Z}}$ be the regime based on the training data $\boldsymbol{Z}$. We want to seek a bootstrap estimator of the value function associated with $d_{\boldsymbol{Z}}$.

Ordinary bootstrap samples $\boldsymbol{Z}^b = (\boldsymbol{Z}_1^b, \cdots, \boldsymbol{Z}_n^b)$ are generated with replacement from $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)$. A total of B such samples are independently drawn, say $\boldsymbol{Z}^1, \cdots, \boldsymbol{Z}^B$. Let N_i^b be the number of times that $\boldsymbol{Z}_i$ is included in the b-th bootstrap sample and define,

$$I_i^b = \begin{cases} 1 & \text{if } N_i^b = 0, \\ 0 & \text{if } N_i^b > 0. \end{cases}$$

Define $d_i^b = d_{\boldsymbol{Z}^b}(\boldsymbol{X}_i)$ and $d_i = d_{\boldsymbol{Z}}(\boldsymbol{X}_i)$. Then, the leave-one-out bootstrap value is

$$\widehat{\mathcal{V}}^{(1)} = \frac{\sum_{i=1}^n R_i \widehat{P}_i / \pi(A_i, \boldsymbol{X}_i)}{\sum_{i=1}^n \widehat{P}_i / \pi(A_i, \boldsymbol{X}_i)}, \qquad \text{where } \widehat{P}_i = \frac{\sum_b I_i^b \mathbb{I}(A_i = d_i^b)}{\sum_b I_i^b}.$$

The resubstitution value is the value function of $d_{\boldsymbol{Z}}$ estimated on the training data,

$$\widehat{\mathcal{V}}^{\text{Resub}} = \frac{\sum_{i=1}^n R_i \mathbb{I}(A_i = d_i) / \pi(A_i, \boldsymbol{X}_i)}{\sum_{i=1}^n \mathbb{I}(A_i = d_i) / \pi(A_i, \boldsymbol{X}_i)}.$$

The .632+ bootstrap value is a combination of the leave-one-out bootstrap value and the resubstitution value,

$$\widehat{\mathcal{V}}^{.632+} = (1 - \widehat{v}) \cdot \widehat{\mathcal{V}}^{\text{Resub}} + \widehat{v} \cdot \widehat{\mathcal{V}}^{(1)}, \tag{4.12}$$

where $\widehat{v}$ is related to the no-information value $\mathcal{V}^0$ of $d_{\boldsymbol{Z}}$, that would apply if $\boldsymbol{X}$ and (R, A) were independent. An estimate of $\mathcal{V}^0$ can be obtained by permuting (R_i, A_i) and $\boldsymbol{X}_j$,

$$\widehat{\mathcal{V}}^0 = \frac{\sum_{i=1}^n \sum_{j=1}^n R_i \mathbb{I}(A_i = d_j) / \pi(A_i, \boldsymbol{X}_i)}{\sum_{i=1}^n \sum_{j=1}^n \mathbb{I}(A_i = d_j) / \pi(A_i, \boldsymbol{X}_i)}.$$

Then, the weight $\widehat{v}$ is given by

$$\widehat{v} = \frac{.632(\widehat{\mathcal{V}}^{\text{Resub}} - \widehat{\mathcal{V}}^0)_+}{.632(\widehat{\mathcal{V}}^{\text{Resub}} - \widehat{\mathcal{V}}^0)_+ + .368(\widehat{\mathcal{V}}^{.632+} - \widehat{\mathcal{V}}^0)_+},$$

where $(\cdot)_+$ is the positive part. Here we define $0/0 = 0$.

Now we give the formula to compute the .632+ bootstrap estimator of the value function. It is straightforward to use this estimator in parameter tuning. Taking the elastic-net penalized MRWL as an example, for a predefined pair of λ_1 and λ_2, we calculate the associated .632+ bootstrap value. Then we compare the estimated values from all predefined pairs of parameters to choose the pair giving the largest value. When ties occur, we favor the parameters with a parsimonious regime.

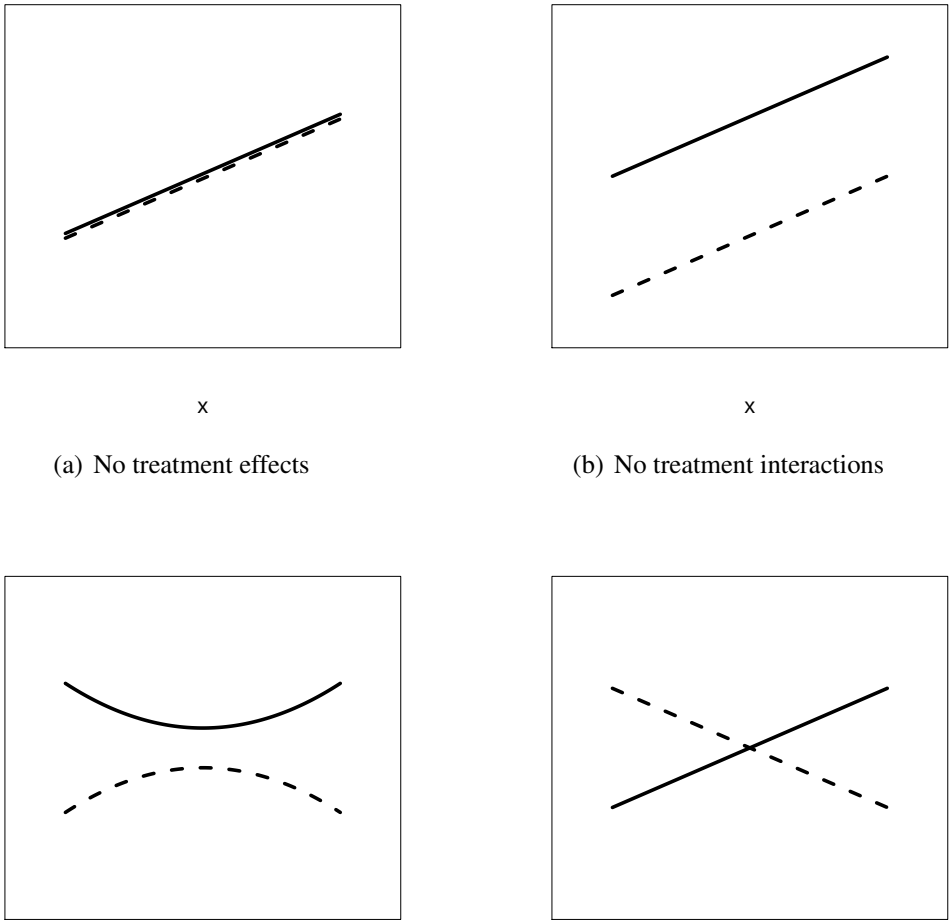
4.2.6 A permutation test for qualitative interaction

For personalized medicine, the distinction between quantitative and qualitative treatment-covariate interactions is important. A qualitative interaction arises when the direction of the treatment effect changes among different subsets of patients, while for the quantitative interaction, the treatment effect varies only in the magnitude, but not the direction.

Figure 4.1 shows possible relationships between treatment and covariates. Actually (A) is a special case of (B). However, case (A) of no treatment effects is important for our proposed permutation test. We will see later that (A) lies on the boundary between the null and alternative hypotheses.

There are currently a few qualitative interaction tests that can be used to test qualitative interactions on prespecified subgroups (Gail and Simon 1985, Zelterman 1990, Li and Chan 2006). These tests were not designed in an exploratory fashion to find the subgroups presenting qualitative interactions.

From the viewpoint of optimal treatment regimes, the common characteristic of (A)$\sim$(C) is that their optimal treatment regimes are non-informative. That is, traditional "one size fits all" approach still works when there are no qualitative interactions. Thus the null and



(a) No treatment effects

(b) No treatment interactions

(c) Quantitative interactions

(d) Qualitative interactions

Figure 4.1: *Different types of treatment-covariate relationships. The data consists of a single covariate X, treatment assignment $A = 1$ or -1, and outcome R. The conditional outcomes $\mathbb{E}(R|X = x, A = 1)$ and $\mathbb{E}(R|X = x, A = -1)$ are represented by solid lines (———) and dashed lines (– – –), respectively.*

alternative hypotheses for qualitative interactions are

$$\begin{aligned} H_0 : &\qquad \mathcal{V}(d^*) = \mathcal{V}(d_0^*) \\ H_a : &\qquad \mathcal{V}(d^*) > \mathcal{V}(d_0^*) \end{aligned} \tag{4.13}$$

A straightforward approach is to estimate the optimal treatment regime $\widehat{d}^*$ and the op-

timal non-informative regime $\widehat{d}_0^*$. The comparison between two estimated regimes can be performed using the test described in Murphy (2005). However, this method inflates the Type-I error rate. When we estimate $\widehat{d}^*$, we implicitly compare with many other regimes, say $d_1, \cdots, d_m$. Hence the hypotheses in (4.13) are equivalent to

$$\begin{aligned} H_0: &\quad \mathcal{V}(d_0^*) \geq \mathcal{V}(d_j), \quad j = 1, \cdots, m. \\ H_a: &\quad \mathcal{V}(d_0^*) < \mathcal{V}(d_j), \quad \text{for some } j \in \{1, \cdots, m\}. \end{aligned} \tag{4.14}$$

Multiplicity has to be considered. However, taking a linear regime as an example, according to Sauer's Lemma (Cover 1965), we simultaneously consider up to $C(n, p)$ linear regimes in comparison, where

$$C(n, p) = 2 \sum_{k=0}^{p-1} \binom{n-1}{k}.$$

When the sample size $n = 100$ and dimensionality $p = 5$, m in (4.14) is just $C(n, p)$, which is about 8×10^6. Traditional multiple comparison approaches would lack of power.

In this work, we propose a permutation test using the elastic-net penalized MRWL. It is easy to see that the non-informative regime corresponds to the solution with $\widehat{\boldsymbol{w}} = \mathbf{0}$. When $\widehat{\boldsymbol{w}}$ is close to $\mathbf{0}$, its value function is close to $\mathcal{V}(d_0^*)$. On the contrary, under the null hypothesis, the Bayes rule is a non-informative regime. When λ_1 is large enough, we can always find a unique solution with $\widetilde{\boldsymbol{w}} = \mathbf{0}$ in (4.9). By Theorem 4.2.1, we have $\boldsymbol{w}_n \to \mathbf{0}$ in probability. Thus we use the test statistic,

$$T = \max_{j \in \{1, \cdots, p\}} |\widehat{w}_j|. \tag{4.15}$$

Its distribution is complicated. We permute $\boldsymbol{X}_i$ to break their connections with (R_i, A_i) to generate the null distribution of T. The procedure of the proposed permutation test is summarized as follows.

1) Calculate $\widehat{R}_i$ by (4.2).

2) Tune the parameters λ_1 and λ_2 in MRWL to give the largest .632+ bootstrap value.

3) Estimate a linear regime $\widehat{f}(\boldsymbol{x}) = \widehat{w}^T\boldsymbol{x} + \widehat{b}$ by MRWL with the tuned parameters λ_1^* and λ_2^*.

4) The observed statistic $T^{\text{obs}} = \max_{j\in\{1,\cdots,p\}} |\widehat{w}_j|$.

5) Fix $(\widehat{R}_i, A_i)$, and permute $\boldsymbol{X}_i$ to generate the null distribution.

   (i) Random permute $\boldsymbol{X}_i$ to generate a sample $\{\boldsymbol{X}_1^m, \cdots, \boldsymbol{X}_n^m\}$.

   (ii) Tune the parameters λ_1 and λ_2 in MRWL to give the largest .632+ bootstrap value on the training data $\{(\widehat{R}_i, A_i, \boldsymbol{X}_i^m), i = 1, \cdots, n\}$.

   (iii) Estimate a linear regime $\widehat{f}^m(\boldsymbol{x}) = \widehat{w}^{mT}\boldsymbol{x} + \widehat{b}^m$ with the tuned parameters λ_1^{*b} and λ_2^{*b} on the training data $\{(\widehat{R}_i, A_i, \boldsymbol{X}_i^m), i = 1, \cdots, n\}$.

   (iv) The test statistic on the permuted sample is $T^m = \max_{j\in\{1,\cdots,p\}} |\widehat{w}_j^m|$.

   (v) Repeat step (i)-(iv) M times.

6) The p-value is $\frac{\sum_{m=1}^{M} \mathbb{I}(T^m \geq T^{\text{obs}})+1}{M+1}$.

4.3 Simulation studies

In this section, we use simulation studies to investigate performance of the proposed permutation test.

4.3.1 Type-I error rates

In the simulations, we generated 5-dimensional vectors of clinical covariates $\boldsymbol{X} = (x_1, \cdots, x_5)^T$, consisting of independent uniform random variables $U(-1, 1)$. The treatment A was generated from $\{-1, 1\}$ independently of $\boldsymbol{X}$ with $\pi(A, \boldsymbol{X}) = 0.5$. The response R was normally distributed with mean $Q_0 = \mu_0(\boldsymbol{X}) + \delta_0(\boldsymbol{X}) \cdot A$ and standard

deviation 1, where $\mu_0(\boldsymbol{X})$ is the common effect for clinical covariates, and $\delta_0(\boldsymbol{X}) \cdot A$ is the treatment-covariate interaction. To evaluate the performance of our proposed permutation test on controlling Type I errors, we considered two scenarios with different choices of $\mu_0(\boldsymbol{X})$ and $\delta_0(\boldsymbol{X})$. We estimated the residuals using a simple weighted linear regression model. In Scenario I, the linear model for residuals was correctly specified, while in Scenario II, it was misspecified.

(I): $\mu_0(\boldsymbol{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$.

I(A): $\delta_0(\boldsymbol{X}) = 0$.

I(B): $\delta_0(\boldsymbol{X}) = 0.45$.

I(C): $\delta_0(\boldsymbol{X}) = 0.25(x_1^2 + x_2^2)$.

(II): $\mu_0(\boldsymbol{X}) = 1 + x_1^2 + x_2^2 + 2x_3^2 + 0.5x_4^2$.

II(A): $\delta_0(\boldsymbol{X}) = 0$.

II(B): $\delta_0(\boldsymbol{X}) = 0.32$.

II(C): $\delta_0(\boldsymbol{X}) = 0.18(x_1^2 + x_2^2)$.

In each scenario, three cases were presented for (A) no treatment effects, (B) no treatment-covariate interactions, and (C) quantitative interactions, respectively. The coefficients in (B) were chosen to reflect a medium effect size according to Cohen's d index (Cohen 1988); and the coefficients in (C) reflect a small effect size. The Cohen's d index is defined as the standardized difference in mean responses between two treatment arms, that is,

$$es = \frac{|\mathbb{E}(R|A=1) - \mathbb{E}(R|A=-1)|}{\sqrt{[Var(R|A=1) + Var(R|A=-1)]/2}}.$$

The effect size was tentatively defined as "small" by Cohen (1988) if the Cohen's d index is 0.2, "medium" if the index is 0.5 and "large" if the index is 0.8.

Table 4.1: *Proportions of significant p values and Type-I error rates in Scenario I and II.*

	Scenario I				Scenario II			
Sample size	A	B	C	Type-I error	A	B	C	Type-I error
$n = 100$	0.040	0.025	0.031	0.040	0.047	0.029	0.041	0.047
$n = 200$	0.049	0.024	0.043	0.049	0.058	0.038	0.044	0.058
$n = 400$	0.056	0.007	0.032	0.056	0.053	0.028	0.045	0.053

For every case, we generated $1,000$ Monte Carlo datasets. On each dataset, we performed the proposed permutation test with $M = 1,000$. The parameters in MRWL were tuned using .632+ bootstrap with $B = 20$. The nominal significance level is $\alpha = 0.05$. For each case, we varied the sample sizes from $n = 100$, 200, to 400. The proportions of significant p values among $1,000$ Monte Carlo datasets are shown in Table 4.1. By definition, the Type-I error rate in each scenario is the maximal proportion of significant p values among its associated cases (A), (B) and (C). The Type-I error rates are also presented in Table 4.1. As evident from the table, the Type-I error rates are all derived from cases (A). The reason is that the case (A) is just on the boundary between the null and alternative hypotheses. Our proposed permutation test performed well in preserving the significance levels of the test. Even in Scenario II where the model to estimate residuals was misspecified, the permutation test still controlled the Type-I error well.

4.3.2 Power comparison

We used the similar simulation setups as in the Type-I error study. The clinical covariates $\boldsymbol{X}$ were 5-dimensional vectors, consisting of independent uniform random variables $U(-1, 1)$. Two treatment arms were balanced. The response R was normally distributed with mean $Q_0 = \mu_0(\boldsymbol{X}) + \delta_0(\boldsymbol{X}) \cdot A$ and standard deviation 1. We considered two scenarios:

(III) $\mu_0(\boldsymbol{X}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{X}) = \gamma(0.3 - x_1 - x_2)$.

(IV) $\mu_0(\boldsymbol{X}) = 1 + x_1^2 + x_2^2 + 2x_3^2 + 0.5x_4^2$; $\delta_0(\boldsymbol{X}) = \gamma(0.3 - x_1 - x_2)$.

γ controls the level of qualitative treatment-covariate interactions. Again, we used a simple linear regression model to estimate residuals. Thus, in Scenario III, the linear model was correctly specified, while in Scenario IV, it was misspecified.

For every scenario, we generated $1,000$ Monte Carlo datasets. On each dataset, we performed the proposed permutation test with $M = 1,000$, and tuned the parameters using .632+ bootstrap with $B = 20$. The sample sizes were varied from $n = 100$, 200, to 400. The results are reported in Table 4.2, with $\alpha = 0.05$ as the nominal significance level of the test.

We also considered the Gail-Simon test (Gail and Simon 1985) and the Zelterman test (Zelterman 1990) as comparisons. The Gail-Simon test is a widely used likelihood ratio test for qualitative interactions. However, it is well acknowledged that the Gail-Simon test is not very powerful especially when treatment effects are close to 0 (Zelterman 1990, Piantadosi and Gail 1993). Zelterman (1990) constructed an unbiased and locally most powerful test by modifying the Gail-Simon test. An interesting discussion between the Gail-Simon and Zelterman tests can be found in Silvapulle and Sen (2005, Section 9.3). Since both only apply to predefined subgroups, we used them in an oracle form. Specifically, we divided the data into two disjoint subgroups according to the sign of $(x_1 + x_2 - 0.3)$, and then tested qualitative interactions across the subgroups. The results are also given in Table 4.2.

Both the Gail-Simon and Zelterman tests failed to control the Type-I error (when $\gamma = 0$). The Gail-Simon test is very conservative and the Zelterman test is too liberal. As shown in previous section, our proposed permutation test controlled Type-I error well. The

Table 4.2: *Power comparison between the permutation test and the oracle Gail-Simon test in Scenario III and IV.*

		γ						
Sample size	Test	0	0.1	0.2	0.3	0.4	0.5	0.6
		Scenario III						
$n = 100$	Permutation test	0.040	0.050	0.123	0.250	0.464	0.647	0.790
	Oracle Gail-Simon test	0.005	0.009	0.019	0.065	0.111	0.180	0.260
	Oracle Zelterman test	0.129	0.124	0.139	0.170	0.190	0.242	0.322
$n = 200$	Permutation test	0.049	0.106	0.307	0.599	0.852	0.955	0.988
	Oracle Gail-Simon test	0.006	0.022	0.052	0.123	0.236	0.386	0.528
	Oracle Zelterman test	0.138	0.141	0.157	0.212	0.310	0.429	0.545
$n = 400$	Permutation test	0.056	0.161	0.562	0.906	0.991	0.999	0.999
	Oracle Gail-Simon test	0.008	0.028	0.132	0.269	0.465	0.648	0.787
	Oracle Zelterman test	0.128	0.143	0.217	0.315	0.485	0.653	0.788
		Scenario IV						
$n = 100$	Permutation test	0.047	0.057	0.095	0.172	0.306	0.460	0.630
	Oracle Gail-Simon test	0.008	0.013	0.037	0.083	0.167	0.285	0.434
	Oracle Zelterman test	0.119	0.123	0.149	0.163	0.223	0.329	0.452
$n = 200$	Permutation test	0.058	0.088	0.197	0.378	0.646	0.846	0.950
	Oracle Gail-Simon test	0.006	0.017	0.075	0.218	0.405	0.605	0.740
	Oracle Zelterman test	0.119	0.132	0.166	0.271	0.432	0.618	0.740
$n = 400$	Permutation test	0.053	0.116	0.351	0.738	0.928	0.989	0.997
	Oracle Gail-Simon test	0.006	0.045	0.197	0.473	0.704	0.859	0.932
	Oracle Zelterman test	0.139	0.169	0.263	0.508	0.709	0.859	0.932

poor performance of the Gail-Simon and Zelterman tests is probably due to the violation of assumptions for them. Both tests assume that the responses on both arms in each subgroups are independent, and the treatment effects follow normal distributions. These assumptions were not satisfied in the simulation studies.

The powers of our proposed permutation test, the Gail-Simon and Zelterman tests increase, as γ increases and as the sample size increases. As evident from the table, the permutation test was more powerful than both the Gail-Simon and Zelterman tests when $\gamma \geq 0.3$, even though the permutation test did not know the true boundary for qualitative

interactions. Under the misspecified model for residuals, the permutation test in Scenario IV still retained good power compared to either the Gail-Simon test or the Zelterman test. However, when compared to the permutation test in Scenario III, the test in Scenario IV is not that powerful. It suggests that an appropriate model for estimating residuals may increase the power of the proposed permutation test.

4.4 Data analysis

In this section, we applied our proposed permutation test to two clinical trial studies. The first one is the Nefazodone-CBASP clinical trial (Keller et al. 2000), and the second is the EPIC randomized clinical trial (Treggiari et al. 2011).

4.4.1 Nefazodone-CBASP trial

The Nefazodone-CBASP trial was designed to compare three different treatments for patients suffering chronic depression. Patients with non-psychotic chronic major depressive disorder (MDD) were randomized in a 1:1:1 ratio to either Nefazodone (NFZ), cognitive behavioral-analysis system of psychotherapy (CBASP), or the combination of Nefazodone and CBASP (COMB). The primary endpoint was the score on the 24-item Hamilton Rating Scale for Depression (HRSD). Lower HRSD is preferred. We used 50 pre-treatment covariates as in Zhao et al. (2012). Patients with missing covariates were excluded from the analyses. The data used here consisted of 647 patients: 216 in NFZ, 220 in CBASP, and 211 in COMB. Each clinical covariate was linearly scaled to the range $[-1, +1]$.

We carried out pairwise comparisons to test if there exist qualitative interactions between NFZ and CBASP, between NFZ and COMB, and between CBASP and COMB, respectively. We run the proposed permutation test with $M = 10,000$, and tuned the pa-

rameters using .632+ bootstrap with $B = 50$. Figure 4.2 shows the null distributions of the test statistic and the observed statistics. The corresponding p values are 0.511 (NFZ vs CBASP), 1 (NFZ vs COMB), and 1 (CBASP vs COMB), respectively. According to the data, there are no qualitative interactions presenting in the data.

4.4.2 EPIC cystic fibrosis trial

The trial was designed to determine the optimal anti-pseudomonal treatment strategy in children with cystic fibrosis (CF) who recently acquired Pseudomonas aeruginosa (*Pa*). Patients were randomized in a 1:1 ratio to one of two maintenance treatment strategies: cycled therapy and culture-based therapy. There were two endpoints over the course of the 18-month study. One was the number of *Pa* positive cultures from scheduled follow-up quarterly cultures, and the other is the number of pulmonary exacerbations (PE) requiring any (intravenous, inhaled, or oral) antibiotic use or hospitalization during the study. We considered 7 baseline clinical covariates. Patients with missing covariates were excluded from the analyses. The data used here consisted of 283 patients: 141 in the cycled therapy, and 142 in the culture-based therapy. Each clinical covariate was linearly scaled to the range $[-1, +1]$

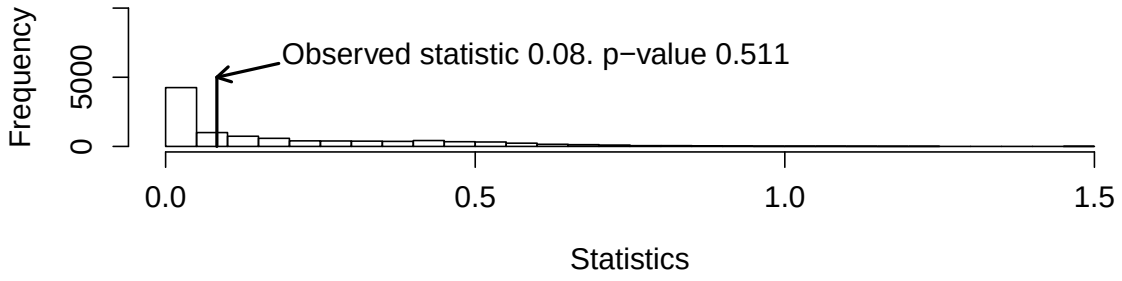
For the *Pa*-related endpoint, we used the ratio of the number of *Pa* negative cultures to the number of *Pa* cultures over the follow up period as the outcome, and treated it as continuous. A linear regression model was used to estimate residuals. We performed the proposed permutation test with $M = 10,000$, and tuned the parameters using .632+ bootstrap with $B = 50$. The null distribution is shown in Figure 4.3. The corresponding p value is 0.158. Zhou et al. (2015) identified a treatment regime only involving one covariate for the *Pa* endpoint. However, according to the permutation test here, their finding is probably spurious. This example demonstrates the risk of data-driven methods in the analysis. A

rigorous test is required to check the validity of the obtained regimes.

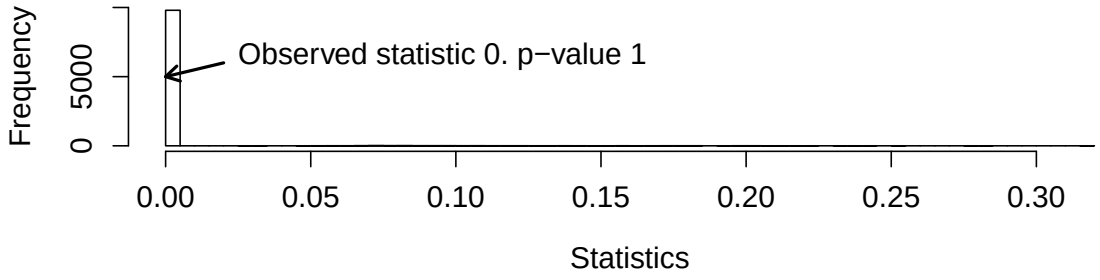
The number of PEs during the study was a rate variable. We used a Poisson regression model to estimate residuals. The same values of M and B were used to carry out the proposed permutation test. The null distribution is also given in Figure 4.3. The corresponding p value for the observed test statistic is 0.039. Although there is no significant difference between the two maintenance treatment strategies at the population level for the PE outcome (p-value 0.280), qualitative interactions do exist. Correctly identifying the qualitative interactions may improve the clinical practice.

4.5 Conclusion

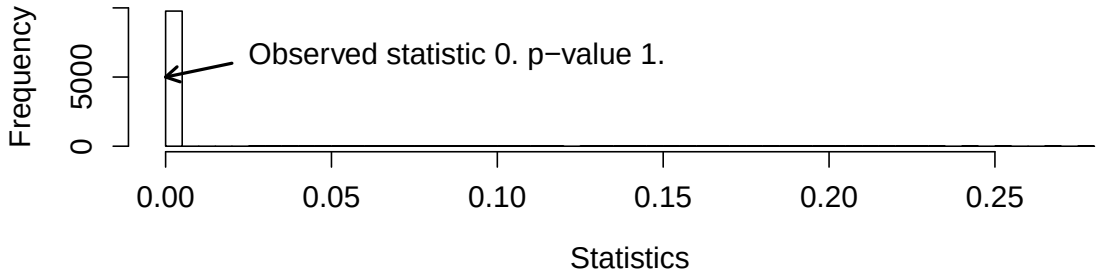
In this work, we propose a permutation test for qualitative interactions. In the test, we first estimate the optimal treatment regime using mirrored residual weighted learning, and then apply a permutation test to determine the difference between the estimated regime and non-informative regime. From the simulation studies and case studies, our proposed permutation test controls the Type-I error rate, and is also powerful to detect the qualitative interaction.



(a) NFZ vs CBASP

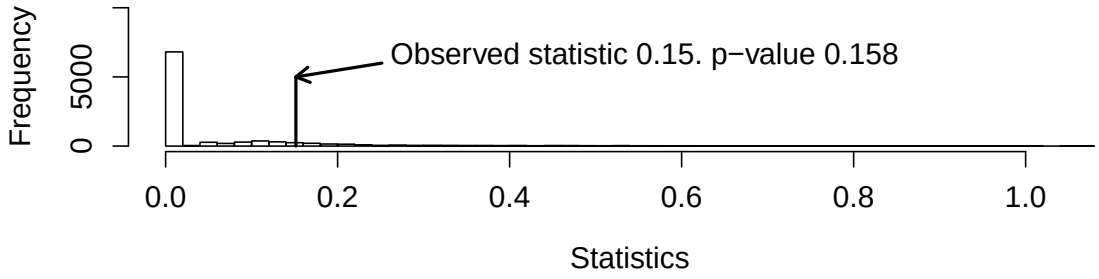


(b) NFZ vs COMB

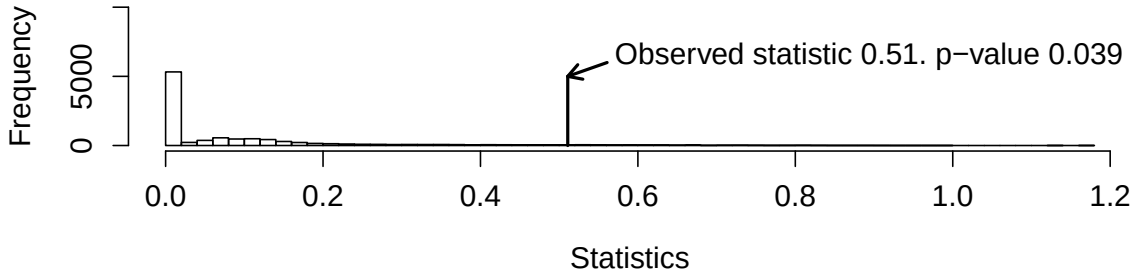


(c) CBASP vs COMB

Figure 4.2: *Histograms of the null distributions and observed statistics for the proposed permutation test on the Nefazodone-CBASP trial data.*



(a) *Pa* endpoint



(b) PE endpoint

Figure 4.3: *Histograms of the null distributions and observed statistics for the proposed permutation test on the EPIC trial data.*

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH PLAN

Three methods for optimal treatment regimes are proposed in the dissertation: nearest neighbor rules, residual weighted learning (RWL), and mirrored residual weighted learning (MRWL). A permutation test for qualitative interactions is also proposed by using MRWL. There are some subsequent works that we would like to explore or finish in the near future:

- RWL is a powerful tool to detect optimal treatment regimes. However, it can only apply to two-arm trials. In practice, some trials involve multiple arms, *i.e.* have three or more treatment arms. It would thus be worthwhile to extend RWL to multiple-arm trials.
- Recently, subgroup identification has been received more attention in the pharmaceutical industry, since it may help to salvage some new drugs and improve overall success. It is interesting to extend the proposed methods in this dissertation to the area of subgroup identification. Inference related to subgroup findings is very challenging in this topic.
- We have proposed a permutation test for qualitative interaction. However, due to the computational cost, our proposal is limited to linear regimes. In practical use, nonlinear optimal treatment regimes do exist. It would be worthwhile to investigate qualitative interactions with nonlinear treatment regimes. The nearest neighbor regime proposed in this dissertation is a good candidate to address the nonlinear regime in a permutation test.

APPENDIX A: PROOFS FOR CHAPTER 2

We prove Theorem 2.2.1 and 2.2.2 in Appendix A.1 and A.2. The proofs are based on theoretical results for nearest neighbor rules in regression. For completeness, we collect the theorems and lemmas needed in the proofs in Appendix A.3.

A.1 Proof of Theorem 2.2.1

The following lemma shows that consistency of $\widehat{m}_\ell(\boldsymbol{x})$, $\ell = 1, \cdots, L$, guarantees consistency of the rule d^{NN}.

Lemma A.1.1. *The k-NN rule in (2.3) satisfies the following bound for any distribution P for $(\boldsymbol{X}, A, R)$,*

$$\mathcal{V}(d^*) - \mathcal{V}(d^{NN}) \leq \sum_{\ell=1}^{L} \int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \mu(d\boldsymbol{x}).$$

Proof of Lemma A.1.1: Note that the value function of any rule d,

$$\mathcal{V}(d) := \mathbb{E}(R|A = d(\boldsymbol{X})) = \mathbb{E}\left(\frac{R\mathbb{I}(A = d(\boldsymbol{X}))}{\pi_A(\boldsymbol{X})}\right).$$

Refer to Qian and Murphy (2011) and Zhao et al. (2012) regarding the derivation. Thus, by fixing $\boldsymbol{x} \in \mathbb{R}^p$, we have

$$\begin{aligned} & \mathbb{E}\left(\frac{R\mathbb{I}(A = d^*(\boldsymbol{X}))}{\pi_A(\boldsymbol{X})}\big|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R\mathbb{I}(A = d^{NN}(\boldsymbol{X}))}{\pi_A(\boldsymbol{X})}\big|\boldsymbol{X} = \boldsymbol{x}\right) \\ = & \sum_{\ell=1}^{L} m_\ell(\boldsymbol{x}) \left(\mathbb{I}(d^*(\boldsymbol{x}) = \ell) - \mathbb{I}(d^{NN}(\boldsymbol{x}) = \ell)\right) \\ = & \ m_{\ell_1}(\boldsymbol{x}) - m_{\ell_2}(\boldsymbol{x}), \end{aligned}$$

where $\ell_1 = d^*(\boldsymbol{x})$ and $\ell_2 = d^{NN}(\boldsymbol{x})$, and the expectation $\mathbb{E}$ is with respect to P for

$(\boldsymbol{X}, A, R)$. By the construction of $d^{NN}(\boldsymbol{x})$, we have

$$\begin{aligned} & m_{\ell_1}(\boldsymbol{x}) - m_{\ell_2}(\boldsymbol{x}) \\ \leq \; & (m_{\ell_1}(\boldsymbol{x}) - \widehat{m}_{\ell_1}(\boldsymbol{x})) - (m_{\ell_2}(\boldsymbol{x}) - \widehat{m}_{\ell_2}(\boldsymbol{x})) \\ \leq \; & \sum_{\ell=1}^{L} |m_\ell(\boldsymbol{x}) - \widehat{m}_\ell(\boldsymbol{x})| \,. \end{aligned}$$

The desired result follows by taking expectation over $\boldsymbol{X}$ on both sides. □

Now it is sufficient to prove, for any $\ell \in \{1, \cdots, L\}$,

$$\int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \mu(d\boldsymbol{x}) \to 0$$

in probability or almost surely, as $n \to \infty$. We start from a simpler k-nearest neighbor rule, for $\ell \in \{1, \cdots, L\}$,

$$\widehat{m}'_\ell(\boldsymbol{x}) = \sum_{i=1}^{k} R_{(i,n)}(\boldsymbol{x}) \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{k\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}. \tag{A.1}$$

The relationship between $\widehat{m}_\ell(\boldsymbol{x})$ and $\widehat{m}'_\ell(\boldsymbol{x})$ is that

$$\widehat{m}_\ell(\boldsymbol{x}) = \frac{\widehat{m}'_\ell(\boldsymbol{x})}{\frac{1}{k}\sum_{i=1}^{k} \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}}.$$

By the law of large numbers, the denominator

$$\frac{1}{k}\sum_{i=1}^{k} \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)} \to 1 \quad \text{a.s.}$$

as $k \to \infty$. Thus it is now sufficient to prove, for any $\ell \in \{1, \cdots, L\}$,

$$\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \mu(d\boldsymbol{x}) \to 0$$

in probability or almost surely, as $n \to \infty$.

For weak consistency, we will prove a slightly stronger result,

$$\mathbb{E}\left(\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \mu(d\boldsymbol{x})\right) \to 0.$$

We proceed by checking the conditions of Stone's Theorem in Appendix A.3. The weights

$$V_{n,i}^{\ell}(\boldsymbol{x}) = \begin{cases} \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{k\pi_{\ell}\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}, & \text{if } i \leq k, \\ 0 & \text{if } i > k. \end{cases}$$

Conditions (ii) and (iv) of Stone's Theorem are automatically satisfied. Condition (v) is obvious since $k \to \infty$ and when $i \leq k$, $V_{n,i}^{\ell}(\boldsymbol{X}) \to 1/k$ in probability. For condition (iii) observe that, for each $a > 0$,

$$\begin{aligned} & \mathbb{E}\left\{\sum_{i=1}^{n} |V_{n,i}^{\ell}(\boldsymbol{X})| \mathbb{I}(||\boldsymbol{X}_i - \boldsymbol{X}|| > a)\right\} \\ = & \int \mathbb{E}\left\{\sum_{i=1}^{n} |V_{n,i}^{\ell}(\boldsymbol{x})| \mathbb{I}(||\boldsymbol{X}_i - \boldsymbol{x}|| > a)\right\} \mu(d\boldsymbol{x}) \\ \leq & \int \mathbb{E}\left\{\frac{1}{k\zeta}\sum_{i=1}^{k} \mathbb{I}(||\boldsymbol{X}_{(i,n)}(\boldsymbol{x}) - \boldsymbol{x}|| > a)\right\} \mu(d\boldsymbol{x}) \\ \leq & \frac{1}{\zeta}\int P(||\boldsymbol{X}_{(k,n)}(\boldsymbol{x}) - \boldsymbol{x}|| > a)\mu(d\boldsymbol{x}). \end{aligned}$$

For $\boldsymbol{x} \in support(\mu)$, when $k/n \to 0$, Lemma A.3.1 implies $P(||\boldsymbol{X}_{(k,n)}(\boldsymbol{x}) - \boldsymbol{x}|| > a) \to 0$. Then the dominated convergence theorem implies condition (iii). Finally, we consider condition (i). For any nonnegative measurable function f with $\mathbb{E}(f(\boldsymbol{X})) < \infty$ and any n,

$$\begin{aligned} & \mathbb{E}\left\{\sum_{i=1}^{n} V_{n,i}^{\ell}(\boldsymbol{X}) f(\boldsymbol{X}_i)\right\} \\ \leq & \frac{1}{k\zeta}\mathbb{E}\left\{\sum_{i=1}^{k} f(\boldsymbol{X}_{(i,n)}(\boldsymbol{X}))\right\} \\ \leq & \frac{\gamma_d}{\zeta}\mathbb{E}(f(\boldsymbol{X})). \end{aligned}$$

The last inequality is due to Lemma A.3.4. Condition (i) is verified. By Stone's theorem,

$$\mathbb{E}\int |\widehat{m}'_{\ell}(\boldsymbol{x}) - m_{\ell}(\boldsymbol{x})|\mu(d\boldsymbol{x}) \to 0.$$

Weak consistency of the kNN rule (2.3) follows by Lemma A.1.1.

We next show strong consistency for bounded outcomes. By the Borel-Cantelli Lemma, it suffices to show the following theorem. The proof follows an idea in Devroye et al. (1996, Chapter 11).

Theorem A.1.1. *For any distribution P for $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1), (A3), and $|R| \leq M < \infty$ for some constant M, if $k \to \infty$ and $k/n \to 0$, then for every $\epsilon > 0$ there exists an n_0 such that for $n \geq n_0$*

$$P\left(\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| d(\boldsymbol{x}) > \epsilon\right) \leq 2\exp(-cn\epsilon^2),$$

where $c > 0$ depends only upon the dimension p, M and ζ.

Proof of Theorem A.1.1: Fix $\boldsymbol{x} \in \mathbb{R}^p$. Denote $\rho_n(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{X}_{(k,n)}(\boldsymbol{x})||$. Also define $\rho_n^*(\boldsymbol{x})$ as the solution of the equation $\frac{k}{n} = \mu(S_{\boldsymbol{x},\rho_n^*(\boldsymbol{x})})$. Since distance ties occur with probability zero in μ, the solution always exists. And define the rule

$$\widehat{m}_\ell^*(\boldsymbol{x}) = \sum_{i=1}^n R_i \frac{\mathbb{I}(A_i = \ell)}{k\pi_\ell(\boldsymbol{X}_i)} \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right).$$

Consider the following decomposition,

$$|\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \leq |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{m}_\ell^*(\boldsymbol{x})| + |\widehat{m}_\ell^*(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|.$$

For the first term on the right-hand side, we obtain,

$$\begin{aligned}
& |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{m}_\ell^*(\boldsymbol{x})| \\
= \; & \frac{1}{k}\Big| \sum_{i=1}^n R_i \frac{\mathbb{I}(A_i = \ell)}{\pi_\ell(\boldsymbol{X}_i)} \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n(\boldsymbol{x})\right) - \sum_{i=1}^n R_i \frac{\mathbb{I}(A_i = \ell)}{\pi_\ell(\boldsymbol{X}_i)} \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right) \Big| \\
= \; & \frac{1}{k}\Big| \sum_{i=1}^n R_i \frac{\mathbb{I}(A_i = \ell)}{\pi_\ell(\boldsymbol{X}_i)} \Big(\mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n(\boldsymbol{x})\right) - \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right) \Big) \Big| \\
\leq \; & \frac{1}{k} \sum_{i=1}^n \Big| R_i \frac{\mathbb{I}(A_i = \ell)}{\pi_\ell(\boldsymbol{X}_i)} \Big(\mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n(\boldsymbol{x})\right) - \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right) \Big) \Big| \\
\leq \; & \frac{M}{k\zeta} \sum_{i=1}^n \Big| \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n(\boldsymbol{x})\right) - \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right) \Big| \\
= \; & \frac{M}{\zeta}\Big| \frac{1}{k} \sum_{i=1}^n \mathbb{I}\left(||\boldsymbol{x} - \boldsymbol{X}_i|| \leq \rho_n^*(\boldsymbol{x})\right) - 1 \Big| = \frac{M}{\zeta}\Big| \frac{1}{k} \sum_{i=1}^n \mathbb{I}\left(\boldsymbol{X}_i \in S_{\boldsymbol{x},\rho_n^*(\boldsymbol{x})}\right) - 1 \Big|.
\end{aligned}$$

Denote $\widehat{s}(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^n \mathbb{I}\left(\boldsymbol{X}_i \in S_{\boldsymbol{x},\rho_n^*(\boldsymbol{x})}\right)$. Thus,

$$|\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})| \leq \frac{M}{\zeta}|\widehat{s}(\boldsymbol{x}) - 1| + |\widehat{m}_\ell^*(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|. \tag{A.2}$$

Observe that $\mathbb{E}(\widehat{s}(\boldsymbol{x})) = 1$, then we have,

$$\begin{aligned}\mathbb{E}\left\{\int |\widehat{s}(\boldsymbol{x}) - 1|\mu(d\boldsymbol{x})\right\} &\leq \int \sqrt{\mathbb{E}\left\{\left(\widehat{s}(\boldsymbol{x}) - 1\right)^2\right\}}\mu(d\boldsymbol{x}) \\ &= \int \sqrt{\frac{n}{k^2}\mathrm{Var}\left(I(\boldsymbol{X} \in S_{\boldsymbol{x},\rho_n^*(\boldsymbol{x})})\right)}\mu(d\boldsymbol{x}) \leq \frac{1}{\sqrt{k}}.\end{aligned}$$

Thus we obtain,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left(\int |\widehat{s}(\boldsymbol{x}) - 1|\mu(d\boldsymbol{x})\right) &= 0, \\ \text{and} \quad \lim_{n\to\infty} \mathbb{E}\left(\int |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{m}^*_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right) &= 0.\end{aligned}$$

We already showed that

$$\lim_{n\to\infty} \mathbb{E}\left(\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right) = 0.$$

So we have,

$$\lim_{n\to\infty} \mathbb{E}\left(\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right) = 0.$$

Fix $\epsilon > 0$. Then we can find an n_0 such that, for $n \geq n_0$,

$$\begin{aligned}\mathbb{E}\left(\int |\widehat{s}(\boldsymbol{x}) - 1|\mu(d\boldsymbol{x})\right) &< \frac{\zeta}{8M}\epsilon, \\ \text{and} \quad \mathbb{E}\left(\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right) &< \frac{\epsilon}{8}.\end{aligned}$$

Then, by (A.2), we have, when $n \geq n_0$,

$$\begin{aligned}& P(\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) > \epsilon) \qquad \text{(A.3)} \\ \leq\ & P\left(\int |\widehat{s}(\boldsymbol{x}) - 1|\mu(d\boldsymbol{x}) - \mathbb{E}\int |\widehat{s}(\boldsymbol{x}) - 1|\mu(d\boldsymbol{x}) > \frac{\zeta}{4M}\epsilon\right) \\ & + P\left(\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) - \mathbb{E}\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) > \frac{1}{2}\epsilon\right).\end{aligned}$$

We will use McDiarmid's inequality (Devroye et al. 1996, Theorem 9.2) to bound each term on the right-hand side of (A.3). Fix an arbitrary realization of the data $(\boldsymbol{x}_j, a_j, r_j)_{j=1}^n$. Replace $(\boldsymbol{x}_i, a_i, r_i)$ by $(\boldsymbol{x}'_i, a'_i, r'_i)$, changing the value of $\widehat{m}^*_\ell(\boldsymbol{x})$ to $\widehat{m}^*_{\ell,i}(\boldsymbol{x})$. Thus

$$\left|\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) - \int |\widehat{m}^*_{\ell,i}(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right| \leq \int |\widehat{m}^*_\ell(\boldsymbol{x}) - \widehat{m}^*_{\ell,i}(\boldsymbol{x})|\mu(d\boldsymbol{x}).$$

And

$$|\widehat{m}^*_\ell(\boldsymbol{x})-\widehat{m}^*_{\ell,i}(\boldsymbol{x})| = \frac{1}{k}\Big| r_i \frac{\mathbb{I}(a_i=\ell)}{\pi_\ell(\boldsymbol{x}_i)}\mathbb{I}\left(||\boldsymbol{x}-\boldsymbol{x}_i|| \le \rho^*_n(\boldsymbol{x})\right) - r'_i \frac{\mathbb{I}(a'_i=\ell)}{\pi_\ell(\boldsymbol{x}'_i)}\mathbb{I}\left(||\boldsymbol{x}-\boldsymbol{x}'_i|| \le \rho^*_n(\boldsymbol{x})\right)\Big|$$

is bounded by $2M/(k\zeta)$, and can differ from zero only if $||\boldsymbol{x}-\boldsymbol{x}_i|| \le \rho^*_n(\boldsymbol{x})$ or $||\boldsymbol{x}-\boldsymbol{x}'_i|| \le \rho^*_n(\boldsymbol{x})$. Note that $||\boldsymbol{x}-\boldsymbol{x}_i|| \le \rho^*_n(\boldsymbol{x})$ if and only if $\mu(S_{\boldsymbol{x},||\boldsymbol{x}-\boldsymbol{x}_i||}) \le k/n$. By Lemma A.3.3, the measure of such $\boldsymbol{x}$ is bounded by $\gamma_p k/n$. Thus by McDiarmid's inequality,

$$P\left(\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) - \mathbb{E}\int |\widehat{m}^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) > \frac{1}{2}\epsilon\right) \le \exp\left(-\frac{n\epsilon^2\zeta^2}{32M^2\gamma_p^2}\right).$$

Similarly,

$$\Big|\int |\widehat{s}(\boldsymbol{x})-1|\mu(d\boldsymbol{x}) - \int |\widehat{s}_i(\boldsymbol{x})-1|\mu(d\boldsymbol{x})\Big| \le \int |\widehat{s}(\boldsymbol{x})-\widehat{s}_i(\boldsymbol{x})|\mu(d\boldsymbol{x}),$$

and

$$|\widehat{s}(\boldsymbol{x})-\widehat{s}_i(\boldsymbol{x})| = \frac{1}{k}\Big|\mathbb{I}\left(||\boldsymbol{x}-\boldsymbol{x}_i|| \le \rho^*_n(\boldsymbol{x})\right) - \mathbb{I}\left(||\boldsymbol{x}-\boldsymbol{x}'_i|| \le \rho^*_n(\boldsymbol{x})\right)\Big|$$

is bounded by $1/k$. By McDiarmid's inequality again,

$$P\left(\int |\widehat{s}(\boldsymbol{x})-1|\mu(d\boldsymbol{x}) - \mathbb{E}\int |\widehat{s}(\boldsymbol{x})-1|\mu(d\boldsymbol{x}) > \frac{\zeta}{4M}\epsilon\right) \le \exp\left(-\frac{n\epsilon^2\zeta^2}{32M^2\gamma_p^2}\right).$$

The desired result follows by (A.3) with $c = \dfrac{\zeta^2}{32M^2\gamma_p^2}$. □

Now we prove (ii), strong consistency for unbounded R. A counterpart of Lemma 5 in Devroye et al. (1994) is needed for the setting of optimal treatment regimes. The proof follows the idea in Györfi (1991).

Lemma A.1.2. *Consider the k-nearest neighbor estimate $\widehat{m}'_\ell(\boldsymbol{x})$ in (A.1). Then*

$$\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \to 0$$

almost surely for all distributions of $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1)∼(A3) if the following two conditions are satisfied:

(a) $\int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \to 0$ almost surely for all distributions of $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1) and (A3) with bounded R.

(b) There exists a constant $c > 0$ such that, for all distributions of $(\boldsymbol{X}, A, R)$ satisfying assumptions (A2) and (A3),

$$\limsup_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k} \int |R_{(i,n)}(\boldsymbol{x})|\mu(d\boldsymbol{x}) \leq c\mathbb{E}|R| \quad a.s. \tag{A.4}$$

Proof of Lemma A.1.2: For an arbitrary M, let

$$T_i = \begin{cases} R_i & \text{if } |R_i| \leq M, \\ M\text{sign}(R_i) & \text{otherwise,} \end{cases}$$

for $i = 1, \cdots, n$. T is defined similarly. Let $\widehat{t}_\ell(\boldsymbol{x})$ be the functions $\widehat{m}'_\ell(\boldsymbol{x})$, respectively, when the R_i is replaced by T_i, for $i = 1, \cdots, n$. Denote $t_\ell(\boldsymbol{x}) = \mathbb{E}(T|\boldsymbol{X} = \boldsymbol{x}, A = \ell)$ for $\ell = 1, \cdots, L$. Now,

$$\begin{aligned} & \limsup_{n\to\infty} \int |\widehat{m}'_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\ \leq\ & \limsup_{n\to\infty} \int |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{t}_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) + \limsup_{n\to\infty} \int |\widehat{t}_\ell(\boldsymbol{x}) - t_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\ & + \int |t_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}). \end{aligned}$$

For the first term on the right-hand side, we have,

$$\begin{aligned} & \limsup_{n\to\infty} \int |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{t}_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\ \leq\ & \limsup_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k} \int |R_{(i,n)}(\boldsymbol{x}) - T_{(i,n)}(\boldsymbol{x})| \frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x}) = \ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}\mu(d\boldsymbol{x}) \\ \leq\ & \limsup_{n\to\infty} \frac{1}{k\zeta}\sum_{i=1}^{k} \int |R_{(i,n)}(\boldsymbol{x}) - T_{(i,n)}(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\ \leq\ & \frac{c}{\zeta}\mathbb{E}|R - T| \quad a.s. \end{aligned}$$

The last inequality is due to condition (b) since $\mathbb{E}|R-T| < \infty$. The second term converges almost surely to zero by condition (a). By Jensen's inequality, the third term satisfies,

$$\int |t_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) = \mathbb{E}\Big|\mathbb{E}(R-T|X, A = \ell)\Big| \leq \mathbb{E}\left(|R-T|\Big|A = \ell\right) \leq \frac{1}{\zeta}\mathbb{E}|R-T|.$$

Thus we have,

$$\limsup_{n\to\infty} \int |\widehat{m}'_\ell(\boldsymbol{x}) - \widehat{t}_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}) \le \frac{c+1}{\zeta}\mathbb{E}|R-T| \quad a.s.$$

By the dominated convergence theorem, $\mathbb{E}|R-T| \to 0$ as $M \to \infty$. The desired result now follows as $M \to \infty$. □

For strong consistency in (ii), since we have already proved strong consistency for bounded R, it is enough to prove (A.4).

We need some geometric properties of nearest neighborhood. Define a cone $C(\boldsymbol{x}, \boldsymbol{s})$ to be the collection of all $\boldsymbol{x}' \in \mathbb{R}^p$ for which either $\boldsymbol{x}' = \boldsymbol{x}$ or angle$(\boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{s}) \le \pi/6$. Let S be a minimal subset of $\mathbb{R}^p$ such that a collection of cones $C(\boldsymbol{x}, \boldsymbol{s})$ for $\boldsymbol{s} \in S$ covers $\mathbb{R}^p$. By Lemma A.3.2, such an S exists, and its cardinality $|S|$ is γ_p. Let A_i be the collection of all $\boldsymbol{x} \in \mathbb{R}^p$ such that $\boldsymbol{X}_i$ is one of its k nearest neighbors. Define the sets $C_{i,\boldsymbol{s}} = C(\boldsymbol{X}_i, \boldsymbol{s})$ for $i = 1, \cdots, n$ and $\boldsymbol{s} \in S$. Let $B_{i,\boldsymbol{s}}$ be the subset of $C_{i,\boldsymbol{s}}$ consisting of all $\boldsymbol{x}$ that are among k nearest neighbors of $\boldsymbol{X}_i$ in the set $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \cdots, \boldsymbol{X}_n, \boldsymbol{x}\} \bigcap C_{i,\boldsymbol{s}}$. If the number of $\boldsymbol{X}_j$'s ($j \ne i$) contained in $C_{i,\boldsymbol{s}}$ is fewer than k, then $B_{i,\boldsymbol{s}} = C_{i,\boldsymbol{s}}$.

Observe that, by Lemma A.3.5 and Lemma A.3.6,

$$\limsup_{n\to\infty} \frac{n}{k} \max_i \mu(A_i) \le \limsup_{n\to\infty} \frac{n}{k} \max_i \sum_{\boldsymbol{s}\in S} \mu(B_{i,\boldsymbol{s}}) \le \sum_{\boldsymbol{s}\in S} \limsup_{n\to\infty} \frac{n}{k} \max_i \mu(B_{i,\boldsymbol{s}}) \le 2\gamma_p.$$

Then, we have,

$$\begin{aligned}
& \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} \int |R_{(i,n)}(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\
= \; & \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{n} |R_i|\mu(A_i) \\
\le \; & \limsup_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} |R_i|\right) \limsup_{n\to\infty} \left(\frac{n}{k} \max_i \mu(A_i)\right) \\
\le \; & 2\gamma_p \limsup_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} |R_i|\right) = 2\gamma_p \mathbb{E}|R| \quad a.s.
\end{aligned}$$

Thus strong consistency in (ii) follows from Lemma A.1.2 and Lemma A.1.1. The proof of Theorem 2.2.1 is complete.

REMARK: In practical use, we prefer Stone's estimate in (2.4) to break distance ties. Consider a simpler rule,

$$\widetilde{m}'_\ell(\boldsymbol{x}) = \frac{1}{k}\sum_{i\in A_n(\boldsymbol{x})} R_{(i,n)}(\boldsymbol{x})\frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)} + \frac{k-|A_n(\boldsymbol{x})|}{k|B_n(\boldsymbol{x})|}\sum_{i\in B_n(\boldsymbol{x})} R_{(i,n)}(\boldsymbol{x})\frac{\mathbb{I}(A_{(i,n)}(\boldsymbol{x})=\ell)}{\pi_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right)}.$$

When $k\to\infty$, $\widetilde{m}'_\ell$ is asymptotically equivalent to Stone's estimate $\widetilde{m}_\ell$ in (2.4). Assumption (A3) has a connotation of breaking distance ties randomly as demonstrated in the main paper. If the assumption does not hold, a small uniform variable $U \sim uniform[0,\epsilon]$ independent of $(\boldsymbol{X}, A, R)$ may be added to the vector $\boldsymbol{X}$. We may perform the k-nearest neighbor rule on $(\boldsymbol{X}, U)$. By Jensen's inequality,

$$\mathbb{E}\int_0^\epsilon\int |\widehat{m}'_\ell(\boldsymbol{x},u)-m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})du \geq \mathbb{E}\int\left|\mathbb{E}\left(\int_0^\epsilon \widehat{m}'_\ell(\boldsymbol{x},u)du\Big|D_n\right)-m_\ell(\boldsymbol{x})\right|\mu(d\boldsymbol{x}).$$

Fixing the data $D_n=\{(\boldsymbol{X}_i,A_i,R_i): i=1,\cdots,n\}$, we can always find a small enough ϵ such that

$$\mathbb{E}\left(\int_0^\epsilon \widehat{m}'_\ell(\boldsymbol{x},u)du\Big|D_n\right)=\widetilde{m}'_\ell(\boldsymbol{x}).$$

Thus $\widetilde{m}'_\ell$ is better than $\widehat{m}'_\ell$ on $(\boldsymbol{X},U)$, and then Stone's tie-breaking rule $\widetilde{m}_\ell$ in (2.4) is asymptotically better than random tie-breaking.

A.2 Proof of Theorem 2.2.2

By Lemma A.1.1,

$$\mathbb{E}\left\{\left(\mathcal{V}(d^*)-\mathcal{V}(d^{NN})\right)^2\right\} \leq L\sum_{\ell=1}^{L}\mathbb{E}\left\{\left(\int|\widehat{m}_\ell(\boldsymbol{x})-m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right)^2\right\}.$$

So it suffices to show the following theorem for the bound on $\mathbb{E}\left\{(\int|\widehat{m}_\ell(\boldsymbol{x})-m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x}))^2\right\}$, for $\ell=1,\cdots,L$.

Theorem A.2.1. *For any distribution P for $(\boldsymbol{X}, A, R)$ satisfying assumptions (A1′)∼(A4′), and $\ell = 1, \cdots, L$,*

(i) If $p = 1$,

$$\mathbb{E}\left\{\left(\int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right)^2\right\} \leq c^2\sigma^2\frac{1}{k} + 16c\rho^2C^2\frac{k}{n}. \tag{A.5}$$

(ii) If $p = 2$,

$$\mathbb{E}\left\{\left(\int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right)^2\right\} \leq c^2\sigma^2\frac{1}{k} + 8c\rho^2C^2\frac{k}{n}\left(1 + \log\left(\frac{n}{k}\right)\right). \tag{A.6}$$

(iii) If $p \geq 3$,

$$\mathbb{E}\left\{\left(\int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right)^2\right\} \leq c^2\sigma^2\frac{1}{k} + \frac{8c\rho^2C^2}{1 - 2/p}\left\lfloor\frac{n}{k}\right\rfloor^{-\frac{2}{d}}. \tag{A.7}$$

Proof of Theorem A.2.1: Let

$$m^*_\ell(\boldsymbol{x}) = \mathbb{E}\left(\widehat{m}_\ell(\boldsymbol{x})|\boldsymbol{X}_1, A_1, \cdots, \boldsymbol{X}_n, A_n\right) = \sum_{i=1}^k W^\ell_{n,i}(\boldsymbol{x})m_\ell\left(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\right).$$

The last equality is due to the fact that $W^\ell_{n,i}(\boldsymbol{x}) = 0$ if $A_{(i,n)}(\boldsymbol{x}) \neq \ell$. We have the decomposition

$$\begin{aligned}
&\mathbb{E}\left\{\left(\int |\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})|\mu(d\boldsymbol{x})\right)^2\right\} \leq \mathbb{E}\int\left(\widehat{m}_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})\right)^2\mu(d\boldsymbol{x}) \\
= \;& \mathbb{E}\int\left(\widehat{m}_\ell(\boldsymbol{x}) - m^*_\ell(\boldsymbol{x})\right)^2\mu(d\boldsymbol{x}) + \mathbb{E}\int\left(m^*_\ell(\boldsymbol{x}) - m_\ell(\boldsymbol{x})\right)^2\mu(d\boldsymbol{x}).
\end{aligned}$$

For the first term on the right-hand side,

$$\begin{aligned}
&\mathbb{E}\int\left(\widehat{m}_\ell(\boldsymbol{x}) - m^*_\ell(\boldsymbol{x})\right)^2\mu(d\boldsymbol{x}) \\
= \;& \mathbb{E}\int\left(\sum_{i=1}^k W^\ell_{n,i}(\boldsymbol{x})\Big(R_{(i,n)}(\boldsymbol{x}) - m_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)\Big)\right)^2\mu(d\boldsymbol{x}) \\
= \;& \mathbb{E}\int\sum_{i=1}^k\left(W^\ell_{n,i}(\boldsymbol{x})\right)^2\Big(R_{(i,n)}(\boldsymbol{x}) - m_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)\Big)^2\mu(d\boldsymbol{x}) \\
= \;& \mathbb{E}\int\sum_{i=1}^k\left(W^\ell_{n,i}(\boldsymbol{x})\right)^2\sigma^2_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big)\mu(d\boldsymbol{x}) \\
\leq \;& c^2\sigma^2\frac{1}{k}.
\end{aligned}$$

For the second term,

$$\begin{aligned}
& \mathbb{E}\int \big(m_\ell^*(\boldsymbol{x}) - m_\ell(\boldsymbol{x})\big)^2 \mu(d\boldsymbol{x}) \\
= \;& \mathbb{E}\int \left(\sum_{i=1}^{k} W_{n,i}^{\ell}(\boldsymbol{x})\Big(m_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big) - m_\ell(\boldsymbol{x})\Big)\right)^2 \mu(d\boldsymbol{x}) \\
\leq \;& \mathbb{E}\int \sum_{i=1}^{k}\Big(W_{n,i}^{\ell}(\boldsymbol{x})\Big)^2 \sum_{i=1}^{k}\Big(m_\ell\big(\boldsymbol{X}_{(i,n)}(\boldsymbol{x})\big) - m_\ell(\boldsymbol{x})\Big)^2 \mu(d\boldsymbol{x}) \\
\leq \;& cC^2\mathbb{E}||\boldsymbol{X}_{(k,n)}(\boldsymbol{X}) - \boldsymbol{X}||^2
\end{aligned}$$

The desired results in Theorem A.2.1 now follow directly from Lemma A.3.7. □

When $p = 1$, take $k \propto n^{1/2}$, and the right-hand side of (A.5) is $O(n^{-1/2})$. When $p = 2$, take $k \propto n^{1/2-\epsilon}$ for any $\epsilon > 0$, and the right-hand side of (A.6) is $O(n^{-1/2+\epsilon})$. When ϵ is very small, its rate of convergence will arbitrarily be close to $1/2$. When $p \geq 3$, take $k \propto n^{2/(p+2)}$, and the right-hand side of (A.7) is $O(n^{-2/(p+2)})$. Theorem 2.2.2 is now proved.

A.3 Background on k-NN regression

The setup in this section is for regression analysis, and is different with the setup in the main paper. In regression analysis one considers a random vector $(\boldsymbol{X}, Y)$, where $\boldsymbol{X}$ is $\mathbb{R}^p$-valued, and Y is $\mathbb{R}$-valued. Let D_n be the set of observed data defined by

$$D_n = \{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\},$$

where $(\boldsymbol{X}_1, Y_1)$, $\cdots$, $(\boldsymbol{X}_n, Y_n)$ and $(\boldsymbol{X}, Y)$ are independent and identically distributed (i.i.d.) random variables. Let $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. In the regression problem one wants to use the data D_n in order to construct an estimate $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ of the regression function m. Here $\widehat{m}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{x}, D_n)$ is a measurable function of $\boldsymbol{x}$ and the data. We first state Stone's Theorem (Stone 1977). The theorem was applied to prove consistency

of kernel and nearest neighbor estimates in the literature (Devroye et al. 1996, Györfi et al. 2002). The theorem considers a regression function estimate taking the form

$$\widehat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} W_{n,i}(\boldsymbol{x}) Y_i,$$

where the weights $W_{n,i}(\boldsymbol{x}) = W_{n,i}(\boldsymbol{x}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n) \in \mathbb{R}$ depend on $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$.

Theorem A.3.1 (Stone's Theorem)**.** *Assume that the following conditions are satisfied for any distribution of $\boldsymbol{X}$:*

(i) There is a constant c such that for every nonnegative measurable function f satisfying $\mathbb{E}f(\boldsymbol{X}) < \infty$ and any n,

$$\mathbb{E}\left\{\sum_{i=1}^{n} |W_{n,i}(\boldsymbol{X})| f(\boldsymbol{X}_i)\right\} \leq c\mathbb{E}f(\boldsymbol{X}).$$

(ii) There is a $D \geq 1$ such that

$$\mathbb{P}\left\{\sum_{i=1}^{n} |W_{n,i}(\boldsymbol{X})| \leq D\right\} = 1,$$

for all n.

(iii) For all $a > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{\sum_{i=1}^{n} |W_{n,i}(\boldsymbol{X})| \mathbb{I}(||\boldsymbol{X}_i - \boldsymbol{X}|| > a)\right\} = 0.$$

(iv)

$$\sum_{i=1}^{n} W_{n,i}(\boldsymbol{X}) \to 1 \quad \textit{in probability.}$$

(v)

$$\max_i |W_{n,i}(\boldsymbol{X})| \to 0 \quad \textit{in probability.}$$

Then the corresponding regression function estimate $\widehat{m}$ converges in mean to m, i.e.,

$$\mathbb{E}\left(\int |\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| \mu(d\boldsymbol{x})\right) \to 0$$

for all distributions of $(\boldsymbol{X}, Y)$ with $\mathbb{E}|Y| < \infty$.

We fix $x \in \mathbb{R}^p$, and reorder the observed data $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$ according to increasing values of $||\boldsymbol{X}_i - \boldsymbol{x}||$. The reordered data sequence is denoted by

$$(\boldsymbol{X}_{(1,n)}(\boldsymbol{x}), Y_{(1,n)}(\boldsymbol{x})), \cdots, (\boldsymbol{X}_{(n,n)}(\boldsymbol{x}), Y_{(n,n)}(\boldsymbol{x})).$$

Thus $\boldsymbol{X}_{(k,n)}(\boldsymbol{x})$ is the k-th nearest neighbor of $\boldsymbol{x}$.

We introduce some results on the nearest neighborhood of $\boldsymbol{x}$ which are useful in proving theorems in the main paper. Denote the probability measure for $\boldsymbol{X}$ by μ. In this section, we assume that distance ties occur with probability zero in μ. Let $S_{\boldsymbol{x},\epsilon}$ be the closed ball centered at $\boldsymbol{x}$ of radius $\epsilon > 0$. Define $support(\mu) = \{\boldsymbol{x} : \text{for all } \epsilon > 0, \mu(S_{\boldsymbol{x},\epsilon}) > 0\}$.

Lemma A.3.1 (Lemma 5.1 in Devroye et al. (1996)). *If $\boldsymbol{x} \in support(\mu)$ and $\lim_{n\to\infty} k/n = 0$, then $||\boldsymbol{X}_{(k,n)}(\boldsymbol{x}) - \boldsymbol{x}|| \to 0$ with probability one.*

Let us define the cone $C(\boldsymbol{x}, \boldsymbol{s})$ to be the collection of all $\boldsymbol{x}' \in \mathbb{R}^p$ for which either $\boldsymbol{x}' = \boldsymbol{x}$ or $\text{angle}(\boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{s}) \leq \pi/6$. The following lemma shows a finite set of such cones may cover $\mathbb{R}^p$.

Lemma A.3.2 (Lemma 5.5 in Devroye et al. (1996)). *There exists a finite set $S \subset \mathbb{R}^p$ such that*

$$\mathbb{R}^p = \bigcup_{\boldsymbol{s}\in S} C(\boldsymbol{x}, \boldsymbol{s}),$$

regardless of how $\boldsymbol{x} \in \mathbb{R}^p$ is picked. Furthermore, define γ_p as the minimal number of elements in S. Then γ_p depends only on the dimension p, and

$$\gamma_p \leq \left(1 + 2\sqrt{2 - \sqrt{3}}\right)^p - 1.$$

The next several lemma will enable us to establish weak and strong consistency of nearest neighbor rules.

Lemma A.3.3 (Lemma 11.1 in Devroye et al. (1996)). *Let $B_a(\boldsymbol{x}') = \{\boldsymbol{x} : \mu(S_{\boldsymbol{x},||\boldsymbol{x}-\boldsymbol{x}'||}) \leq a\}$. Then for all $\boldsymbol{x}' \in \mathbb{R}^p$,*

$$\mu(B_a(\boldsymbol{x}')) \leq \gamma_p a.$$

Lemma A.3.4 (Lemma 5.3 in Devroye et al. (1996)). *For any integrable function f, any n, and any $k \leq n$,*

$$\sum_{i=1}^{k} \mathbb{E}\left(|f(\boldsymbol{X}_{(i,n)}(\boldsymbol{X}))|\right) \leq k\gamma_p \mathbb{E}(|f(\boldsymbol{X})|).$$

Let A_i be the collection of all $\boldsymbol{x} \in \mathbb{R}^p$ such that $\boldsymbol{X}_i$ is one of its k nearest neighbors. Let S be a minimal subset of $\mathbb{R}^p$, such that a collection of cones $C(\boldsymbol{x}, \boldsymbol{s})$ for $\boldsymbol{s} \in S$ covers $\mathbb{R}^p$. Thus $\gamma_p = |S|$, the cardinality of this set. Define the sets $C_{i,\boldsymbol{s}} = C(\boldsymbol{X}_i, \boldsymbol{s})$. Let $B_{i,\boldsymbol{s}}$ be the subset of $C_{i,\boldsymbol{s}}$ consisting of all $\boldsymbol{x}$ that are among the k nearest neighbors of $\boldsymbol{X}_i$ in the set $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \cdots, \boldsymbol{X}_n, \boldsymbol{x}\} \bigcap C_{i,\boldsymbol{s}}$.

Lemma A.3.5 (Lemma 6 in Devroye et al. (1994)). *If $\boldsymbol{x} \in A_i$, then $\boldsymbol{x} \in \bigcup_{\boldsymbol{s}\in S} B_{i,\boldsymbol{s}}$, and thus*

$$\mu(A_i) \leq \sum_{\boldsymbol{s}\in S} \mu(B_{i,\boldsymbol{s}}).$$

Lemma A.3.6 (Lemma 8 in Devroye et al. (1994)). *If $k/\log(n) \to \infty$ and $k/n \to 0$, then*

$$\limsup_{n\to\infty} \frac{n}{k} \max_i \mu(B_{i,\boldsymbol{s}}) \leq 2 \qquad a.s.$$

Devroye et al. (1994) applied an additional independent random variable to break distance ties. It is easy to translate the proofs of the previous two lemmas for the case where distance ties occur with probability zero in μ, so we can skip the proofs. The next lemma is helpful to show the rate of convergence in the main paper.

Lemma A.3.7 (Corollary 6 in Biau et al. (2010)). *Suppose that μ has a compact support with diameter 2ρ. Then*

(i) If $p = 1$,

$$\mathbb{E}||\boldsymbol{X}_{(i,n)} - \boldsymbol{X}||^2 \leq \frac{16\rho^2 i}{n}.$$

(ii) If $p = 2$,

$$\mathbb{E}||\boldsymbol{X}_{(i,n)} - \boldsymbol{X}||^2 \leq \frac{8\rho^2 i}{n}\left(1 + \log\left(\frac{n}{i}\right)\right).$$

(iii) If $p \geq 3$*,*

$$\mathbb{E}||\boldsymbol{X}_{(i,n)} - \boldsymbol{X}||^2 \leq \frac{8\rho^2 \lfloor n/i \rfloor^{-\frac{2}{p}}}{1 - 2/p}.$$

APPENDIX B: PROOFS FOR CHAPTER 3

Proof of Lemma 3.2.1

Proof. Note that,

$$\begin{aligned} & \mathbb{E}\left(\frac{\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}\right) \\ = \;& \mathbb{E}\left(\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big)|\boldsymbol{X}, A=1\right) + \mathbb{E}\left(\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big)|\boldsymbol{X}, A=-1\right) = 1. \quad \text{(B.8)} \end{aligned}$$

The desired result follows easily. □

Proof of Theorem 3.2.2

Proof. For any measurable function g,

$$\begin{aligned} & Var\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) \\ = \;& Var\left(\frac{R-\widetilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) + Var\left(\frac{\widetilde{g}(\boldsymbol{X})-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) \\ & +2Cov\left(\frac{R-\widetilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big), \frac{\widetilde{g}(\boldsymbol{X})-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right). \end{aligned}$$

It suffices to show that the covariance term is zero. Applying (B.8), we have

$$\begin{aligned} & \mathbb{E}\left(\frac{R-\widetilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) \\ = \;& \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) - \widetilde{g}(\boldsymbol{X})\mathbb{E}\left(\frac{\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}\right) \\ = \;& 0. \end{aligned}$$

Note that

$$\begin{aligned} \widetilde{g}(\boldsymbol{X}) &= \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) \\ &= \mathbb{E}(R|\boldsymbol{X}, A=1)\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big) + \mathbb{E}(R|\boldsymbol{X}, A=-1)\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big). \quad \text{(B.9)} \end{aligned}$$

Thus we have,

$$\begin{aligned}
& \mathbb{E}\left(\frac{R-\widetilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})^2}\mathbb{I}\big(A\neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) \\
= \;& \mathbb{E}\left(R-\widetilde{g}(\boldsymbol{X})|\boldsymbol{X},A=1\right)\mathbb{I}\big(d(\boldsymbol{X})\neq 1\big)/\pi(1,\boldsymbol{X}) \\
& +\mathbb{E}\left(R-\widetilde{g}(\boldsymbol{X})|\boldsymbol{X},A=-1\right)\mathbb{I}\big(d(\boldsymbol{X})\neq -1\big)/\pi(-1,\boldsymbol{X}) \\
= \;& 0.
\end{aligned}$$

The desired result follows easily. □

Proof of Theorem 3.3.1

Proof. Given $\boldsymbol{X}=\boldsymbol{x}$, for any measurable function f, similar reasoning to that used in the proof of Lemma 3.2.1 yields,

$$\mathbb{E}\left(\frac{T\big(Af(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}=\boldsymbol{x}\right)=2.$$

Then the conditional T-risk is

$$\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af(\boldsymbol{X}))\big|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \;& \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)T(f(\boldsymbol{x}))+\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)T(-f(\boldsymbol{x}))-2g(\boldsymbol{x}) \\
= \;& \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)T(f(\boldsymbol{x})) \\
& +2\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)-2g(\boldsymbol{x}). \qquad \text{(B.10)}
\end{aligned}$$

If $\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=1\right]-\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=-1\right]>0$, any function $f(\boldsymbol{x})\geq 1$ minimizes the conditional T-risk; similarly, if $\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=1\right]-\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=-1\right]<0$, any function $f(\boldsymbol{x})\leq -1$ minimizes the conditional T-risk. For either case, $sign(f^*_{T,g})=d^*$.

For the second part, by applying (B.10),

$$\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Ad^*(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right)-\mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \;& \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\left(T(d^*(\boldsymbol{x}))-T(f^*_{T,g}(\boldsymbol{x}))\right)=0.
\end{aligned}$$

The desired result follows by taking expectations on both sides. □

Proof of Theorem 3.3.2

Proof. Given $\boldsymbol{X} = \boldsymbol{x}$. By applying (B.10), for any measurable function f, we have

$$\begin{aligned} & \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Af(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) \\ = \; & \left(\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)\right)\left(T(f(\boldsymbol{x})) - T(f^*_{T,g}(\boldsymbol{x}))\right). \end{aligned}$$

Similarly,

$$\begin{aligned} & \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})} \mathbb{I}(A \neq sign(f(\boldsymbol{X})))|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})} \mathbb{I}(A \neq d^*(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) \\ = \; & \left(\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)\right)\left(\mathbb{I}(sign(f(\boldsymbol{x})) \neq 1) - \mathbb{I}(d^*(\boldsymbol{x}) \neq 1)\right). \end{aligned}$$

From the proof of Theorem 3.3.1, when $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) > \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$, $f^*_{T,g}(\boldsymbol{x}) \geq 1$ and $d^*(\boldsymbol{x}) = 1$, so $T(f^*_{T,g}(\boldsymbol{x})) = 0$ and $\mathbb{I}(d^*(\boldsymbol{x}) \neq 1) = 0$; when $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) < \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$, $f^*_{T,g}(\boldsymbol{x}) \leq -1$ and $d^*(\boldsymbol{x}) = -1$, so $T(f^*_{T,g}(\boldsymbol{x})) = 2$ and $\mathbb{I}(d^*(\boldsymbol{x}) \neq 1) = 1$. Note that, for any measurable function f, $1 \geq T(f(\boldsymbol{x})) - \mathbb{I}(sign(f(\boldsymbol{x})) \neq 1) \geq 0$. Thus it is easy to check that when $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) > \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$,

$$T(f(\boldsymbol{x})) - T(f^*_{T,g}(\boldsymbol{x})) \geq \mathbb{I}(sign(f(\boldsymbol{x})) \neq 1) - \mathbb{I}(d^*(\boldsymbol{x}) \neq 1),$$

and when $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) < \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$,

$$T(f(\boldsymbol{x})) - T(f^*_{T,g}(\boldsymbol{x})) \leq \mathbb{I}(sign(f(\boldsymbol{x})) \neq 1) - \mathbb{I}(d^*(\boldsymbol{x}) \neq 1).$$

So, for either case, we have

$$\begin{aligned} & \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})} \mathbb{I}(A \neq sign(f(\boldsymbol{X})))|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R}{\pi(A, \boldsymbol{X})} \mathbb{I}(A \neq d^*(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) \\ \leq \; & \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Af(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right). \end{aligned}$$

The desired result follows by taking expectations on both sides. □

Proof of Theorem 3.3.3

Proof. Let $L(h,b) = (R - g(\boldsymbol{X}))T(A(h(\boldsymbol{X}) + b))/\pi(A, \boldsymbol{X})$. For simplicity, we denote f_{D_n,λ_n}, h_{D_n,λ_n} and b_{D_n,λ_n} by f_n, h_n and b_n, respectively. By the definition of h_{D_n,λ_n} and b_{D_n,λ_n}, we have, for any $h \in \mathcal{H}_K$ and $b \in \mathbb{R}$,

$$\mathbb{P}_n(L(h_n, b_n)) \leq \mathbb{P}_n(L(h_n, b_n)) + \frac{\lambda_n}{2}||h_n||_K^2 \leq \mathbb{P}_n(L(h,b)) + \frac{\lambda_n}{2}||h||_K^2,$$

where $\mathbb{P}_n$ denotes the empirical measure of the observed data. Then, $\limsup_n \mathbb{P}_n(L(h_n, b_n)) \leq \mathbb{P}(L(h,b)) = \mathcal{R}_{T,g}(h+b)$ with probability 1. This implies

$$\limsup_n \mathbb{P}_n(L(h_n, b_n)) \leq \inf_{h \in \mathcal{H}_K, b \in \mathbb{R}} \mathcal{R}_{T,g}(h+b) \leq \mathbb{P}(L(h_n, b_n))$$

with probability 1. It suffices to show $\mathbb{P}_n(L(h_n, b_n)) - \mathbb{P}(L(h_n, b_n)) \to 0$ in probability.

We first obtain a bound for $||h_n||_K$. Since $\mathbb{P}_n(L(h_n, b_n)) + \lambda_n||h_n||_K^2/2 \leq \mathbb{P}_n(L(h,b)) + \lambda_n||h||_K^2/2$, for any $h \in \mathcal{H}_K$ and $b \in \mathbb{R}$, we can choose $h = 0$ and $b = 0$ to obtain, $\mathbb{P}_n(L(h_n, b_n)) + \lambda_n||h_n||_K^2/2 \leq \mathbb{P}_n((R - g(\boldsymbol{X}))/\pi(A, \boldsymbol{X}))$. Note that $0 \leq T(u) \leq 2$. We thus have,

$$\lambda_n||h_n||_K^2 \leq 2\mathbb{P}_n(|R - g(\boldsymbol{X})|/\pi(A, \boldsymbol{X})) \leq 2M_0.$$

Let $M_1 = \sqrt{2M_0}$. Then the $\mathcal{H}_K$ norm of $\sqrt{\lambda_n}h_n$ is bounded by M_1.

Next we obtain a bound for b_n. We claim that there is a global solution (h_n, b_n) such that $h_n(\boldsymbol{x}_i) + b_n \in [-1, 1]$ for some i. Suppose there is a global solution (h'_n, b'_n) such that $|h'_n(\boldsymbol{x}_i) + b'_n| > 1$ for all i. Let $\delta = |h'_n(\boldsymbol{x}_{i_0}) + b'_n| = \min_{1 \leq i \leq n} |h'_n(\boldsymbol{x}_i) + b'_n| > 1$. Then let $h_n = h'_n$ and $b_n = b'_n - (\delta - 1)sign(h'_n(\boldsymbol{x}_{i_0}) + b'_n)$. It is easy to check that $h_n(\boldsymbol{x}_{i_0}) + b_n = 1$ if $h'_n(\boldsymbol{x}_{i_0}) + b'_n > 1$, and $h_n(\boldsymbol{x}_{i_0}) + b_n = -1$ if $h'_n(\boldsymbol{x}_{i_0}) + b'_n < -1$; furthermore when $i \neq i_0$, $h_n(\boldsymbol{x}_i) + b_n \geq 1$ if $h'_n(\boldsymbol{x}_i) + b'_n > 1$, and $h_n(\boldsymbol{x}_i) + b_n \leq -1$ if $h'_n(\boldsymbol{x}_i) + b'_n < -1$. So $T(h_n(\boldsymbol{x}_i) + b_n) = T(h'_n(\boldsymbol{x}_i) + b'_n)$ for all i. Hence (h_n, b_n) is a global solution and satisfies our claim. Now if a solution (h_n, b_n) satisfies our claim, we then have,

$$|b_n| \leq 1 + |h_n(\boldsymbol{x}_{i_0})| \leq 1 + ||h_n||_\infty.$$

Note that $||h||_\infty \leq C_K||h||_K$. We have,

$$|\sqrt{\lambda_n}b_n| \leq \sqrt{\lambda_n} + C_K\sqrt{\lambda_n}||h_n||_K.$$

Since $\lambda_n \to 0$, and C_K and $\sqrt{\lambda_n}||h_n||_K$ are both bounded, we have $|\sqrt{\lambda_n}b_n|$ is bounded too. Let the bound be M_2, *i.e.* $|\sqrt{\lambda_n}b_n| \leq M_2$.

Note that the class $\{\sqrt{\lambda_n}h : ||\sqrt{\lambda_n}h||_K \leq M_1\}$ is a Donsker class. So $\{\sqrt{\lambda_n}(h+b) : ||\sqrt{\lambda_n}h||_K \leq M_1, |\sqrt{\lambda_n}b| \leq M_2\}$ is also P-Donsker. Consider the function

$$T_\lambda(u) = \begin{cases} 2\sqrt{\lambda} & \text{if} \quad u < -\sqrt{\lambda}, \\ 2\sqrt{\lambda} - \frac{1}{\sqrt{\lambda}}(\sqrt{\lambda}+u)^2 & \text{if} \quad -\sqrt{\lambda} \leq u < 0, \\ \frac{1}{\sqrt{\lambda}}(\sqrt{\lambda}-u)^2 & \text{if} \quad 0 \leq u < \sqrt{\lambda}, \\ 0 & \text{if} \quad u \geq \sqrt{\lambda}. \end{cases}$$

We have $T_\lambda(\sqrt{\lambda}u) = \sqrt{\lambda}T(u)$. Since $T_\lambda(u)$ is a Lipschitz continuous function with Lipschitz constant equal to 2, and $\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$ is bounded, the class $\{\sqrt{\lambda_n}L(h,b) : ||\sqrt{\lambda_n}h||_K \leq M_1, |\sqrt{\lambda_n}b| \leq M_2\}$ is also P-Donsker. Therefore,

$$\sqrt{n\lambda_n}(\mathbb{P}_n - \mathbb{P})L(h_n, b_n) = O_p(1).$$

Consequently, from $n\lambda_n \to \infty$, $\mathbb{P}_n(L(h_n,b_n)) - \mathbb{P}(L(h_n,b_n)) \to 0$ in probability. □

Proof of Lemma 3.3.4

Proof. Fix any $0 < \epsilon < 1$. $d^*(\boldsymbol{x}) = sign(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1))$ is measurable. Since μ is regular, using Lusin's theorem in measure theory, we know that $d^*(\boldsymbol{x})$ can be approximated by a continuous function $f'(\boldsymbol{x}) \in C(\mathcal{X})$ such that $\mu(f'(\boldsymbol{x}) \neq d^*(\boldsymbol{x})) \leq \frac{\epsilon}{4M}$. Thus

$$
\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af'(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right)-\mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Ad^*(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \; & \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\left(T(f'(\boldsymbol{x}))-T(d^*(\boldsymbol{x}))\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
& \mathcal{R}_{T,g}(f')-\mathcal{R}^*_{T,g}=|\mathcal{R}_{T,g}(f')-\mathcal{R}_{T,g}(d^*)| \\
= \; & \left|\int\left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\left(T(f'(\boldsymbol{x}))-T(d^*(\boldsymbol{x}))\right)\mu(d\boldsymbol{x})\right| \\
\leq \; & \int\left|\left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\right|\left|T(f'(\boldsymbol{x}))-T(d^*(\boldsymbol{x}))\right| \\
& \qquad \mathbb{I}(f'(\boldsymbol{x})\neq d^*(\boldsymbol{x}))\mu(d\boldsymbol{x}).
\end{aligned}
$$

Since $|R|\leq M$ and $0\leq T(u)\leq 2$,

$$
\mathcal{R}_{T,g}(f')-\mathcal{R}^*_{T,g}<\epsilon
$$

Since K is universal, there exist a function $f''\in\mathcal{H}_K$ such that $||f''-f'||_\infty<\frac{\epsilon}{4M}$. Note that $T(\cdot)$ is Lipschitz continuous with Lipschitz constant 2. Similarly,

$$
\begin{aligned}
& |\mathcal{R}_{T,g}(f'')-\mathcal{R}_{T,g}(f')| \\
= \; & \left|\int\left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\left(T(f''(\boldsymbol{x}))-T(f'(\boldsymbol{x}))\right)\mu(d\boldsymbol{x})\right| \\
\leq \; & 2\int\left|\left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\right|\left|f'(\boldsymbol{x})-f'(\boldsymbol{x})\right|\mu(d\boldsymbol{x})<\epsilon.
\end{aligned}
$$

By combining the two inequalities, we have

$$
\mathcal{R}_{T,g}(f'')-\mathcal{R}^*_{T,g}<2\epsilon.
$$

Noting that $f''\in\mathcal{H}_K$ and letting $\epsilon\to 0$, we obtain the desired result. □

Proof of Theorem 3.3.6

Proof. The proof follows the idea in Devroye et al. (1996, Theorem 7.2). Since the proof is very similar, we only provide a sketch to save space.

Let $b = 0.b_1b_2b_3\cdots$ be a real number on $[0,1]$ with the given binary expansion, and let B be a random variable uniformly distributed on $[0,1]$ with expansion $B = 0.B_1B_2B_3\cdots$. Let us restrict ourselves to a random variable $\boldsymbol{X}$ with the support $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ where $\boldsymbol{x}_i \in \mathcal{X}$. For simplicity, we recode the support of $\boldsymbol{X}$ as $\{1, 2, \cdots\}$. Let

$$P(\boldsymbol{X} = i) = p_i, \quad i \geq 1, \tag{B.11}$$

where $p_1 \geq p_2 \geq \cdots > 0$, and $\sum_{i=n+1}^{\infty} p_i \geq \max(8c_n, 32np_{n+1})$ for every n. Such p_i's exist by Devroye et al. (1996, Lemma 7.1). Let $A \in \{1, -1\}$ be a binomial variable with $\pi(A, \boldsymbol{X}) = 0.5$. For a given b, set $R = AM$ if $b_{\boldsymbol{X}} = 1$, and $R = -AM$ if $b_{\boldsymbol{X}} = 0$. Then the Bayes rule is $d^*(\boldsymbol{X}) = (2 * b_{\boldsymbol{X}} - 1)$. Thus each $b \in [0,1]$ describes a different distribution of $(\boldsymbol{X}, A, R)$. Introduce the shortened notation $D_n = \{(\boldsymbol{X}_1, A_1, R_1), \cdots, (\boldsymbol{X}_n, A_n, R_n)\}$. Let d_n be a rule generated by data D_n. Define $d_n^i = d_n(i)$ for $i = 1, \cdots, n$. Let $\Delta\mathcal{R}_n(b)$ be the excess risk of the rule d_n for the distribution parametrized by b, and $\Delta\mathcal{R}_n(B)$ be the excess risk of the rule d_n for the random distribution.

$$\begin{aligned}
& \Delta\mathcal{R}_n(B) \\
= \;& \mathbb{E}\left[\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}(A \neq d_n(\boldsymbol{X}))\Big|B\right] - \mathbb{E}\left[\frac{R}{\pi(A, \boldsymbol{X})}\mathbb{I}(A \neq d^*(\boldsymbol{X}))\Big|B\right] \\
= \;& \mathbb{E}\Big[\big(\mathbb{E}(R|B, \boldsymbol{X}, A = 1) - \mathbb{E}(R|B, \boldsymbol{X}, A = -1)\big)\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 1) - \mathbb{I}(d^*(\boldsymbol{X}) \neq 1)\big)\Big|B\Big] \\
= \;& 2M\mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq d^*(\boldsymbol{X}))\big|B\big) \\
= \;& 2M\mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 2B_{\boldsymbol{X}} - 1)\big).
\end{aligned}$$

Let $L_n(B) = \mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 2B_{\boldsymbol{X}} - 1)\big)$. Then we have,

$$L_n(B) = \sum_{i=1}^{\infty} p_i \mathbb{I}(d_n^i \neq 2B_i - 1).$$

Following the same arguments used in Devroye et al. (1996, Theorem 7.2), we have

$$P(L_n(B) < 2c_n|D_n) \leq P(\sum_{i=n+1}^{\infty} p_i B_i < 2c_n) \leq e^{-2n}.$$

Hence we have

$$\begin{aligned}\sup_b \inf_n \mathbb{E}\left(\frac{L_n(b)}{2c_n}\right) &\geq \mathbb{E}\left(\mathbb{E}\left(\inf_n\left(\frac{L_n(b)}{2c_n}\Big|\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots\right)\right)\right)\\ &\geq \mathbb{E}\left(1-\sum_{i=1}^{\infty}\mathbb{E}(P(L_n(B)<2c_n|D_n)|\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots)\right)\\ &\geq 1-\sum_{i=1}^{\infty}e^{-2n}=\frac{e^2-2}{e^2-1}>\frac{1}{2}.\end{aligned}$$

Here we are omitting many steps. Refer to Devroye et al. (1996, Theorem 7.2) for details. The conclusion is that there exists a b for which $\Delta\mathcal{R}_n(b)\geq 2Mc_n$, $n=1,2,\cdots$. □

Proof of Theorem 3.3.7

Proof. Define the random variable $S=\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$. We consider a probability measure on the triplet $(\boldsymbol{X},A,S)$ instead of on $(\boldsymbol{X},A,R)$. Let $D_n=\{\boldsymbol{X}_i,A_i,S_i\}_{i=1}^n$ be independent random variables with the same distribution as $(\boldsymbol{X},A,S)$. Let $\mathbb{P}_n$ be the empirical measure on D_n. For simplicity, we denote f_{D_n,λ_n}, h_{D_n,λ_n} and b_{D_n,λ_n} by f_n, h_n and b_n, respectively. Let $(\widetilde{h}_{\lambda_n},\widetilde{b}_{\lambda_n})$ be a solution of the following optimization problem:

$$\min_{h\in\mathcal{H}_K,b\in\mathbb{R}}\frac{\lambda_n}{2}||h||_K^2+\mathcal{R}_{T,g}(h+b).$$

Let $L(h,b)=ST(A(h(\boldsymbol{X})+b))$. Then

$$\begin{aligned}&\mathcal{R}_{T,g}(f_n)-\mathcal{R}_{T,g}(f_{T,g}^*)\\ \leq\ &(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n)+(\frac{\lambda_n}{2}||h_n||^2+\mathbb{P}_nL(h_n,b_n))-(\frac{\lambda_n}{2}||\widetilde{h}_{\lambda_n}||^2+\mathbb{P}_nL(\widetilde{h}_{\lambda_n},\widetilde{b}_{\lambda_n}))\\ &+(\mathbb{P}_n-\mathbb{P})L(\widetilde{h}_{\lambda_n},\widetilde{b}_{\lambda_n})+\mathcal{A}(\lambda_n)\\ \leq\ &(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n)+(\mathbb{P}_n-\mathbb{P})L(\widetilde{h}_{\lambda_n},\widetilde{b}_{\lambda_n}))+\mathcal{A}(\lambda_n).\end{aligned}$$

We first estimate the second term by the Hoeffding inequality (Steinwart and Christmann 2008, Theorem 6.10). Since $|L(h,b)|\leq 2M_0$, we thus have, with probability at least $1-\delta/2$,

$$(\mathbb{P}_n-\mathbb{P})L(\widetilde{h}_{\lambda_n},\widetilde{b}_{\lambda_n})\leq M_0\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{B.12}$$

By the arguments used in the proof of Theorem 3.3.3, we have $||h_n||_K \leq \sqrt{\frac{2M_0}{\lambda_n}}$ and $|b_n| \leq 1 + C_K\sqrt{\frac{2M_0}{\lambda_n}}$. Then let $\mathcal{F} = \{(h,b) \in \mathcal{H}_K \times \mathbb{R} : ||h||_K \leq \sqrt{\frac{2M_0}{\lambda_n}}, |b| \leq 1 + C_K\sqrt{\frac{2M_0}{\lambda_n}}\}$. Let $\widetilde{L}(h,b) = S\Big[T(A(h(\boldsymbol{X})+b))-1\Big]$. For the first term, $(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n)$,

$$\begin{aligned}(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) &\leq \sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)L(h,b) \\ &= \sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) + (\mathbb{P}-\mathbb{P}_n)L(0,0).\end{aligned}$$

When an $(\boldsymbol{x}_i, a_i, s_i)$ triplet changes, the random variable $\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b)$ can change by no more than $\frac{2M_0}{n}$. McDiarmid's inequality (Bartlett and Mendelson 2002, Theorem 9) then implies that with probability at least $1-\delta/4$,

$$\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) \leq \mathbb{E}\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) + M_0\sqrt{\frac{2\log(4/\delta)}{n}}.$$

A similar argument, together with the fact that $\mathbb{E}\mathbb{P}_n L(0,0) = \mathbb{P}L(0,0)$, shows that with probability at least $1-\delta/2$,

$$(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) \leq \mathbb{E}\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Let $D_n' = \{\boldsymbol{X}_i', A_i', S_i'\}_{i=1}^n$ be an independent sample with the same distribution as $(\boldsymbol{X}, A, S)$. Let $\mathbb{P}_n'$ denote the empirical measure on D_n'. Let σ be a uniform $\{\pm 1\}$-valued random variable, and $\sigma_1, \ldots, \sigma_n$ be n independent copies of σ. Then we have

$$\begin{aligned}\mathbb{E}\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) &= \mathbb{E}\sup_{(h,b)\in\mathcal{F}}\mathbb{E}\Big(\mathbb{P}_n'\widetilde{L}(h,b) - \mathbb{P}_n\widetilde{L}(h,b)\Big|D_n\Big) \\ &\leq 2\mathbb{E}\sup_{(h,b)\in\mathcal{F}}\mathbb{P}_n\sigma\widetilde{L}(h,b) \\ &\leq 2\mathbb{E}\mathbb{E}\Big(\sup_{(h,b)\in\mathcal{F}}|\mathbb{P}_n\sigma\widetilde{L}(h,b)|\Big|S_i, A_i, i=1,\ldots,n\Big) \\ &\leq \frac{16M_0}{n}\mathbb{E}\sup_{(h,b)\in\mathcal{F}}\Big|\sum_{i=1}^n\sigma_i\big(h(\boldsymbol{X}_i)+b\big)\Big|.\end{aligned}$$

The last inequality is due to the contraction inequality (Ledoux and Talagrand 1991, Corollary 3.17). The preceding can be further majorized by using Lemma 22 in Bartlett and

Mendelson (2002),

$$\begin{aligned}
\mathbb{E}\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\widetilde{L}(h,b) &\leq \frac{16M_0}{n}\mathbb{E}\sup_{(h,b)\in\mathcal{F}}\Big|\sum_{i=1}^{n}\sigma_i h(\boldsymbol{X}_i)\Big| + \frac{16M_0}{n}(1+C_K\sqrt{\frac{2M_0}{\lambda_n}})\mathbb{E}\Big|\sum_{i=1}^{n}\sigma_i\Big| \\
&\leq \frac{16M_0}{\sqrt{n}}\sqrt{\frac{2M_0}{\lambda_n}}C_K + \frac{16M_0}{\sqrt{n}}(1+C_K\sqrt{\frac{2M_0}{\lambda_n}}) \\
&= \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}).
\end{aligned}$$

Then we have that with probability at least $1-\delta/2$,

$$(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) \leq \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{B.13}$$

By the assumption, (B.12), and (B.13), we obtain that with probability at least $1-\delta$,

$$\mathcal{R}_{T,g}(f_n)-\mathcal{R}^*_{T,g} \leq M_0\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}} + c\lambda_n^{\beta}.$$

Let $\lambda_n = n^{-\frac{1}{2\beta+1}}$. By Theorem 3.3.2, we obtain the final result that with probability at least $1-\delta$,

$$\mathcal{R}(sign(f_n)) - \mathcal{R}^* \leq \widetilde{c}\sqrt{\log(4/\delta)}n^{-\frac{\beta}{2\beta+1}}.$$

Here $\widetilde{c} = M_0\left(16+3\sqrt{2}+32C_K\sqrt{2M_0}\right)+c$. This completes the proof. □

Proof of Lemma 3.3.9

Proof. We first introduce a lemma. It is revised from Lemma 4.1 in Steinwart and Scovel (2007) to adapt to our settings for individualized treatment rules.

Lemma B.3.8. *Let $\mathcal{X}$ be the closed unit ball of the Euclidean space $\mathbb{R}^p$, and P be a distribution on $\mathcal{X}\times\mathcal{A}\times\mathcal{M}$ with regular marginal distribution on $\boldsymbol{X}$. Recall $\delta(\boldsymbol{x}) = \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)$ for $\boldsymbol{x}\in\mathcal{X}$. On $\acute{\mathcal{X}} := 3\mathcal{X}$ we define*

$$\acute{\delta}(\boldsymbol{x}) = \begin{cases} \delta(\boldsymbol{x}) & \text{if } ||\boldsymbol{x}|| \leq 1, \\ \delta\left(\dfrac{\boldsymbol{x}}{||\boldsymbol{x}||}\right) & \text{otherwise,} \end{cases}$$

where $||\cdot||$ *is the Euclidean norm. We also write* $\acute{\mathcal{X}}^+ = \{\boldsymbol{x} \in \acute{\mathcal{X}} : \acute{\delta}(\boldsymbol{x}) > 0\}$*, and* $\acute{\mathcal{X}}^- = \{\boldsymbol{x} \in \acute{\mathcal{X}} : \acute{\delta}(\boldsymbol{x}) < 0\}$*. Finally let* $B(\boldsymbol{x}, r)$ *denote the open ball of radius* r *about* $\boldsymbol{x}$ *in* $\mathbb{R}^p$*. Then for* $\boldsymbol{x} \in \mathcal{X}^+$*, we have* $B(\boldsymbol{x}, \tau_{\boldsymbol{x}}) \subset \acute{\mathcal{X}}^+$*, and for* $\boldsymbol{x} \in \mathcal{X}^-$*, we have* $B(\boldsymbol{x}, \tau_{\boldsymbol{x}}) \subset \acute{\mathcal{X}}^-$.

The proof is simple, and is the same as that of Lemma 4.1 in Steinwart and Scovel (2007). We omit the proof here. In the lemma, the support is enlarged to ensure that all balls of the form $B(\boldsymbol{x}, \tau_{\boldsymbol{x}})$ are contained in the enlarged support. We return to the proof of Lemma 3.3.9.

Let $L_2(\mathbb{R}^p)$ be the L_2-space on $\mathbb{R}^p$ with respect to Lebesque measure, and $\mathcal{H}_\sigma(\mathbb{R}^p)$ be the RKHS of the Gaussian RBF kernel K_σ. The linear operator $V_\sigma : L_2(\mathbb{R}^p) \to \mathcal{H}_\sigma(\mathbb{R}^p)$ defined by

$$V_\sigma \ell(\boldsymbol{x}) = \frac{(2\sigma)^{d/2}}{\pi^{d/4}} \int_{\mathbb{R}^p} e^{-2\sigma^2 ||\boldsymbol{x}-\boldsymbol{y}||^2} \ell(\boldsymbol{y}) d\boldsymbol{y}, \qquad \ell \in L_2(\mathbb{R}^p),\ \boldsymbol{x} \in \mathbb{R}^p,$$

is an isometric isomorphism (Steinwart et al. 2006). Thus we have,

$$\mathcal{A}(\lambda) \leq \inf_{\ell \in L_2(\mathbb{R}^p)} \frac{\lambda}{2} ||\ell||^2_{L_2(\mathbb{R}^p)} + \mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g}. \tag{B.14}$$

With the notation of Lemma B.3.8 we fix a measurable $\acute{f}_P : \acute{\mathcal{X}} \to [-1, 1]$ that satisfies $\acute{f}_P = 1$ on $\acute{\mathcal{X}}^+$, $\acute{f}_P = -1$ on $\acute{\mathcal{X}}^-$, and $\acute{f}_P = 0$ otherwise. For $\ell := (\sigma^2/\pi)^{p/4} \acute{f}_P$, we immediately obtain,

$$||\ell||_{L_2(\mathbb{R}^p)} \leq \left(\frac{81\sigma^2}{\pi}\right)^{p/4} \theta(p), \tag{B.15}$$

where $\theta(p)$ denotes the volume of $\mathcal{X}$. As shown in the proof of Theorem 3.3.2, we have

$$\mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g} = \mathbb{E}(|\delta(\boldsymbol{x})| \cdot |T(V_\sigma \ell(\boldsymbol{x})) - T(d^*(\boldsymbol{x}))|) \leq 2\mathbb{E}(|\delta(\boldsymbol{x})| \cdot |V_\sigma \ell(\boldsymbol{x}) - d^*(\boldsymbol{x})|).$$

Following the same derivations as in the proof of Theorem 2.7 of Steinwart and Scovel (2007), we also obtain

$$|V_\sigma \ell(\boldsymbol{x}) - d^*(\boldsymbol{x})| \leq 8e^{-\sigma^2 \tau^2_{\boldsymbol{x}}/(2p)}.$$

The geometric noise assumption yields

$$\mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g} \leq 16\mathbb{E}(|\delta(\boldsymbol{x})|e^{-\sigma^2\tau_{\boldsymbol{x}}^2/(2p)}) \leq 16C(2p)^{qp/2}\sigma^{-qp}. \tag{B.16}$$

Combining (B.14), (B.15) and (B.16) yields

$$\mathcal{A}(\lambda) \leq \left(\frac{81\sigma^2}{\pi}\right)^{p/2} \theta^2(p)\lambda/2 + 16C(2p)^{qp/2}\sigma^{-qp}.$$

The desired result now follows by taking $\sigma = \lambda^{-\frac{1}{(q+1)p}}$. □

APPENDIX C: SUPPLEMENTS FOR CHAPTER 5

C.1 Details of mirrored residual weighted learning (MRWL)

As explained in Section 4.2.3, MRWL is to find d^* by solving the following optimization problem,

$$d^* \in \arg\min_d \mathbb{E}\left(\frac{|\widehat{R}|}{\pi(A,\boldsymbol{X})}\mathbb{I}\Big(A\text{sign}(\widehat{R}) \neq d(\boldsymbol{X})\Big)\right),$$

where negative residuals are reflected to positive, and accordingly their treatments are switched to opposites. We seek to minimize the following empirical risk to find a decision function f,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\mathbb{I}\Big(A_i\text{sign}(\widehat{R}_i) \neq d_f(\boldsymbol{X}_i)\Big).$$

Similarly with OWL and RWL, we instead seek the decision function f by minimizing a regularized surrogate risk,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)f(\boldsymbol{X}_i)\Big) + \frac{\lambda}{2}||f||^2, \tag{C.17}$$

where $\phi(\cdot)$ is a continuous surrogate loss function, $||f||$ is some norm for f, and λ is a tuning parameter controlling the trade-off between empirical risk and complexity of the decision function f. As the weights $\frac{|\widehat{R}_i|}{\pi(A_i,\boldsymbol{X}_i)}$ are nonnegative, convex surrogate can be employed for efficient computation. In this work, we apply the Huberized hinge loss function (Wang et al. 2008),

$$\phi(u) = \begin{cases} 0 & \text{if } u \geq 1, \\ \frac{1}{4}(1-u)^2 & \text{if } -1 \leq u < 1, \\ -u & \text{if } u < -1. \end{cases}$$

Other loss functions, such as the hinge loss, can be also applied in MRWL. The Huberized hinge loss is smooth everywhere. Hence it has computational advantages in optimization.

We derive an algorithm for linear MRWL, and then generalize it to the case of nonlinear learning through kernel mapping.

C.1.1 Linear decision rules for MRWL

Suppose that the decision function $f(\boldsymbol{x})$ is a linear function of $\boldsymbol{x}$, *i.e.* $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$. Then the associated regime will assign a subject with clinical covariates $\boldsymbol{x}$ into treatment 1 if $\boldsymbol{w}^T\boldsymbol{x} + b > 0$ and -1 otherwise. We define $||f||$ as the Euclidean norm of $\boldsymbol{w}$. Then the optimization problem (C.17) can be rewritten as

$$\min_{\boldsymbol{w},b} \quad \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{R}_i|}{\pi(A_i, \boldsymbol{X}_i)}\phi\Big(A_i\text{sign}(\widehat{R}_i)(\boldsymbol{w}^T\boldsymbol{X}_i + b)\Big).$$

It is a smooth unconstrained convex optimization problem. There are many efficient numerical methods for solving this problem. One example is the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Nocedal 1980). L-BFGS is a quasi-Newton method that approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using a limited amount of computer memory. When we obtain the solution $(\widehat{\boldsymbol{w}}, \widehat{b})$, the decision function is $\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{w}}^T\boldsymbol{x} + \widehat{b}$, and thus the estimated optimal regime is the sign of $\widehat{f}(\boldsymbol{x})$.

C.1.2 Nonlinear decision rule for MRWL

For a nonlinear decision rule, the decision function $f(\boldsymbol{x})$ is represented by $h(\boldsymbol{x}) + b$ with $h(\boldsymbol{x}) \in \mathcal{H}_K$ and $b \in \mathbb{R}$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel function K. The kernel function $K(\cdot, \cdot)$ is a positive definite function mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. The norm in $\mathcal{H}_K$, denoted by $||\cdot||_K$, is induced by the following inner product:

$$< f, g >_K = \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

for $f(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, \boldsymbol{x}_i)$ and $g(\cdot) = \sum_{j=1}^{m} \beta_j K(\cdot, \boldsymbol{x}_j)$. Then minimizing (C.17) can be rewritten as

$$\min_{h,b} \quad \frac{\lambda}{2}||h||_K^2 + \frac{1}{n}\sum_{i=1}^{n} \frac{|\widehat{R}_i|}{\pi(A_i, \boldsymbol{X}_i)} \phi\Big(A_i \text{sign}(\widehat{R}_i)\big(h(\boldsymbol{X}_i) + b\big)\Big). \tag{C.18}$$

Due to the representer theorem (Kimeldorf and Wahba 1971), the nonlinear problem can be reduced to finding finite-dimensional coefficients v_i, and $h(\boldsymbol{x})$ can be represented as $\sum_{j=1}^{n} v_j K(\boldsymbol{x}, \boldsymbol{x}_j)$. Thus the problem (C.18) is transformed to

$$\min_{\boldsymbol{v},b} \quad \frac{\lambda}{2}\sum_{i,j=1}^{n} v_i v_j K(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{n}\sum_{i=1}^{n} \frac{|\widehat{R}_i|}{\pi(A_i, \boldsymbol{X}_i)} \phi\Big(A_i \text{sign}(\widehat{R}_i)\big(\sum_{j=1}^{n} v_j K(\boldsymbol{X}_i, \boldsymbol{X}_j) + b\big)\Big). \tag{C.19}$$

It is still a smooth unconstrained convex optimization problem, and can be solve by, for example, L-BFGS. When we obtain the solution $(\widehat{\boldsymbol{v}}, \widehat{b})$, the decision function is $\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \widehat{v}_j K(\boldsymbol{x}, \boldsymbol{x}_j) + \widehat{b}$, and thus the estimated optimal regime is the sign of $\widehat{f}(\boldsymbol{x})$. Note that if we choose a linear kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T\boldsymbol{z}$, the obtained regime reduces to the previous linear regime. The most widely used nonlinear kernel in practice is the Gaussian Radial Basis Function (RBF) kernel, that is,

$$K_\sigma(\boldsymbol{x}, \boldsymbol{z}) = \exp\Big(-\sigma^2||\boldsymbol{x} - \boldsymbol{z}||^2\Big),$$

where $\sigma > 0$ is a free parameter whose inverse $1/\sigma$ is called the width of K_σ.

C.2 Simulation studies for comparison between RWL and MRWL

In this section, we carried out simulation studies to compare MRWL with RWL. The simulation setups were the same as those in Zhou et al. (2015). Specifically, we generated 5-dimensional vectors of clinical covariates $x_1, \cdots, x_5$, consisting of independent uniform random variables $U(-1, 1)$. The treatment A was generated from $\{-1, 1\}$ independently of X with $P(A = 1) = 0.5$. That is, $\pi(a, \boldsymbol{x}) = 0.5$ for all a and $\boldsymbol{x}$. The response R was normally distributed with mean $Q_0 = \mu_0(\boldsymbol{x}) + \delta_0(\boldsymbol{x}) \cdot a$ and standard deviation 1, where

$\mu_0(\boldsymbol{x})$ is the common effect for clinical covariates, and $\delta_0(\boldsymbol{x}) \cdot a$ is the interaction between treatment and clinical covariates. We considered four scenarios with different choices of $\mu_0(\boldsymbol{x})$ and $\delta_0(\boldsymbol{x})$:

(1) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 1.8(0.3 - x_1 - x_2)$.

(2) $\mu_0(\boldsymbol{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$; $\delta_0(\boldsymbol{x}) = 1.3(x_2 - 2x_1^2 + 0.3)$.

(3) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2 + 0.5x_4^2$; $\delta_0(\boldsymbol{x}) = 3.8(0.8 - x_1^2 - x_2^2)$.

(4) $\mu_0(\boldsymbol{x}) = 1 + x_1^2 + x_2^2 + 2x_3^2$; $\delta_0(\boldsymbol{x}) = 10(1 - x_1^2 - x_2^2)(x_1^2 + x_2^2 - 0.2)$.

The decision boundaries for all scenarios are determined by x_1 and x_2 only. The scenarios have different decision boundaries in truth. The decision boundary is a line in Scenario 1, a parabola in Scenario 2, a circle in Scenario 3, and a ring in Scenario 4.

We compared the performance of the following four methods:

(1) RWL using the Gaussian RBF kernel (RWL-Gaussian).

(2) RWL using the linear kernel (RWL-Linear).

(1) MRWL using the Gaussian RBF kernel (MRWL-Gaussian).

(2) MRWL using the linear kernel (MRWL-Linear).

Residuals were estimated by a linear main effects model for all methods.

For each scenario, we considered two sample sizes for training datasets: $n = 100$ and $n = 400$, and repeated the simulation 500 times. All four methods had at least one tuning parameter. To be consistent with the simulation studies in Zhou et al. (2015), we applied a 10-fold cross-validation procedure to tune parameters. In the comparison, the performances of four methods were evaluated by two criteria: the first criterion was the

value function under the estimated optimal treatment regime when we applied to an independent and large test dataset; the second criterion was the misclassification error rate of the estimated optimal treatment regime from the true optimal treatment regime on the large test dataset. Specifically, a test set with 10,000 observations was simulated to assess the performance. The estimated value function under any regime d was given by $\mathbb{P}_n^*[R\mathbb{I}(A = d(\boldsymbol{X}))/\pi(A, \boldsymbol{X})]/\mathbb{P}_n^*[\mathbb{I}(A = d(\boldsymbol{X}))/\pi(A, \boldsymbol{X})]$ (Murphy 2005), where $\mathbb{P}_n^*$ denoted the empirical average using the test data; the misclassification rate under any regime d was given by $\mathbb{P}_n^*[\mathbb{I}(d(\boldsymbol{X}) \neq d^*(\boldsymbol{X}))]$, where d^* was the true optimal regime which was known when generating the simulated data.

The simulation results are presented in Table C.1. We reported the mean and standard deviation of value functions and misclassification rates over 500 replications. Scenario 1 was a linear example. The model specification for RWL-Linear and MRWL-Linear was correct, and they provided good performance. Both RWL and MRWL, with either the linear kernel or the Gaussian RBF kernel, yielded similar performance.

The decision boundaries in the remaining three scenarios were nonlinear. RWL-Linear and MRWL-Linear were misspecified, and hence they did not perform very well in these three scenarios. We focused on the comparison between RWL-Gaussian and MRWL-Gaussian. In Scenarios 2 and 3, RWL-Gaussian and MRWL-Gaussian achieved similar performance. It was challenging to find the optimal treatment regime in Scenario 4 since the decision boundary was complicated. When the sample size was small ($n = 100$), the performances of RWL-Gaussian and MRWL-Gaussian were similar. However, when the sample size increased, RWL-Gaussian showed better performance than MRWL-Gaussian.

Table C.1: *Mean (std) of empirical value functions and misclassification rates evaluated on independent test data for four simulation scenarios with 5 covariates. The best value function and minimal misclassification rate for each scenario and sample size combination are in bold.*

	$n = 100$		$n = 400$	
	Value	Misclassification	Value	Misclassification
Scenario 1 (Optimal value 2.25)				
RWL-Linear	2.19 (0.04)	0.09 (0.04)	**2.23 (0.01)**	0.05 (0.02)
RWL-Gaussian	2.17 (0.06)	0.11 (0.04)	2.22 (0.02)	0.05 (0.02)
MRWL-Linear	**2.20 (0.04)**	**0.08 (0.03)**	**2.23 (0.01)**	**0.04 (0.02)**
MRWL-Gaussian	2.17 (0.07)	0.10 (0.05)	2.22 (0.02)	0.06 (0.03)
Scenario 2 (Optimal value 1.96)				
RWL-Linear	1.66 (0.08)	0.26 (0.04)	1.74 (0.03)	0.23 (0.02)
RWL-Gaussian	1.75 (0.09)	0.20 (0.05)	1.90 (0.03)	0.10 (0.03)
MRWL-Linear	1.69 (0.08)	0.25 (0.04)	1.75 (0.02)	0.23 (0.01)
MRWL-Gaussian	**1.77 (0.09)**	**0.19 (0.05)**	**1.91 (0.03)**	**0.09 (0.03)**
Scenario 3 (Optimal value 3.88)				
RWL-Linear	3.15 (0.13)	0.34 (0.03)	3.28 (0.04)	0.31 (0.01)
RWL-Gaussian	3.62 (0.12)	0.19 (0.04)	3.82 (0.04)	0.10 (0.02)
MRWL-Linear	3.02 (0.10)	0.37 (0.02)	3.03 (0.04)	0.37 (0.01)
MRWL-Gaussian	**3.67 (0.10)**	**0.18 (0.04)**	**3.84 (0.03)**	**0.09 (0.02)**
Scenario 4 (Optimal value 3.87)				
RWL-Linear	2.42 (0.14)	0.54 (0.09)	2.49 (0.10)	0.58 (0.08)
RWL-Gaussian	**2.68 (0.28)**	**0.43 (0.08)**	**3.49 (0.08)**	**0.22 (0.03)**
MRWL-Linear	2.42 (0.15)	0.55 (0.08)	2.49 (0.10)	0.59 (0.06)
MRWL-Gaussian	2.61 (0.20)	0.45 (0.06)	3.22 (0.10)	0.29 (0.04)

BIBLIOGRAPHY

Allen, G. I. (2013), "Automatic Feature Selection via Weighted Kernels and Regularization," *Journal of Computational and Graphical Statistics*, 22, 284–299.

Ambroise, C. and McLachlan, G. J. (2002), "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Natl. Acad. Sci.*, 99, 6562–6566.

An, L. T. H. and Tao, P. D. (1997), "Solving a Class of Linearly Constrained Indefinite Quadratic Problems by D.C. Algorithms," *Journal of Global Optimization*, 11, 253–285.

Argyriou, A., Hauser, R., Micchelli, C. A., and Pontil, M. (2006), "A DC-programming algorithm for kernel selection." in *ICML*, eds. Cohen, W. W. and Moore, A., ACM, vol. 148 of *ACM International Conference Proceeding Series*, pp. 41–48.

Atiya, A. F. (2005), "Estimating the Posterior Probabilities Using the K-Nearest Neighbor Rule." *Neural Computation*, 17, 731–740.

Baron, G., Perrodeau, E., Boutron, I., and Ravaud, P. (2013), "Reporting of analyses from randomized controlled trials with multiple arms: a systematic review," *BMC Medicine*, 11, 84.

Bartlett, P. L. and Mendelson, S. (2002), "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results." *Journal of Machine Learning Research*, 3, 463–482.

Beygelzimer, A., Kakade, S., and Langford, J. (2006), "Cover Trees for Nearest Neighbor," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA: ACM, pp. 97–104.

Biau, G., Cérou, F., and Guyader, A. (2010), "On the Rate of Convergence of the Bagged Nearest Neighbor Estimate," *Journal of Machine Learning Research*, 11, 687–712.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995), "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific and Statistical Computing*, 16, 1190–1208.

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011), "Analysis of randrandom comparative clinical trial data for personalized treatment selection," *Biostatistics*, 12, 270–282.

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006), "Trading convexity for scalability," in *Proceedings of the 23rd international conference on Machine learning*, New York, NY, USA: ACM, ICML '06, pp. 201–208.

Cover, T. and Hart, P. (1967), "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, 13, 21–27.

Cover, T. M. (1965), "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Transactions on Electronic Computers*, EC-14, 326–334.

Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: The L1 View*, New York: John Wiley.

Devroye, L., Gyorfi, L., Krzyzak, A., and Lugosi, G. (1994), "On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates," *The Annals of Statistics*, 22, 1371–1385.

Devroye, L., Györfi, L., and Lugosi, G. (1996), *A Probabilistic Theory of Pattern Recognition*, Springer.

Dietterich, T. G. and Bakiri, G. (1995), "Solving Multiclass Learning Problems via Error-correcting Output Codes," *Journal of Artificial Intelligence Research*, 2, 263–286.

Domeniconi, C., Peng, J., and Gunopulos, D. (2002), "Locally Adaptive Metric Nearest-Neighbor Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1281–1285.

Efron, B. and Tibshirani, R. (1997), "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548–560.

Fix, E. and Hodges, J. (1951), "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties," Tech. Rep. Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas.

Fleming, T. and Harrington, D. (1991), *Counting Processes and Survival Analysis*, Wiley.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977), "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Transactions on Mathematical Software*, 3, 209–226.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Fung, G. M. and Mangasarian, O. L. (2004), "A Feature Selection Newton Method for Support Vector Machine Classification," *Computational Optimization and Applications.*, 28, 185–202.

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

Gong, P., Zhang, C., Lu, Z., Huang, J., and Ye, J. (2013), "A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems." in *ICML*, vol. 28 of *JMLR Proceedings*, pp. 37–45.

Grandvalet, Y. and Canu, S. (2002), "Adaptive Scaling for Feature Selection in SVMs." in *Advances in Neural Information Processing System*, eds. Becker, S., Thrun, S., and Obermayer, K., MIT Press, pp. 553–560.

Gunter, L., Zhu, J., and Murphy, S. (2011), "Variable Selection for Qualitative Interactions," *Statis-*

tical Methodology, 8, 42–55.

Györfi, L. (1981), "The rate of convergence of kn-NN regression estimates and classification rules." *IEEE Transactions on Information Theory*, 27, 362–364.

— (1991), "Universal consistencies of a regression estimate for unbounded regression functions," in *Nonparametric Functional Estimation and Related Topics*, ed. Roussas, G., Kluwer Academic Publisher, pp. 329–338.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002), *A distribution-free theory of nonparametric regression*, Springer series in statistics, New York, Berlin, Paris: Springer.

Hall, P., Park, B. U., and Samworth, R. J. (2008), "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, 36, 2135–2152.

Hastie, T. and Tibshirani, R. (1996), "Discriminant Adaptive Nearest Neighbor Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 607–616.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning*, Springer.

Havlir, D. V., Kendall, M. A., Ive, P., Kumwenda, J., Swindells, S., Qasba, S. S., Luetkemeyer, A. F., Hogg, E., Rooney, J. F., Wu, X., Hosseinipour, M. C., Lalloo, U., Veloso, V. G., Some, F. F., Kumarasamy, N., Padayatchi, N., Santos, B. R., Reid, S., Hakim, J., Mohapi, L., Mugyenyi, P., Sanchez, J., Lama, J. R., Pape, J. W., Sanchez, A., Asmelash, A., Moko, E., Sawe, F., Andersen, J., and Sanne, I. (2011), "Timing of Antiretroviral Therapy for HIV-1 Infection and Tuberculosis," *New England Journal of Medicine*, 365, 1482–1491.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003), "A practical guide to support vector classification," Tech. rep., Department of Computer Science, National Taiwan University, Taipei 106, Taiwan.

Keller, M., Mccullough, J., Klein, D., Arnow, B., Dunner, D., Gelenberg, A., Markowitz, J., Nemeroff, C., Russell, J., Thase, M., Trivedi, M., and Zajecka, J. (2000), "A comparision of Nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression," *The New England Journal of Medicine*, 342, 1462–1470.

Kimeldorf, G. and Wahba, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of Mathematical analysis and applications*, 33, 82–95.

Lafferty, J. and Wasserman, L. (2008), "Rodeo: Sparse, greedy nonparametric regression," *The Annals of Statistics*, 36, 28–63.

Ledoux, M. and Talagrand, M. (1991), *Probability in Banach Spaces: isoperimetry and processes*, Springer.

Lee, Y., Lin, Y., and Wahba, G. (2004), "Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, 99, 67–81.

Li, J. and Chan, I. S. F. (2006), "Detecting Qualitative Interactions in clinical trials: an extension of range test," *Journal of Biopharmaceutical Statistics*, 16, 831–841.

Lin, Y. and Zhang, H. H. (2006), "Component selection and smoothing in multivariate nonparametric regression," *The Annals of Statistics*, 34, 2272–2297.

Lindenbaum, M., Markovitch, S., and Rusakov, D. (2004), "Selective Sampling for Nearest Neighbor Classifiers," *Machine Learning*, 54, 125–152.

Liu, Y., Wang, Y., Kosorok, M., Zhao, Y., and Zeng, D. (2015), "Augmented multistage outcome-weighted learning for estimating personalized dynamic treatment regimes," Manuscript.

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007), "Demystifying optimal dynamic treatment regimes," *Biometrics*, 63, 447–455.

Morales, J. L. and Nocedal, J. (2011), "Remark on 'Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-scale Bound Constrained Optimization'," *ACM Transactions on Mathematical Software*, 38, 7:1–7:4.

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

— (2005), "An experimental design for the development of adaptive treatment strategies." *Statistics in medicine*, 24.

Newey, W. and McFadden, D. (1986), "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, eds. R., F. E. and D., M., Elsevier, vol. 4, chap. 36, 1st ed., pp. 2111–2245.

Nocedal, J. (1980), "Updating Quasi-Newton Matrices with Limited Storage," *Mathematics of Computation*, 35, 773–782.

Nocedal, J. and Wright, S. J. (2006), *Numerical Optimization*, New York: Springer, 2nd ed.

Peto, R. (1982), "Statistical aspects of cancer trials," in *Treatment of Cancer*, ed. Halnan, K. E., London: Chapman & Hall, pp. 867–871.

Piantadosi, S. and Gail, M. H. (1993), "A comparison of the power of two tests for qualitative interactions," *Statistics in Medicine*, 12, 1239–1248.

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

Rakotomamonjy, A. (2003), "Variable selection using svm based criteria," *J. Mach. Learn. Res.*, 3, 1357–1370.

Robins, J. (1986), "A new approach to causal inference in mortality studies with a sustained exposure periodłapplication to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

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

— (1997), "Causal Inference from Complex Longitudinal Data," in *Latent Variable Modeling and Applications to Causality*, ed. Berkane, M., Springer New York, vol. 120 of *Lecture Notes in Statistics*, pp. 69–117.

— (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, eds. Lin, D. and Heagerty, P., Springer New York, vol. 179 of *Lecture Notes in Statistics*, pp. 189–326.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

Rudin, W. (1987), *Real and Complex Analysis, 3rd Ed.*, New York, NY, USA: McGraw-Hill, Inc.

Samworth, R. J. (2012), "Optimal weighted nearest neighbour classifiers," *The Annals of Statistics*, 40, 2733–2763.

Schmidt, M. (2010), "Graphical model structure learning with l1-regularization," Ph.D. thesis, The University of British Columbia.

Shaw, A. T., Kim, D.-W., Mehra, R., Tan, D. S., Felip, E., Chow, L. Q., Camidge, D. R., Vansteenkiste, J., Sharma, S., De Pas, T., Riely, G. J., Solomon, B. J., Wolf, J., Thomas, M., Schuler, M., Liu, G., Santoro, A., Lau, Y. Y., Goldwasser, M., Boral, A. L., and Engelman, J. A. (2014), "Ceritinib in ALK-Rearranged Non-Small-Cell Lung Cancer," *New England Journal of Medicine*, 370, 1189–1197.

Silvapulle, M. J. and Sen, P. K. (2005), *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, John Wiley & Sons.

Steinwart, I. and Christmann, A. (2008), *Support Vector Machines*, Springer.

Steinwart, I., Hush, D. R., and Scovel, C. (2006), "An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels." *IEEE Transactions on Information Theory*, 52, 4635–4643.

Steinwart, I. and Scovel, C. (2007), "Fast rates for support vector machines using Gaussian kernels," *The Annals of Statistics*, 35, 575–607.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–620.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990), "Martingale-based residuals for survival models," *Biometrika*, 77, 147–160.

Tian, L., Alizadeh, A., Gentles, A., and Tibshirani, R. (2012), "A Simple Method for Detecting Interactions between a Treatment and a Large Number of Covariates," *ArXiv e-prints*.

Tibshirani, R. (1994), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Treggiari, M. M., Retsch-Bogart, G., Mayer-Hamblett, N., Khan, U., Kulich, M., Kronmal, R., Williams, J., Hiatt, P., Gibson, R. L., Spencer, T., Orenstein, D., Chatfield, B. A., Froh, D. K., Burns, J. L., Rosenfeld, M., and Ramsey, B. W. (2011), "Comparative efficacy and safety of 4 randomized regimens to treat early pseudomonas aeruginosa infection in children with cystic fibrosis," *Archives of Pediatrics & Adolescent Medicine*, 165, 847–856.

Treggiari, M. M., Rosenfeld, M., Mayer-Hamblett, N., Retsch-Bogart, G., Gibson, R. L., Williams, J., Emerson, J., Kronmal, R. A., and Ramsey, B. W. (2009), "Early anti-pseudomonal acquisition in young patients with cystic fibrosis: Rationale and design of the EPIC clinical trial and observational study," *Contemporary Clinical Trials*, 30, 256 – 268.

Vapnik, V. N. (1998), *Statistical Learning Theory*, New York: John Wiley and Sons.

Wang, L., Zhu, J., and Zou, H. (2008), "Hybrid huberized support vector machines for microarray classification and gene selection." *Bioinformatics*, 24, 412–419.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2000), "Feature Selection for SVMs," in *Advances in Neural Information Processing System*, MIT Press, pp. 668–674.

Wu, Y. and Liu, Y. (2007), "Robust truncated-hinge-loss support vector machines," *Journal of the American Statistical Association*, 102, 974–983.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., and Shao, J. (2015), "Regularized outcome weighted subgroup identification for differential treatment effects," *Biometrics*, In press.

Yang, Y. and Zou, H. (2013), "An Efficient Algorithm for Computing the HHSVM and Its Generalizations," *Journal of Computational and Graphical Statistics*, 22, 396–415.

Yuille, A. L. and Rangarajan, A. (2003), "The concave-convex procedure," *Neural Computation*, 15, 915–936.

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. A. (1991), "Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials," *Journal of the American Medical Association*, 266, 93–98.

Zelterman, D. (1990), "On tests for qualitative interactions," *Statistics & Probability Letters*, 10, 59–63.

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

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018.

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

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, 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, 1422–1433.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, in press.

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*, in press, http://arxiv.org/pdf/1508.03179.pdf.

Zhu, J. and Hastie, T. (2004), "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, 5, 427–443.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003), "1-Norm Support Vector Machines," in *Advances in Neural Information Processing System*.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.