INCORPORATING BENEFIT-RISK CONSIDERATION AND FEATURE SELECTION INTO OPTIMAL DYNAMIC TREATMENT REGIMENS

Mochuan Liu

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
2023

Approved by:

Donglin Zeng

Anna Kahkoska

Quefeng Li

Xianming Tan

Yuanjia Wang

ABSTRACT

Mochuan Liu: Incorporating Benefit-Risk Consideration and Feature Selection into Optimal Dynamic Treatment Regimens
(Under the direction of Donglin Zeng)

Optimal dynamic treatment regimen (DTR) is one of the most important strategies in precision medicine, which sequentially assigns the best treatment to patients based on their evolving health status to maximize the cumulative outcome. For many chronic diseases, treatments are often multifaceted where aggressive treatments with a higher beneficial reward are usually accompanied by an elevated risk of adverse outcomes, and ideal DTRs should both yield a higher beneficial gain while avoiding unnecessary risk. In addition, it is often that among many possible tailoring variables, only a small subset is essential for treatment, and identifying these variables is particularly important for developing sparse DTRs, which are useful in practice.

To address these challenges, in the first project we propose a new machine learning-based method to learn the optimal DTRs that maximize patients' cumulative reward but at each stage, the acute short-term risk induced by the treatments is controlled lower than a pre-specified threshold. We show that this multistage-constrained problem can be decomposed into a series of single-stage single-constrained problems, which can be efficiently solved using a backward algorithm. We provide theoretical guarantees for the method and demonstrate the performance via simulation studies and an application to a clinical trial for T2D patients (DURABLE study).

In the second project, we develop a general approach to estimate the optimal DTRs that maximize patients' cumulative reward but lead to a cumulative risk no higher than a pre-specified threshold. This procedure converts the problem into solving unconstrained DTRs problems, which can be accommodated to existing DTRs methods. Furthermore, we propose an estimation procedure (MRL) to solve the decision rules across all stages simultaneously. The method is justified via theoretical guarantees, simulation studies, and an application to the DURABLE study.

In the third project, we develop a new machine learning-based method by extending and adding an L1-penalty to the MRL framework to implement variable selection while learning optimal DTRs across

all stages contingently. A DC algorithm is developed to solve the L1-MRL problem efficiently and the performance is demonstrated via simulation studies and application to an observational electronic health record (EHR) data of T2D patients.

To my parents,

Qing Xu and Jun Liu,

who keep encouraging and supporting me

to get through this hard journey in my life.

ACKNOWLEDGEMENTS

Writing the Ph.D. acknowledgement is absolutely the toughest part of my dissertation, at least for me but probably for every (former) Ph.D. candidate, as there are so many people I would like to say thanks to, without whom I might have probably already given up my dream of pursuing a Ph.D. degree and reluctantly let the regret stay with me for the rest of my life. But even though challenging, I would still like to try to first sincerely express my utmost gratitude to Professor Donglin Zeng, my Ph.D. advisor, who unselfishly shared his bright ideas and wisdom with me, which, plus my humble efforts, compose the largest part of this not-too-bad dissertation work. It is really hard to find an appropriate way to best deliver my highest gratitude, but luckily I found a wonderful sentence, which was once used in an elegant proof of a charming theorem, that with slight modifications makes it possible: *I have discovered a truly marvelous way to deliver my gratitude, which this short paragraph is too narrow to contain.*

The completion of this work also relies on the help from many other people. Among them, I would like to sincerely thank Professor Yuanjia Wang at the first place, who provided tremendous, vital and irreplaceable help and contributions to all the projects presented in this dissertation. I would also like to thank my committee members, Professor Quefeng Li, Professor Anna Kahkoska, and Professor Xianming Tan, who provide valuable suggestions and comments that helped to improve and enrich our work decently. There are also many people whose contributions cannot be properly reflected in the dissertation but without whom I will never be able to proceed with my Ph.D. career. Among them, I want to express my best appreciation to Professor Yu Gu (HKU), Professor Xinming An (UNC-Chapel Hill), and Professor Samuel McLean (UNC-Chapel Hill), who kindly offered financial support that helped me to keep pursuing my Ph.D. degree. Lastly, I would also like to express my sincere gratitude to Professor Lu Wang at the University of Michigan at Ann Arbor, who firmly persuaded me to keep pursuing a Ph.D. degree during the hardest time in my life when my road ahead was shrouded under the mist of failure and uncertainty that finally leads to who I am today.

All my friends I met and get acquainted with at UNC-Chapel Hill are also definitely and inevitably part of the success that I achieve today. Their friendship and encouragement are the most indispensable and irreplaceable sources of joy and happiness which cure the loneliness, frustration, and depression that one

must face during one's Ph.D. career. Unfortunately, there are so many people that I truly want to convey my best wishes, whose names I cannot list all here, and I would like to sincerely thank all of them from the bottom of my heart. The support from my family is also crucial that backed me up till today, and I would also like to thank all my family members (except my parents who I have reserved the entire dedication section solely for them to express my deepest love) for their concern and support.

I would also like to say thanks to all staff in the Department of Biostatistics at UNC-Chapel Hill who serve and assist me to deal with all the tedious but mandatory daily affairs during my Ph.D. career. And finally, I would also like to specially say thanks to all staff working at Jade Palace, Rasa Malaysia, Taipei 101, and the sushi chefs at Harris Teeter in Chatham, whose names I may probably never have the chance to learn, but their devotion to their work and their customers have made my life in Chapel Hill much more tasty, enjoyable and comfortable. A Ph.D. dissertation is never a solely individual achievement or a simple collaboration between a student and his/her advisors but is an integrated effort from numerous people who work hard in their own positions and dedicate themselves to their own responsibilities. Completing a Ph.D. dissertation is one of the typical examples that nothing can be achieved by merely struggling oneself for one's best, and I hope that my endeavor for my future career, wherever it might be, could also help someone to pursue and achieve his/her own career, which will be my acknowledgement left for myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DTR(s)	Dynamic Treatment Regimen(s)
SMART	Sequential Multiple Assignment Randomized Trial
SUTV	Stable Unit Treatment Value (Assumption)
NUC	No Unmeasured Confounders (Assumption)
(A)OWL	(Augmented) Outcome Weighted Learning
SVM	Support Vector Machine
RKHS	Reproducing Kernel Hilbert Space

CHAPTER 1: INTRODUCTION

1.1 Study background

Personalized medicine aims at tailoring treatments to patients so that treatments are best suited to each individual via leveraging patient's health heterogeneity during treatment design (Kosorok and Laber, 2019; Roberts et al., 2020). The awareness of personalized medicine was driven by the recognition that for many complex diseases, there is no optimal *one-size-fits-all* treatment that can best fit all patients sharing the same diagnosis, and the advocacy of personalized medicine can be traced back to late 20th centuries (Sørensen, 1996; Longford and Nelder, 1999). Many data-driven strategies have been developed to substantiate the idea of personalized medicine in the past two decades, and among them *dynamic treatment regimens (DTRs)* focuses on finding the optimal sequence of decision rules based on patients' evolving health status so that patients' cumulative medical reward is maximized (Chakraborty and Moodie, 2013; Laber et al., 2014).

For many DTRs studies, the statistical problem is usually simplified as an optimization problem where the unique goal is to find the optimal DTRs that maximize patients' cumulative beneficial rewards. However, from the application perspective, it has been debated that the potential side effect, treatment cost and patient's personal preference should also be fully concerned and addressed when personalizing treatments to patients, which is particularly important for treatments of chronic diseases such as cancer and diabetes (Krzyszczyk et al., 2018; Chung et al., 2020). Only a few recent studies have ever been proposed to tackle the optimal DRTs problem when restrictions such as safety or budget limit should also be assessed and controlled. Motivated by the demand from real applications, in this dissertation, we focus on developing new DTRs methods to incorporate additional constraint(s) in DTRs estimation.

In this section, we provide a general review of existing standard optimal DTRs methods and recent developments of DTRs methods with consideration of additional restrictions in literature. The remaining sections are organized as follow: in Section 1.2, we review existing standard DTRs methods which aims at maximizing certain beneficial reward; in Section 1.3, we go over existing DRTs studies where additional requirements of DTRs need to be satisfied; in Section 1.4, we discuss the main motivation of three main problems studied in this dissertation and the outline of remaining chapters.

1.2 Review of standard optimal DTRs methods

1.2.1 Statistical framework of standard DTRs problem

For a standard T-stage DTRs problem, the data consists of $\{(H_t, A_t, Y_t)\}_{t=1}^T$. Here, $H_t \in \mathcal{H}_t \subseteq \mathbb{R}^{d_t}$ denotes patients' feature variables at stage t which can be leveraged to determine t-stage treatment assignment, $A_t \in \mathcal{A}_t$ denotes the observed treatment assignment at stage t with $\mathcal{A}_t$ denoting the set of available treatments, and $Y_t \in \mathbb{R}$ denotes the instant reward by the end of stage t. In practice, we assume that $H_{t+1} = (H_t, A_t, O_t)$ which includes all previous stages' feature variables H_t, treatment assignment of previous stage A_t and additional time-dependent covariates O_t which are only observed after treatment assignment happened. In this work, we only focus on the problem where $\{\mathcal{A}_t\}_{t=1}^T$ are discrete sets. For convenience, we further reduce the problem to the case when only two treatments are available at each stage, denoted as $\{-1, 1\}$, i.e., $\mathcal{A}_t = \{-1, 1\}$ for $t = 1, ..., T$. In the end, we assume that higher outcome Y_t indicates better treatment performance and the cumulative reward gained at final stage T is measured by $Y = \sum_{t=1}^T Y_t$.

We say

$$\mathcal{D} = (\mathcal{D}_1, ..., \mathcal{D}_t) : \mathcal{H}_1 \times \cdots \times \mathcal{H}_T \to \mathcal{A}_1 \times \cdots \times \mathcal{A}_T \text{ where } \mathcal{D}_t : \mathcal{H}_t \to \mathcal{A}_t$$

is a sequence of DTRs. Under $\mathcal{D}$, patients with health status $h_t \in \mathcal{H}_t$ at stage t will be assigned with treatment $\mathcal{D}_t(h_t)$. The optimal DTRs problem aims at finding the optimal rules $\mathcal{D}^*$ such that

$$\mathcal{D}^* \in \arg\max_{\mathcal{D}} E^{\mathcal{D}}[Y]$$

where $E^{\mathcal{D}}[\cdot]$ denotes the expectation under $A_t = \mathcal{D}_t(H_t)$ for $t = 1, ..., T$. In other words, we would like to learn decision rules $\mathcal{D}^*$ so that patient's average cumulative reward will be maximized by following treatments $A_t = \mathcal{D}^*(H_t)$ at each stage.

Additional assumptions need to be imposed to ensure that $E^{\mathcal{D}}[Y]$ is learnable given observed data. To this end, we introduce some useful causal inference notations. Let $\bar{A}_t = (A_1, ..., A_t)$ denote the sequence of observed treatments and $\bar{a}_t = (a_1, ..., a_t) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$ denote a sequence of determined treatments combination prior to stage t. We use $X(\bar{a}_t)$ to denote the potential outcome of an arbitrary random variable X under treatment combination $\bar{a}_t$. Throughout this work, we also use $p(A_t = a_t | H_t = h_t)$ to denote the

treatment assignment probability of $A_t = a_t$ given $H_t = h_t$. For the standard DTRs problem, we assume that the following three assumptions hold:

- *Stable Unit Treatment Value (SUTV)*: at each stage, the subject's outcomes are not influenced by other subjects' treatments allocation, i.e., for any $t = 1, ..., T$,

$$Y_t = Y_t(\bar{a}_t) \quad \text{given} \quad \bar{A}_t = \bar{a}_t$$

  holds for any $\bar{a}_t \in \{-1, +1\}^t$.

- *No Unmeasured Confounders (NUC)*: for any $t = 1, ..., T$,

$$A_t \perp\!\!\!\perp (O_{t+1}(\bar{a}_t), Y_{t+1}(\bar{a}_{t+1}), ..., O_T(\bar{a}_{T-1}), Y_T(\bar{a}_T)) \big| H_t$$

  holds for any $\bar{a}_T \in \{-1, +1\}^T$.

- *Positivity*: for any $t = 1, ..., T$, there exists universal constants $0 < c_1 \leq c_2 < 1$ such that

$$c_1 \leq p(A_t = 1|H_t) \leq c_2 \quad \text{for } H_t \text{ a.s.}$$

Assumptions above are standard causal assumptions in many DTRs literature (Chakraborty and Moodie, 2013). SUTV was first proposed in Rubin (1980) and is equivalent to the *no interference between units* assumption introduced in Cox (1992). The SUTV assumption ensures that each subject in the data is independent and not affected by other subjects. In particular, SUTV is a reasonable assumption when data is collected from a randomized trial drawn from a large population. The NUC assumption is sometimes also known as *sequential randomization assumption*, *sequential ignorability* or *exchangeability* in the literature. Under NUC assumption, the treatment assignment A_t will be independent of all future states and outcomes conditioning on past treatment history H_t (Robins, 1997). Mathematically, under NUC assumption we have

$$p(A_t|H_t, O_{t+1}(\bar{a}_t), Y_{t+1}(\bar{a}_{t+1}), ..., O_T(\bar{a}_{T-1}), Y_T(\bar{a}_T)) = p(A_t|H_t),$$

so at each stage, the treatment assignment can be viewed as a randomized trial with randomization probability $p(A_t|H_t)$. The NUC assumption is automatically guaranteed when data is collected from a simple

randomization trial or *sequential multiple assignment randomized trial (SMART)* (Murphy, 2005*b*). For observational data, the NUC assumption will also approximately holds when all relevant confounders have been observed and controlled. The last positivity assumption is also sometimes referred to *experimental treatment assignment (ETA)* assumption, which ensures that treatments to be evaluated at each stage will have a positive chance to be observed conditioning on the past trajectory. In recent years, new causal frameworks and personalized medicine methods have also been proposed to tackle the problem when either SUTV (Jiang, Wallace and Thompson, 2022), NUC (Kallus and Zhou, 2019; Bennett and Kallus, 2019; Cui and Tchetgen Tchetgen, 2021; Chen and Zhang, 2021; Saghafian, 2022; Rose, Moodie and Shortreed, 2022; Fu et al., 2022), or positivity assumption (van der Laan and Petersen, 2007; Li and Li, 2019; Zhou et al., 2023) is violated.

1.2.2 Regression-based DTRs methods

One type of approach to solving the optimal DTRs problem is known as the regression-based method. For regression-based methods, optimal DTRs are estimated via modeling the expected reward under different treatments or different variants. To illustrate the idea behind regression-based methods, we consider the single-stage DTR problem and use (H, A, Y) to denote the feature variables, observed treatment assignment and reward. The single problem then becomes

$$\max_{\mathcal{D}:\mathcal{H}\to\{-1,1\}} E^{\mathcal{D}}[Y].$$

Define *Q-function* to be $Q(h,a) \coloneqq E[Y|H=h, A=a]$ and let the *conditional average treatment effect (CATE)* to be

$$C(h) \coloneqq Q(h,1) - Q(h,-1).$$

Then when both SUTV, NUC and positivity assumptions hold, it can be verified that the optimal decision rule $\mathcal{D}^*$ is given by

$$\mathcal{D}^*(h) \overset{(i)}{=} \text{sign}(C(h)) \overset{(ii)}{=} \underset{a\in\{-1,1\}}{\arg\max}\ Q(h,a). \tag{1.1}$$

Therefore, (1.1) indicates that the estimation of $\mathcal{D}^*$ can be achieved via either modeling the CATE or Q-function using observed data, which leads to the *structured mean model (SMM)* (Robins, Rotnitzky and Zhao, 1994) and *Q-learning* (Watkins, 1989; Qian and Murphy, 2011) in personalized medicine literature.

Both SMM and Q-learning can be extended to the problem with more than two stages, however, such extension is non-trivial. The main challenge is the existence of so-called *delayed treatment effect* where treatments adopted at stage t may not only affect the instant reward Y_t but can also influence the final outcome Y via affecting future time-dependent covariates including future instant rewards. As a result, decision rules that maximize future reward conditioning on the past observed trajectory may eliminate potential delayed treatments effect and lead to suboptimal rules (Almirall, Ten Have and Murphy, 2010).

The delayed treatment effect can be fully accounted for by using the well-unknown Bellman equation in reinforcement learning literature (Bellman, 1966; Sutton and Barto, 1998). Specifically, for a given DTRs problem, we define the stagewise Q-function to be

$$Q_t(h_t, a_t) \coloneqq E[Y_t + \max_{a_{t+1}\in\{-1,1\}} Q_{t+1}(H_{t+1}, a_{t+1})|H_t = h_t, A_t = a_t], \quad t = 1, ..., T,$$

with $Q_{T+1} = 0$. Then Bellman equation indicates that the optimal decision rules will satisfy

$$\mathcal{D}^*(h_t) = \arg\max_{a_t\in\{-1,1\}} Q_t(h_t, a_t), \quad t = 1, ..., T. \tag{1.2}$$

To extend SMM to $T \geq 2$ using (1.2), we define the stagewise *g-outcome* to be

$$Y_t^{(g)} \coloneqq \sum_{s=t}^{T} Y_s - \sum_{s=t+1}^{T} \text{sign}(C_t(H_t))C_t(H_t)$$

where $C_t(H_t) \coloneqq Q_t(H_t, 1) - Q_t(H_t, -1)$ denotes the t-stage CATE function. By noting that

$$E[Y_t^{(g)}|H_t, A_t] = Q_t(H_t, A_t) = \frac{1}{2}(Q_t(H_1, 1) + Q_t(H_t, -1)) + \frac{1}{2}A_tC_t(H_t),$$

the t-stage optimal treatment rule can be estimated by fitting the semi-parametric model

$$Y_t^{(g)} = m_t(H_t) + \frac{1}{2}A_tC_t(H_t; \beta_t) + e_t^{(g)}; \quad E[e_t^{(g)}|H_t, A_t] = 0, \tag{1.3}$$

where $C_t(H_t;\beta_t)$ is a parametric model of $C_t(H_t)$ indexed by some unknown parameters β_t. The optimal decision rules can be approximated by

$$\widehat{D}_t(h_t) = \text{sign}(C(h_t;\widehat{\beta})).$$

Model (1.3) is a special case of the optimal *structural nested mean model (SNMM)* proposed by Robins (2004). Several approaches have been proposed to implement (1.3) given observed data, including A-learning (Blatt, Murphy and Zhu, 2004; Shi et al., 2018), dynamic weighted ordinary least squares (dWOLS) (Huang, Ning and Wahed, 2014; Wallace and Moodie, 2015), regret regression (Murphy, 2003; Almirall, Ten Have and Murphy, 2010; Henderson, Ansell and Alshibani, 2010; Almirall et al., 2014) and G-estimation (Robins, 2004).

Alternatively, the optimal decision rules can be estimated via modeling the expected reward under each treatment rule according to equation (ii) in (1.1). Following this idea, Qian and Murphy (2011) proposed a backward induction procedure, still namely *Q-learning*, to estimate the optimal decision rules iteratively from the final stage T to the initial stage. Specifically, let t-stage *q-outcome* to be

$$Y_t^{(q)} := Y_t + Q_{t+1}(H_{t+1}, \mathcal{D}^*(H_t))$$

then it can be verified that $E[Y_t^{(q)}|H_t, A_t] = Q_t(H_t, A_t)$. The optimal DTRs can be therefore determined via modeling

$$\widehat{Y}_t^{(q)} = Y_t + \max_{a_t\in\{-1,1\}} \widehat{Q}_{t+1}(H_{t+1}, a_{t+1}),$$

where response variable $\widehat{Y}_t^{(q)}$ can be calculated using the estimated Q-function from stage $t+1$ to T. The estimated rule at stage t is then given by

$$\widehat{D}_t(h_t) = \text{sign}(\widehat{Q}_t(h_t, 1) - \widehat{Q}_t(h_t, -1)).$$

When $Q_t(H_t, A_t)$ is assumed to be linear in terms of (H_t, A_t) and their interaction terms, a typical choice of the linear regression model is assuming that

$$Q_t(H_t, A_t) \sim \alpha_{0t} + H_t\alpha_t + A_t(\beta_{0t} + H_t\beta_t).$$

To lessen the model misspecification, refinements are also explored to mitigate Q-function estimation error, including methods via penalized regression model (Qian and Murphy, 2011; Song, Wang, Zeng and Kosorok, 2015), nonlinear model (Laber, Linn and Stefanski, 2014), generative additive model (Moodie, Dean and Sun, 2014), support vector regression and random tree (Zhao, Kosorok and Zeng, 2009), decision tree-based model and kernel regression (Zhang et al., 2018), bayesian model (Murray, Yuan and Thall, 2018) and robust regression model (Ertefaie et al., 2021). Q-learning in DTRs is also closely related to the policy iteration methods in reinforcement learning literature and useful methods including Q-iteration (Ernst, Geurts and Wehenkel, 2005), deep Q-learning (Mnih et al., 2015) and robust policy search methods (Zhang et al., 2013; Jiang and Li, 2016).

1.2.3 Machine learning-based DTRs methods

Different from regression-based methods, another type of method, known as machine learning-based methods, solves the optimal DTRs problem by directly maximizing a value function. Starting with the single-stage problem, the intuition behind machine learning-based methods is to note that under SUTV, NUC and positivity assumptions, the expected reward under arbitrary decision rule $\mathcal{D}$ can be expressed by the expectation of *inverse probability estimator (IPW)*

$$\mathcal{V}(\mathcal{D}) \coloneqq E\left[Y\frac{\mathbb{I}(A=\mathcal{D}(H))}{p(A|H)}\right], \tag{1.4}$$

assuming that the treatment assignment probability is known. Note that (1.4) is indeed a weighted binary classification problem with true label A and weight $Y/p(A|H)$, using this observation Zhao et al. (2012) proposed to estimate the single-stage optimal DTR problem via solving a weight classification problem. Specifically, assume that $\mathcal{D}$ can be expressed by the sign of some measurable function $f : \mathcal{H} \to \mathbb{R}$ and let $\mathcal{F}$ denotes the set of all measured functions, then one can consider the optimization problem

$$\max_{f\in\mathcal{F}} \mathbb{P}_n\left[Y\frac{\mathbb{I}(Af(H)>0)}{p(A|H)}\right].$$

However, due to the existence of the indicator function, solving the optimization above is NP-hard given finite observed data. To overcome this numerical challenge, following the idea from SVM, Zhao et al. (2012) suggests replacing the indicator function with the hinge loss function defined as $\phi(x) = (1-x)_+$ (Cortes

and Vapnik, 1995) and proposes the outcome weighted learning (OWL), which learns the optimal treatment assignment by solving the surrogate minimization problem

$$\min_{f\in\mathcal{G}} \mathbb{P}_n\left[Y\frac{\phi(Af(H))}{p(A|H)}\right]+\lambda_n\|f\|^2_{\mathcal{G}},$$

where the last term is a typical choice of regularization term to reduce overfitting and $\mathcal{G}$ denotes a subset of $\mathcal{F}$. The same classification framework was studied by Zhang, Tsiatis, Laber and Davidian (2012) and Zhang and Zhang (2018), and regression approaches were proposed to solve the binary classification problem. Other than hinge loss, the use of quadratic loss was studied in Qi and Liu (2018); Shah, Fu and Kosorok (2021) and non-convex loss was explored in Huang and Fong (2014); Qiu, Zeng and Wang (2018); Jiang et al. (2020). Alternatively, when $\mathcal{D}$ is assumed from the class of decision tree, different splitting criteria for searching the optimal decision tree based on (1.4) was also studied in Laber and Zhao (2015); Zhu et al. (2017); Kallus (2017).

For observational studies, treatment assignment probability is usually unknown and needs to be estimated from the observed data. In this case, (1.9) can be modified to accommodate observational data via replacing the true assignment probability $p(A|H)$ by any consistent estimator $\widehat{p}(A|H)$, which can still guarantee that the IPW estimator is consistent. To further improve the accuracy, the *augmented IPW (AIPW)* estimator was studied to replace the standard IPW estimator. The AIPW estimator is defined to be

$$\widehat{\mathcal{V}}_{AIPW}(\mathcal{D}) \coloneqq \mathbb{P}_n\left[Y\frac{\mathbb{I}(A=\mathcal{D}(H))}{\widehat{p}(A|H)}-\frac{\mathbb{I}(A=\mathcal{D}(H))-\widehat{p}(A|H)}{\widehat{p}(A|H)}\widehat{E}[Y|H,A=\mathcal{D}(H)]\right] \tag{1.5}$$

Here, $\widehat{p}(A|H)$ denotes an arbitrary estimator of the treatment assignment probability $p(A|H)$, and $\widehat{E}[Y|H,A]$ denotes an arbitrary estimator of the Q-function. It can be shown that when either $\widehat{p}(A|H)$ or $\widehat{E}[Y|H,A]$ is unbiased, the AIPW will be an unbiased estimator of the true expected reward under treatment rule $\mathcal{D}$. This property is called *doubly robust* (Dudik, Langford and Li, 2011; Zhang, Tsiatis, Laber and Davidian, 2012) and an interpretation based on missing data imputation was provided in Robins, Rotnitzky and Zhao (1994). Different methods were proposed to estimate the optimal decision rule based on (1.5), including classification approach (Zhang, Tsiatis, Davidian, Zhang and Laber, 2012), genetic algorithm approach (Zhang et al., 2015), decision tree-based approach (Tao, Wang and Almirall, 2018; Athey and Wager, 2021; Zhou, Athey and Wager, 2022) and deep learning approach (Liang, Lu and Song, 2018). Further refinements of AIPW to

efficiently learn optimal rule was also studied in Zhao et al. (2019); Liang et al. (2021) assuming that the treatment assignment probability or Q-function is only biasedly estimated for some of the subjects.

The single-stage OWL proposed by Zhao et al. (2012) can also be extended to tackle the DTRs problem with $T \geq 2$. Analogous to the extension of single-stage Q-learning to multistage problems, the extension of OWL can be achieved by sequentially excluding subjects whose future treatment rules were not optimal and decomposing the problem into a series of single-stage weight classification problems. Specifically, for $T \geq 2$ the IPW of the expected reward under decision rule $\mathcal{D}$ can be given by

$$\widehat{\mathcal{V}}_{IPW}(\mathcal{D}) \coloneqq \mathbb{P}_n \left[Y \frac{\prod_{t=1}^{T} \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right].$$

Based on this observation, Zhao et al. (2015) proposed the *backward O-learning (BOWL)* to approximate the optimal decision functions $\{\widehat{f}_t\}_{t=1}^{T}$ via solving

$$\widehat{f}_t \in \arg\min_{f \in \mathcal{G}_t} \mathbb{P}_n \left[Y \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s \widehat{f}_s(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \phi(A_t f(H_t)) \right] + \lambda_n \|f\|_{\mathcal{G}_t}^2 \tag{1.6}$$

like Q-learning sequentially from $t = T$ to $t = 1$. Here, $\mathcal{G}_t$ denotes a subset of the set of all measurable functions from $\mathcal{H}_t$ to $\mathbb{R}$.

Since subjects whose future treatment rules do not follow the estimated optimal treatment rules will be assigned with weight 0 in expression (1.6), which will eliminate the contribution of these subjects to all early stages' estimation, this leads to a major limitation for BOWL that the sample size will decrease exponentially as induction proceeds from stage T to 1, unless the majority of the subjects received optimal treatments cross all stages which is very unlikely in real data and particularly will not happen in a randomized trial. To address this issue, Liu et al. (2018) proposed the *augmented OWL (AOWL)*. Concretely speaking, give estimated decision functions $(\widehat{f}_t, ..., \widehat{f}_T)$ the augmented estimated Q-function is defined to be

$$\widehat{Q}_{it} = (\sum_{s=t}^{T} Y_{is}) \frac{\prod_{s=t}^{T} \mathbb{I}(A_{is} \widehat{f}_s(H_{is}) > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} - \sum_{s=t}^{T} \frac{\prod_{l=t}^{s-1} \mathbb{I}(A_{il} \widehat{f}_l(H_{il}) > 0)}{\prod_{l=t}^{s-1} p(A_{il}|H_{il})} \left(\frac{\mathbb{I}(A_{is} \widehat{f}_s(H_{is}) > 0)}{p(A_{is}|H_{is})} - 1 \right) \widehat{m}_{t,s}(H_{is}),$$

where $\widehat{m}_{t,s}(H_{is})$ is the weighted least squares of

$$\sum_{i=1}^{n} \frac{\prod_{l=t}^{T} \mathbb{I}(A_{il} \widehat{f}_l(H_{il}) > 0)}{\prod_{l=t}^{T} p(A_{il}|H_{il})} \frac{1 - p(A_{is}|H_{is})}{\prod_{l=t}^{s} p(A_{il}|H_{il})} \left(\sum_{l=t}^{T} Y_{il} - m_{t,s}(H_{is}) \right)^2.$$

Let $\widehat{Y}_{i,t-1} = \widehat{m}_{t-1}(H_{i,t-1})$ be the least squares of

$$\sum_{i=1}^{n}(Y_{i,t-1} + \widehat{Q}_{it} - m_{t-1}(H_{i,t-1}))^2,$$

then the $t-1$ stage's optimal decision function is approximated by solving the optimization problem

$$\min_{f\in\mathcal{G}_t} \frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{R}_{i,t-1}}{p(A_{i,t-1}|H_{i,t-1})}\phi(\widetilde{A}_{i,t-1}f(H_{i,t-1})) + \lambda_{n,t-1}\|f\|^2_{\mathcal{G}_{t-1}},$$

where $\widetilde{R}_{i,t-1} = Y_{i,t-1} + \widehat{Q}_{it} - \widehat{Y}_{i,t-1}$ and $\widetilde{A}_{i,t-1} = A_{i,t-1}\text{sign}(\widetilde{R}_{i,t-1})$. The intuition behind the AOWL is to impute the expected reward under the optimal decision rules for subjects whose future treatment assignments do not follow the optimal rules. In addition, replacing the imputed response variable $Y_{i,t-1} + \widehat{Q}_{it}$ by its residual $\widetilde{R}_{i,t-1}$ will further reduce the variability in response variable and improve the estimation performance. Numerical and theoretical results indicate that AOWL will have better performance than OWL when the sample size is relatively small. For convenience, in the following chapter, we use O-learning to refer to either OWL, BOWL or AOWL when the context is clear.

1.3 DTRs methods in consideration of additional constraint

Different from flourishing studies that focus on standard DTRs, only a few methods have ever been proposed to address the application when additional constraints, such as the potential side effects or the cost of the treatment, must be fulfilled during optimal treatment rules design and many are restricted to single-stage problems.

Among existing literature, most of the studies consider introducing utility function to combine different outcomes into a univariate outcome (Houede et al., 2010; Thall, Nguyen and Estey, 2008; Thall, 2012; Lee et al., 2015; Butler et al., 2018; Luckett et al., 2021). Specifically, for $T = 1$ we let $(R_1, ..., R_K) \in \mathbb{R}^{d_K}$ denote K arbitrary response variables, which can include either beneficial rewards, adverse risks or patients' preference. A utility function is any prespecified function $U : \mathbb{R}^K \to \mathbb{R}$, and in utility-based approaches, the optimal DTRs are estimated via maximizing

$$\max_{\mathcal{D}} \mathbb{P}_n^{\mathcal{D}}\Big[U(R_1, ..., R_K; H)\Big].$$

The utility function can be selected according to different goals that need to be achieved. However, the main limitation of the utility-based approach is that the choice of the utility function is very subjective and how different choices of utility will affect the optimal decision rules is hard to quantify.

One of the key goals of personalized medicine is to reduce the adverse impact of treatments to avoid unnecessary harm treatments may cause to patients. Lizotte, Bowling and Murphy (2012); Laber, Lizotte and Ferguson (2014) proposed algorithms to maximize multiple outcomes through modeling a series of conditional expectations via regression similar to Q-learning, and as pointed out in Kosorok and Moodie (2015), the method proposed in Laber, Lizotte and Ferguson (2014) can be implemented to estimate the optimal treatment regimen to maximize the beneficial reward while controlling unnecessary risk via a grid search procedure, but the computation is intense and lack of theoretical justification to guarantee that the estimated rule is optimal.

Instead, Wang, Fu and Zeng (2018) proposed to incorporate the risk consideration as an additional constraint and estimate the optimal DTR via solving a constrained optimization problem. Specifically, consider a single-stage optimal treatment regimen problem with Y denoting the beneficial reward that needs to be maximized and R denoting the adverse risk that needs to be avoided. Given prespecified risk constraint τ, Wang, Fu and Zeng (2018) considers following *benefit-risk tradeoff* optimal treatment regimen problem

$$\max_{\mathcal{D}} \ E^{\mathcal{D}}[Y], \quad \text{subject to } E^{\mathcal{D}}[R] \leq \tau. \tag{1.7}$$

By imposing the additional constraint, the estimated rule is guaranteed to maximize beneficial reward and meanwhile secure that the expected risk will also be lower than the prespecified risk threshold τ to ensure treatment safety. Analogous to standard DTRs problem, additional causal assumptions need to be made to ensure that (1.7) is learnable given observed data. To this end, we assume that

- For any $a \in \{-1, +1\}^t$, $(Y, R) = (Y(a), R(a)) \quad \text{given} \quad A = a$.
- For any $a \in \{-1, +1\}^t$, $A \perp\!\!\!\perp (Y(a), R(a))|H$.
- There exist $0 < c_1 \leq c_2 < 1$ such that $c_1 \leq p(A|H) \leq c_2$ hold for H almost surely.

The three assumptions above are SUTV, NUC and positivity assumption in the context of benefit-risk consideration when $T = 1$.

Suppose that the baseline feature variable H has a continuous density function, then it has been shown in Wang, Fu and Zeng (2018) that the true optimal rule of (1.9) can be explicitly derived using the same argument for finding the optimal rejection region in the proof of Neyman-Pearson lemma. Specifically, it can be shown that the optimal decision rule $\mathcal{D}^*$ is given by

$$\mathcal{D}^*(H) = \text{sign}(\delta_Y(H) - \gamma^* \delta_R(H)) \tag{1.8}$$

where

$$\delta_Y(H) \coloneqq E[Y|H, A = 1] - E[Y|H, A = -1],$$

$$\delta_R(H) \coloneqq E[R|H, A = 1] - E[R|H, A = -1]$$

are CATE function w.r.t. Y and R with γ^* being a positive constant such that the expected risk is equal to τ under $\mathcal{D}^*$. As a direct observation, expression (1.8) provides a natural estimation method, namely BR-M, to estimate $\mathcal{D}^*$ via modeling the conditional mean $E[Y|H, A = \pm 1]$ and $E[R|H, A = \pm 1]$. Provided with unbiased estimators $\widehat{E}[Y|H, A = \pm 1]$ and $\widehat{E}[R|H, A = \pm 1]$, then function $\delta_Y(H)$ and $\delta_R(H)$ can be approximated by

$$\widehat{\delta}_Y(H) = \widehat{E}[Y|H, A = +1] - \widehat{E}[Y|H, A = -1],$$

$$\widehat{\delta}_R(H) = \widehat{E}[R|H, A = +1] - \widehat{E}[R|H, A = -1],$$

and the optimal rule can be estimated by grid searching $\widehat{\gamma}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{E}[R|H_i, A = 1]\mathbb{I}(\widehat{\delta}_Y(H_i) - \widehat{\gamma}\widehat{\delta}_R(H_i) > 0) + \widehat{E}[R|H_i, A = -1]\mathbb{I}(\widehat{\delta}_Y(H_i) - \widehat{\gamma}\widehat{\delta}_R(H_i) < 0)\Big] \approx \tau.$$

Like Q-learning, the performance of BR-M may significantly worsen if the conditional mean models are not correctly specified.

Alternatively, to avoid the impact of model misspecification, Wang, Fu and Zeng (2018) also proposed a machine learning-based approach, namely BR-O, which directly solves (1.9) as an optimization problem without involving model estimation. By assuming that the optimal decision rule can be expressed by the sign

of a decision function, under three causal assumptions it can be shown that (1.8) can be reformulated as

$$\max_{f\in\mathcal{F}} E\left[Y\frac{\mathbb{I}(Af(H)>0)}{p(A|H)}\right], \quad \text{subject to } E\left[R\frac{\mathbb{I}(Af(H)>0)}{p(A|H)}\right]\leq\tau. \tag{1.9}$$

Again, like O-learning the indicator functions in both the objective and constraint functions make solving the empirical problem of (1.9) NP-hard and numerically unsolvable. To overcome the computational issue, Wang, Fu and Zeng (2018) suggested replacing the indicator functions with appropriate surrogate functions. Let $\phi(x)$ still denote the hinge loss function and let $\psi(x,\eta)$ denote the shifted ramp loss function (Huang, Shi and Suykens, 2014) defined as

$$\psi(x,\eta)=\begin{cases}0, & \text{if } x\leq 0\\ \frac{x+\eta}{\eta}, & \text{if } x\in(0,1)\\ 1, & \text{if } x\geq 1,\end{cases}$$

associated with shifting parameter $\eta\in(0,1]$. Then, we consider the new surrogate problem

$$\begin{aligned}&\arg\min_{f\in\mathcal{F}} E\left[Y\frac{\phi(Af(H))}{p(A|H)}\right]\\ &\text{subject to } E\left[R\frac{\psi(Af(H),\eta)}{p(A|H)}\right]\leq\tau.\end{aligned} \tag{1.10}$$

In (1.10), the hinge loss is a typical choice of surrogate function for 0-1 loss similar to O-learning, while shifted ramp loss $\psi(x,\eta)$ can be viewed as a smooth upper approximation of the indicator function in the constraint, which will converge to $\mathbb{I}(x\geq 0)$ as η goes to 0. Empirically, an estimated rule can be obtained by solving

$$\begin{aligned}&\arg\min_{f\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} Y_i\frac{\phi(A_if(H_i))}{p(A_i|H_i)}+\lambda_n\|f\|_{\mathcal{G}}^2\\ &\text{subject to } \frac{1}{n}\sum_{i=1}^{n} R_i\frac{\psi(A_if(H_i),\eta)}{p(A_i|H_i)}\leq\tau,\end{aligned} \tag{1.11}$$

for sufficient small η. Again, the additional term $\lambda_n\|f\|_{\mathcal{G}}^2$ is a typical choice of regularization term to reduce overfitting. When $\mathcal{G}$ is a RKHS, Wang, Fu and Zeng (2018) shows that the optimization problem can be solved efficiently via the difference of convex functions (DC) algorithm (Tao and An, 1997) and the

optimization problem in each iteration can be reduced to a standard quadratic optimization problem. Recently, this framework was further studied when outcome variable Y is binary in Wang, Zhao and Zheng (2020) and when the constraint is specified by the quantile of the Y in Fang, Wang and Wang (2022). A similar framework was also studied in Pan et al. (2021) and a cross-validation algorithm was proposed to find the optimal rule that satisfies the constraint.

For DTRs with constraint, the same challenge will occur when extending the single-stage method to the multistage problem with $T \geq 2$. Like standard DTR, the optimal decision rule estimated by conditioning on past trajectories may lead to suboptimal rules due to the delayed treatment effect toward the beneficial reward outcome. Moreover, the treatment may also yield delayed treatment effect upon the adverse risk leading the optimal rules estimated stagewisely infeasible when risk restrictions are imposed. Hence, methods designed for the single-stage problem are hard to extend to the problem with more than 2 stages when risk needs to be considered unless further causal assumptions are imposed. Very few methods have been proposed to tackle the multistage DTRs with additional restrictions. Among existing methods, Laber et al. (2018) studied the multistage optimal dosage regimen problem with additional restrictions over safety. A policy iteration-based algorithm was proposed to estimate the optimal DTRs under constraint via modeling a series of Q-function. Illenberger, Spieker and Mitra (2021) considered the same problem from a benefit-cost perspective and proposed another Q-learning-based approach with the additional restriction that the optimal decision rules are from the class of decision tree. However, both two methods rely on accurately modeling the Q-function and are computationally intense without theoretical justification to guarantee that the estimated rules will be optimal under the constraint.

1.4 Contributions and outline

As discussed early, controlling the adverse impact of treatments is one of the main goals that need to be achieved in personalized medicine and few methods have ever been proposed to address this problem. In this work, we contribute to the study of personalized medicine by proposing new DTRs with consideration of additional risk constraint(s). Our work is mainly motivated by the treatment of chronic diseases where treatments that lead to a higher beneficial reward will also cause adverse impacts on patients and need to be avoided during treatment design.

The remaining work is organized as follows:

- In Chapter 2, we propose a new DTRs method, namely benefit-risk DTRs (BR-DTRs), to tackle the DTRs problem where the primary goal is to maximize certain patient's cumulative reward but during each stage, the acute short-term risk induced by the proposed treatment rules should also be controlled lower than prespecified risk restrictions to ensure safety. Numerically, we show that the multistage multi-constraints DTRs problem can be decomposed into a series of single-stage single-constraint optimization problems, which can be efficiently solved using optimization algorithms developed in Wang, Fu and Zeng (2018).

- In Chapter 3, we develop a general framework to handle one type of DTRs problem, namely the cumulative benefit-risk (CBR) problem, where the primary goal is still to find the optimal treatment rules that maximize certain patient's cumulative beneficial reward, but we also require that the proposed treatment rules should not induce a cumulative risk exceeding the prespecified risk restriction. Numerically, we propose a general estimation procedure that will convert the estimation of the constrained DTRs problem into a series of standard unconstrained DTRs problems. To substantiate the estimation, we present how Q-learning and O-learning can be utilized along with the proposed procedure to solve a concrete CBR problem. In addition, we propose a novel estimation framework, namely multistage ramp loss (MRL) learning, to solve each unconstrained DTRs problem with decision rules being jointly estimated across all stages.

- In Chapter 4, we develop a new machine learning-based method, namely L1-MRL, by extending and incorporating an additional L_1-penalty to MRL framework to implement variable selection in addition to learning optimal DTRs. Because of the simultaneous property of MRL, the new method is able to impose cross-stage penalization over the estimated decision rules and a DC algorithm is developed to estimate the optimal rules efficiently.

CHAPTER 2: LEARNING OPTIMAL DYNAMIC TREATMENT REGIMENS SUBJECT TO STAGEWISE ACUTE RISK CONTROLS

2.1 Introduction

As discussed in Chapter 1, existing DTRs method often formalize the problem as learning the optimal decision rules that solely maximize patients' certain beneficial reward outcome. However, for many chronic diseases, treatments are usually multifaceted: the aggressive treatment with a better reward is often accompanied by higher toxicity, leading to the elevated risk of severe and acute side effects or even fatality. For example, the Standards of Medical Care in Diabetes published by the American Diabetes Association (ADA) suggests metformin as first-line initial therapy for all general T2D patients. Intensified insulin therapy should be applied to patients when the patients' A1C level is above the target (American Diabetes Association, 2022*b*). However, according to the UK Prospective Diabetes study, evidence has indicated that many patients who may eventually rely on insulin therapy to achieve ideal A1C level will be likely to experience more hypoglycemic episodes (UKPDS Group, 1998), and the latter can cause neurological impairments, coma, or death (Cryer, Davis and Shamoon, 2003). Another example is corticosteroid therapy adopted by patients with asthma, rheumatoid arthritis, or other immune system disorders. Corticosteroid helps patients to relieve the symptom but will also increase the risk of complications in the short term if patients have another disease via inhibiting patients' immune system (Buchman, 2001; Liu et al., 2013). Therefore, the benefit-risk challenge presented in these chronic diseases entails that the ideal treatment rules should also take into consideration to reduce any short-term risks while maximizing the long-term rewarding outcome.

To respond to the real demand for the treatment of chronic disease, in this chapter, we consider the problem of learning the optimal DTRs in a multistage study, subject to different risk constraints at each stage. Our motivation is to learn the treatment strategy for T2D patients such that the strategy can best control the HbA1c level in the long run but also ensure that the number of adverse events related to metabolic health is controlled. We propose a general framework, namely benefit-risk DTRs (BR-DTRs), by extending the framework developed in Wang, Fu and Zeng (2018) from the single stage to the multiple stages. Specifically, we propose a backward procedure to estimate the optimal treatment rules: at each stage, we maximize the

expected value function under the risk constraint imposed at the current stage. The solution can be obtained by solving a constrained support vector machine problem. Theoretically, we show that the resulting DTRs are Fisher consistent when some proper surrogate functions are used to replace the risk constraints. We further derive the non-asymptotic error bounds for the cumulative reward and stagewise risks associated with the estimated DTRs.

Our contributions are two-fold: first, we propose a general framework to estimate the optimal DTRs under the stagewise risk constraints; the proposed framework reduces to the O-learning for DTRs in Zhao et al. (2015) when there is no risk constraint and reduces to the method in Wang, Fu and Zeng (2018) when there is only one stage. Second, our work establishes the non-asymptotic results for the estimated DTRs for both value and risk functions and such results have never been given before. In particular, we show that support vector machines still yield Fisher consistent treatment rules under a range of risk constraints. Our theory also shows that the convergence rate of the predicted value function is in an order of the cubic root of the sample size and the convergence rate for the risk control has an order of the square root of the sample size.

The remaining chapter is organized as follows. In Section 2.2, we discuss the statistical framework of BR-DTRs and present the complete BR-DTRs algorithm. In Section 2.3, we provide further theoretical justification for BR-DRTs. We demonstrate the performance of BR-DTRs via simulation studies in Section 2.4 and apply the method to a real study of T2D patients in Section 2.5. We discuss the contribution, limitation and future study topics in Section 2.6. The detailed derivation of the DC algorithm is presented in Section 2.7 and the proofs are presented in Section 2.8 and 2.9.

2.2 Method

2.2.1 DTRs under stagewise risk constraints

Consider a T-stage DTRs problem and we use $(Y_1, ..., Y_T)$ to denote the beneficial reward and $(R_1, ..., R_T)$ to denote the risk outcomes at each stage. We assume that $\{(Y_t, R_t)\}_{t=1}^T$ are bounded random variables and a series of dichotomous treatment options are available at each stage, denoted by $A_t \in \{-1, +1\}$. Let $H_1 \subset \cdots \subset H_T$ be the feature variables at stage t, which includes the baseline prognostic variables, intermediate outcomes and any time-dependent covariates information prior to stage t and recall that DTRs

are defined as the sequence of functions

$$\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_T : \mathcal{H}_1 \times \cdots \times \mathcal{H}_T \to \{-1, +1\}^T \text{ where } \mathcal{D}_t : \mathcal{H}_t \mapsto \{-1, +1\}.$$

The goal of BR-DTRs is to find the optimal rule $\mathcal{D}^*$ that maximizes the cumulative reward at the final stage T, while the risk at each stage t is controlled by a pre-specified risk constraint, denoted by τ_t. Mathematically, we aim to solve the following optimization problem

$$\begin{aligned} \max_{\mathcal{D}} \quad & E^{\mathcal{D}}[\sum_{t=1}^{T} Y_t] \\ \text{subject to} \quad & E^{\mathcal{D}}[R_1] \leq \tau_1, \ldots, E^{\mathcal{D}}[R_T] \leq \tau_T, \end{aligned}$$

where $E^{\mathcal{D}}[\cdot]$ still denotes the expectation given $A_t = \mathcal{D}_t(H_t)$ for $t = 1, .., T$.

Analogous to standard DTRs methods reviewed in Section 1.2, additional assumptions are necessary to ensure that the above problem can be solved using the observed data. To this end, we again let $\bar{A}_t = (A_1, .., A_t)$ denote the observed treatment history and $\bar{a}_t = (a_1, ..., a_t) \in \{-1, +1\}^n$ denote any fixed treatment history up to time t, and use $X(\bar{a}_t)$ to denote the potential outcome of variable X under treatment $\bar{a}_t$.

Assumption 2.1 *(Stable Unit Treatment Value (SUTV)) At each stage, the subject's outcomes are not influenced by other subjects' treatments allocation, i.e.,*

$$(Y_t, R_t) = (Y_t(\bar{a}_t), R_t(\bar{a}_t)) \quad \textit{given} \quad \bar{A}_t = \bar{a}_t.$$

Assumption 2.2 *(No Unmeasured Confounders (NUC)) For any $t = 1, ..., T$*

$$A_t \perp\!\!\!\perp (H_{t+1}(\bar{a}_t), ..., H_T(\bar{a}_{T-1}), Y_T(\bar{a}_T), R_T(\bar{a}_T)) \big| H_t.$$

Assumption 2.3 *(Positivity) For any $t = 1, ..., T$, there exists universal constants $0 < c_1 \leq c_2 < 1$ such that*

$$c_1 \leq p(A_t = 1 | H_t) \leq c_2$$

holds for H_t almost surely.

Assumption 2.4 *(Acute Risk) For any $t = 1, ..., T$ and $\bar{a}_t \in \{-1, 1\}^t$, $R_t(\bar{a}_t)$ only depends on a_t. Thus, we can write $R_t(a_t)$ for this potential outcome.*

The first two assumptions are corresponding SUTV and NUC assumption in the context of the DTRs problem with multiple stagewise risk constraints and we repeat the positivity assumption in the standard DTRs problem as Assumption 2.3. In particular, the NUC and positivity assumption will still hold when data is collected from a simple randomized trial or SMART. Assumption 2.4 captures the acute risk property of chronic diseases. That is, for the same individual, the adverse risk in each stage is caused by his/her most recent treatment. As an additional note, we can further assume that R_t is positive and bounded away from zero after shifting both R_t and τ_t by one same constant without changing the problem of interest.

Under all four additional assumptions and suppose $\mathcal{D}_t(H_t) = \text{sign}(f_t(H_t))$ for some measurable decision function f_t, we note that

$$E^{\mathcal{D}}[R_t] = E\left[\frac{R_t \prod_{t=1}^{T} \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] = E\Big[R_t(\text{sign}(f_1), ..., \text{sign}(f_t))\Big]$$

$$= E\Big[R_t(\text{sign}(f_t))\Big] = E\left[\frac{R_t I(A_t f_t(H_t) > 0)}{p(A_t|H_t)}\right].$$

Then according to Qian and Murphy (2011), the original problem can be reformulated as

$$\begin{aligned} &\max_{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\mathcal{F}_T} E\left[\frac{(\sum_{t=1}^{T} Y_t)\prod_{t=1}^{T}\mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \\ &\text{subject to } \; E\left[\frac{R_t \mathbb{I}(A_t f_t(H_t) > 0)}{p(A_t|H_t)}\right] \le \tau_t, \qquad t = 1, ..., T, \end{aligned} \tag{2.1}$$

where $\mathcal{F}_t$ denotes the set of all real value measurable functions from $\mathcal{H}_t \to \mathbb{R}$. Intuitively, following the same arguments as in Zhao et al. (2015), the solution of the above problem can be solved in a backward fashion as follows: let $\{\mathcal{O}_t\}_{t=1}^{T}$ denote the feasible region of the original problem under risk constraints $(\tau_1, ..., \tau_T)$ at stage t, i.e.,

$$\mathcal{O}_t = \left\{ f \in \mathcal{F}_t \middle| E\left[\frac{R_t \mathbb{I}(A_t f(H_t) > 0)}{p(A_t|H_t)}\right] \le \tau_t \right\}, \quad t = 1, ..., T,$$

and define the U-function as

$$U_t(h_t; g_t, g_{t+1}, ..., g_T) = E\left[\frac{(\sum_{j=t}^{T} Y_t)\prod_{j=t}^{T}\mathbb{I}(A_j g_j(H_j) > 0)}{\prod_{j=t}^{T} p(A_j|H_j)} \middle| H_t = h_t\right],$$

where we set $U_{T+1} = 0$. Then the optimal solution to (2.1), denoted by $(g_1^*, ..., g_T^*)$, satisfies

$$g_t^* = \arg\min_{f_t \in \mathcal{O}_t} E\left[\frac{(Y_t + U_{t+1}(H_{t+1}; g_{t+1}^*, ..., g_T^*))\mathbb{I}(A_t f_t(H_t) < 0)}{p(A_t|H_t)}\right]. \tag{2.2}$$

In fact, our later proof for Theorem 2.1 will show that such a backward algorithm leads to the optimal DTRs solving problem (2.1).

2.2.2 Surrogate loss and Fisher consistency

One main difficulty of implementing the framework (2.2) is the existence of the indicator functions in both the objective function and risk constraints, which makes solving the original problem NP-hard. Following the idea in Wang, Fu and Zeng (2018), we propose to replace the indicator function in the objective function by hinge loss function $\phi(\cdot)$ defined as $\phi(x) = (1 - x)_+$, and replace the indicator function in the risk constraint by shifted ramp loss function given by

$$\psi(x, \eta) = \begin{cases} 1, & \text{if } x \geq 0 \\ \frac{x+\eta}{\eta}, & \text{if } x \in (-\eta, 0) \\ 0, & \text{if } x \leq -\eta, \end{cases}$$

where $\eta \in (0, 1]$ is a prespecified shifting parameter that can vary with stage. We then consider the following surrogate problem, namely the BR-DTRs problem,

$$f_t^* = \arg\min_{f_t \in \mathcal{A}_t} E\left[\frac{(Y_t + U_{t+1}(H_{t+1}; f_{t+1}^*, ..., f_T^*))\phi(A_t f_t(H_t))}{p(A_t|H_t)}\right], \tag{2.3}$$

where

$$\mathcal{A}_t = \left\{ f \in \mathcal{F}_t \middle| E\left[\frac{R_t \psi(A_t f(H_t), \eta_t)}{p(A_t|H_t)}\right] \leq \tau_t \right\}$$

from stage $t \in= \{1, ..., T\}$. Equivalently, we replace the 0-1 loss function in the objective function with the hinge loss and replace the indicator function in the risk constraint with the shifted ramp loss function. The shifted ramp loss leads to a smooth and conservative approximation of the risk constraint function when η_t is small.

Our next result shows that the new surrogate problem leads to the DTRs that are Fisher consistent. Before stating the theorem, we define a t-stage pseudo-outcome Q_t as

$$Q_t = Y_t + U_{t+1}(H_{t+1}; f_{t+1}^*, ..., f_T^*),$$

which is the cumulative reward from stage t to T assuming that all treatments have been optimized from stage $t+1$ to T. For any random vector (Y, R, A, H) and for $a = \pm 1$, we use the following notations:

$$m_Y(h,a) = E[Y|H = h, A = a], \qquad \delta_Y(h) = m_Y(h,1) - m_Y(h,-1),$$
$$m_R(h,a) = E[R|H = h, A = a], \qquad \delta_R(h) = m_R(h,1) - m_R(h,-1).$$

Let

$$\tau_{t,\min} = E\left[R_t \frac{\mathbb{I}(A_t \delta_{R_t}(H_t) < 0)}{p(A_t|H_t)}\right],$$
$$\tau_{t,\max} = E\left[R_t \frac{\mathbb{I}(A_t \delta_{Q_t}(H_t) > 0)}{p(A_t|H_t)}\right].$$

In other words, $\tau_{t,\min}$ is the risk under the decision function given by $-\delta_{R_t}(H_t)$, which is the one maximizing the risk regardless of the reward outcome. Thus, $\tau_{t,\min}$ is the minimum risk that one can possibly achieve at stage t. While, $\tau_{t,\max}$ is the risk for the decision function given by $\delta_{Q_t}(H_t)$, which is the one maximizing the reward regardless of the risk. Thus, $\tau_{\max,t}$ is the maximal risk.

Theorem 2.1 *For $t = 1, .., T$ and any fixed $\tau_{t,\min} < \tau_t < \tau_{t,\max}$, suppose that $P(\delta_{Q_t}(H_t)\delta_{R_t}(H_t) = 0) = 0$ and random variable $\delta_{Q_t}(H_t)/\delta_{R_t}(H_t)$ has the distribution function with a continuous density function in the support of H_t. Then for any $\eta_t \in (0, 1]$ and $t = 1, ..., T$, we have sign(f_t^*) = sign(g_t^*) almost surely, and $(f_1^*, ..., f_T^*)$ solves the optimization problem in (2.1).*

When $\tau_t \geq \tau_{t,\max}$, the BR-DTRs problem reduces to a standard DTRs problem and Zhao et al. (2015) shows that the Fisher consistency holds without additional conditions. For $T = 1$, the conditions are similar to Wang, Fu and Zeng (2018), but they assume H_t to have a continuous distribution. Theorem 2.1 basically indicates that when the risk constraints are feasible and assume that the reward difference between two treatments varies continuously with respect to the risk and risk difference, using the surrogate loss leads to the true optimal DTRs for any shifting parameter $\eta_t \in (0, 1]$. We note that this result has never been established before. The complete proof is presented in Section 2.8.

2.2.3 Estimating BR-DTRs using empirical data

Given data $\{(H_{i1}, A_{i1}, Y_{i1}, R_{i1}, ..., H_{iT}, A_{iT}, Y_{iT}, R_{iT})\}_{i=1}^n$ from n i.i.d. patients, we propose to solve the empirical version of the surrogate problem to estimate the optimal DTRs: let

$$\mathcal{A}_{t,n} = \left\{ f \in \mathcal{G}_t \middle| \frac{1}{n} \sum_{i=1}^{n} \frac{R_{it}\psi(A_{it}f(H_{it}), \eta)}{p(A_{it}|H_{it})} \leq \tau_t \right\},$$

then we solve

$$\widehat{f}_t = \arg\max_{f \in \mathcal{A}_{t,n}} \frac{1}{n} \sum_{i=1}^{n} \frac{(\sum_{s=t}^{T} Y_{is}) \prod_{s=t+1}^{T} \mathbb{I}(A_{is}\widehat{f}_s(H_{is}) > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} \phi(A_{it}f(H_{it})) + \lambda_{n,t}\|f\|_{\mathcal{G}_t}^2 \tag{2.4}$$

for $t = T, ..., 1$ in turn. Here, $\|\cdot\|_{\mathcal{G}_t}$ denotes the functional norm associated with functional space $\mathcal{G}_t \subset \mathcal{F}_t$. Again, the last term $\lambda_{n,t}\|f\|_{\mathcal{G}_t}^2$ is a typical choice of penalty term which regularizes the complexity of the estimated optimal decision function to avoid overfitting. Common choices of $\mathcal{G}_t$ include RKHS under a linear kernel where $k(h_i, h_j) = h_i^T h_j$, or a Gaussian radial basis kernel with $k(h_i, h_j) = \exp(-\sigma^2\|h_i - h_j\|^2)$, where σ denotes the inverse of the bandwidth.

To improve the numerical performance of BR-DTRs, we further adopt the augmentation technique used in the AOWL proposed by Liu et al. (2018) introduced in Section 1.2. Specifically, we replace the weights in the objective function and treatment variable respectively by

$$\widehat{Y}_{it} = |Y_{it} + \widehat{Q}_{i,t+1} - \widehat{\mu}_t(H_{it})|, \quad \widehat{A}_{it} = A_{it}\text{sign}(Y_{it} + \widehat{Q}_{i,t+1} - \widehat{\mu}_t(H_{it})). \tag{2.5}$$

Here, $\widehat{Q}_{i,t}$ is the augmented Q-function defined as

$$\begin{aligned} \widehat{Q}_{i,t+1} =& \frac{(\sum_{s=t+1}^{T} Y_{is}) \prod_{s=t+1}^{T} \mathbb{I}(A_{is}\widehat{f}_s(H_{is}) > 0)}{\prod_{s=t+1}^{T} p(A_{is}|H_{is})} \\ &- \sum_{j=t+1}^{T} \left\{ \frac{\prod_{s=t+1}^{j-1} \mathbb{I}(A_{is}\widehat{f}_s(H_{is}) > 0)}{\prod_{s=t+1}^{j-1} p(A_{is}|H_{is})} \left[\frac{\mathbb{I}(A_{ij}\widehat{f}_j(H_{ij}) > 0)}{p(A_{ij}|A_{ij})} - 1 \right] \widehat{\mu}_{t+1,j}(H_{ij}) \right\}, \end{aligned} \tag{2.6}$$

and let $\widehat{Q}_{i,T+1} = 0$. Additionally, $\widehat{\mu}_{t,j}$ is the estimated mean function from solving the weighted least square problem

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_{is}\widehat{f}_s(H_{is}) > 0)}{\prod_{s=t+1}^{T} p(A_{is}|H_{is})} \frac{1 - p(A_{ij}|H_{ij})}{\prod_{s=t+1}^{j} p(A_{is}|H_{is})} (\sum_{s=t+1}^{T} Y_{is} - \mu_{t,j}(H_{ij}))^2, \tag{2.7}$$

Algorithm 1 BR-DTRs

Input: Given training data $(Y_{it}, R_{it}, A_{it}, H_{it})$ and $(\lambda_t, \mathcal{G}_t, \tau_t, \eta)$ for $i = 1, ..., n$ and $t = 1, ..., T$
for $t = T$ to 1 **do**
  for $j = t + 1$ to T **do**
    obtain estimator $\widehat{\mu}_{t,j}$ via minimizing (2.7)
  end for
  if $t = T$ **then** define $\widehat{Q}_{i,T+1} = 0$
  else compute $\widehat{Q}_{i,t+1}$ from (2.6)
  end if
  compute $\widehat{\mu}_t(H_{ij})$ via least square estimator and obtain $\{(\widehat{Y}_{it}, \widehat{A}_{it})\}_{i=1}^n$ via (2.5)
  obtain $\widehat{f}_t$ by solving

$$\begin{aligned} \min_{f \in \mathcal{G}_t} \quad & \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{Y}_{it}}{p(A_{it}|H_{it})} \phi(\widehat{A}_{it} f(H_{it})) + \lambda_{n,t} \|f\|_{\mathcal{G}_t}^2 \\ \text{subject to} \quad & \frac{1}{n} \sum_{i=1}^{n} \frac{R_{it}}{p(A_{it}|H_{it})} \psi(A_{it} f(H_{it}), \eta) \leq \tau_t \end{aligned}$$

  using DC algorithm
end for
Output: $(\widehat{f}_1, ..., \widehat{f}_T)$

and $\widehat{\mu}_t(H_{ij})$ is obtained by minimizing $\sum_{i=1}^n (Y_{it} + \widehat{Q}_{i,t+1} - \mu_t(H_{it}))^2$. The main motivation for this construction was discussed in Section 1.2.2, where Liu et al. (2018) showed that the removal of the main effect, $\widehat{\mu}_t(H_{it})$, could reduce the weight variability without affecting the treatment rule estimation, and that using the augmentation term in constructing $\widehat{Q}_{i,t+1}$ could use the information from all subjects, leading to more efficient estimation for DTRs.

Hence, we propose a backward procedure to estimate the optimal DTRs. First, we solve a single-stage problem using data at stage $t = T$, and then in turn, for $t = T-1, ..., 1$, we solve the constrained optimization problem after plugging in $(\widehat{f}_{t+1}, ..., \widehat{f}_T)$. For the optimization at each stage, we apply the difference of convex functions (DC) algorithm (Tao and An, 1997) which is an iterative process and in each iteration, the update can be reduced to a standard quadratic programming problem. Details of the DC algorithm are provided in Section 2.7.

2.3 Theoretical Results

In this section, we establish the non-asymptotic error rate of the value function and stagewise risks under the estimated decision functions $(\widehat{f}_1, ..., \widehat{f}_T)$. More specifically, for any arbitrary decision functions

$(g_1, ..., g_t)$, the value function of $(g_1, ..., g_T)$ is defined as

$$\mathcal{V}(g_1, ..., g_T) = E\bigg[\frac{(\sum_{t=1}^T Y_t)\prod_{t=1}^T \mathbb{I}(A_t g_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg].$$

We aim at obtaining the non-asymptotic bound for the regret function given by

$$\mathcal{V}(f_1^*, ..., f_T^*) - \mathcal{V}(\widehat{f}_1, ..., \widehat{f}_T)$$

and the stagewise risk difference is given by

$$E\bigg[\frac{R_t \mathbb{I}(A_t \widehat{f}_t(H_t) > 0)}{p(A_t|H_t)}\bigg] - \tau_t,$$

for $t = 1, ..., T$.

We assume that $\{\mathcal{G}_t\}_{t=1}^T$ are the RKHS generated by the Gaussian radial basis kernel, i.e., $\mathcal{G}_t = \mathcal{G}(\sigma_{n,t})$ where $\mathcal{G}(\sigma)$ denotes the Gaussian RKHS associated with bandwidth σ^{-1}. Furthermore, for random variable Q_t, R_t, A_t and H_t, we define for $a, b \in \{-1, 1\}$,

$$H_{a,b} = \bigg\{h \in \mathcal{H} : a\delta_{Q_t}(h) > 0, b f_{t,\tau_t}^*(h) > 0\bigg\}$$

and $\Delta_{t,\tau'}(h) = \sum_{a,b\in\{-1,1\}} \text{dist}(h, \mathcal{H}/H_{a,b}) I(h \in H_{a,b})$, where dist$(\cdot)$ denotes the Euclidean distance and $f_{t,\tau'}^*$ denotes optimal solution of (2.3) at stage t but replace the risk constraint in $\mathcal{A}_t$ by τ'. Recall that

$$Q_t = Y_t + U_{t+1}(H_{t+1}; f_{t+1}^*, ..., f_T^*) = Y_t + U_{t+1}(H_{t+1}; f_{t+1}^*, ..., f_T^*).$$

We assume

Assumption 2.5 *Let P_t denote the joint probability H_t. For given $(\tau_1, ..., \tau_T)$ and any $t = 1, .., T$, there exists universal positive constant $\delta_{0,t} > 0$, $K_t > 0$ and $\alpha_t > 0$ such that for any $\tau' \in [\tau_t - 2\delta_{0,t}, \tau_t + 2\delta_{0,t}] \subset (\tau_{t,\min}, \tau_{t,\max})$ we have*

$$\int_{\mathcal{H}_t} \exp\bigg(-\frac{\Delta_{t,\tau'}(h)^2}{s}\bigg) P_t(dh) \leq K_t s^{\alpha_t d_t/2}$$

holds for all $s > 0$.

Assumption 2.5 is an extension of the Geometric Noise Exponent (GNE) assumption proposed by Steinwart and Scovel (2007) to establish a fast convergence risk bound for standard SVM, and later adopted by Zhao et al. (2012) to derive the risk bound for the DTRs without risk constraints. The assumption can be viewed as a regularization condition of the behavior of samples near the true optimal decision boundary. For a fixed τ_t, α_t can be taken to 1 when $\Delta(h)$ has order less or equal to $O(h)$. When the optimal decision boundary is strictly separated, i.e. $\text{dist}(H_{a,b}, H_{a',b'}) > 0$ for any $a \neq a'$ and $b \neq b'$, by using the fact that $\exp(-t) \leq C_s t^{-s}$ one can check that Assumption 2.5 holds for $\alpha_t = \infty$. When the optimal decision boundary is not strictly separated, it can be shown that Assumption 2.5 can still hold for arbitrary $\alpha_t \in (0, \infty)$ when the marginal distribution of H_t has light density near the optimal decision boundary (see Example 2.4 in Steinwart and Scovel (2007)).

The following theorem gives the non-asymptotic error bound for the value loss and risk difference for the estimated DTRs, assuming that all μ_t and $\mu_{t,j}$ in the augmentation are known. The theorem allows stage-wise shifting parameters to vary with sample size, denoted by $(\eta_{n,1}, ..., \eta_{n,T})$.

Theorem 2.2 *Suppose that Assumption 2.1 to 2.5 and conditions in Theorem 2.1 hold, and H_t is defined on a compact set $\mathcal{H}_t \subset \mathbb{R}^{d_t}$ for $t = 1, ..., T$. Let $\{\nu_t\}$ and $\{\theta_t\}$ be two series of positive constants such that $0 < \nu_t < 2$ and $\theta_t > 0$ for all $t = 1, ..., T$. Then for any $n \geq 1$, $\lambda_{n,t} > 0$, $\sigma_{n,t} > 0$ and $0 < \eta_t \leq 1$, such that $\lambda_{n,t} \to 0, \sigma_{n,t} \to \infty$ and that there exit constants C_1, C_2, C_3 satisfying*

$$C_1 \sigma_n^{-\alpha_t d_t} \eta_{n,t}^{-1} \leq \delta_{0,t}, \;\; C_2 n^{-1} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t} \leq 1,$$

and $\delta_t + C_1 \sigma_{n,t}^{-\alpha_t, d_t} \eta_{n,t}^{-1} + C_3 n^{-1/2} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \left(\frac{M}{c_1 \lambda_{n,t}} + \sigma_{n,t}^{d_t} \right)^{\nu_t/4} \eta_{n,t}^{-\nu_t/2} \leq 2\delta_{0,t}$, it holds

$$|\mathcal{V}(\widehat{f}_1, ..., \widehat{f}_T) - \mathcal{V}(f_1^*, ..., f_T^*)| \leq \sum_{t=1}^{T} (c_1/5)^{1-t} C_t \Big(n^{-1/2} \lambda_{n,t}^{-1/2} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} + \lambda_{n,t} \sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \eta_{n,t}^{-1} + \eta_{n,t} \Big)$$

with probability of at least $1 - \sum_{t=1}^{T} h_t(n, \sigma_{n,t})$, where

$$h_t(n, \sigma_{n,t}) = 2 \exp\left(- \frac{2n\delta_{0,t}^2 c_1^2}{M^2} \right) + 2 \exp\left(- \frac{n\delta_t^2 c_1^2}{2M^2} \right) + \exp\big(- \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t} \big).$$

Moreover, with probability at least $1 - h_t(n, \sigma_{n,t})$*, the risk induced by* $\widehat{f}_t$ *satisfies*

$$E\left[\frac{R_t \mathbb{I}(A_t \widehat{f}_t(H_t) > 0)}{p(A_t|H_t)}\right] \leq \tau_t + \delta_t + C_t n^{-1/2} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} \eta_{n,t}^{-\nu/2}.$$

Here, C_t *denotes some constant only depending on* α_t, K_t*,* d_t*,* ν_t*,* θ_t*,* c_1 *and* M*.*

Theorem 2.2 can be established by first verifying the result for $T = 1$ and then extending the result to $T \geq 2$ using an analogous argument of Theorem 3.4 of Zhao et al. (2015). The risk bound of the value function proved in Theorem 2.2 indicates that the error consists of four parts. The first two terms correspond to the stochastic error and systematic error due to using the empirical estimator to approximate the true objective function and restricting the optimal function within a RKHS in empirical the problem (2.4). The third error term $O(\sigma_{n,t}^{-\alpha_t d_t} \eta_{n,t}^{-1})$ is induced by using the empirical estimator as risk constraints in (2.4). The remaining error has order $O(\eta_{n,t})$ and results from the property that the regret under 0-1 loss function is upper bounded by the regret under hinge loss plus an error term of order $O(\eta)$ when we use the shifted ramp loss to approximate the indicator function in constraints. Due to the existence of the last two error terms, the choice of shifting parameter must be small but bounded away from 0 in order to minimize the regret. The proof of Theorem 2.2 and required preliminary lemmas are provided in Section 2.9.

According to Theorem 2.2, the risk bound of the regret is minimized by setting $\eta_{n,t} = \sigma_{n,t}^{-\alpha_t d_t} \eta_{n,t}^{-1}$, $\lambda_{n,t} \sigma_{n,t}^{d_t} = \sigma_{n,t}^{-\alpha_t d_t} \eta_{n,t}^{-1}$ and $\eta_{n,t} = n^{-1/2} \lambda_{n,t}^{-1/2} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}$, which gives

$$\lambda_{n,t} = O\big(\sigma_{n,t}^{-(\alpha_t+2)d_t/2}\big), \quad \eta_t = O\big(\sigma_{n,t}^{-\alpha_t d_t/2}\big)$$

and

$$\sigma_{n,t} = O\big(n^{\frac{1}{\alpha_t d_t + (\alpha_t+2)d_t/2 + (1-\nu_t/2)(1+\theta_t)d_t}}\big).$$

Consequently, there exists constants $k_1, k_2 > 0$ independent of sample size n such that

$$|\mathcal{V}(\widehat{f}_1, ..., \widehat{f}_T) - \mathcal{V}(f_1^*, ..., f_T^*)| \leq k_1 \sum_{t=1}^{T} (c_1/5)^{1-t} n^{-\frac{\alpha_t d_t}{2\alpha_t d_t + (\alpha_t+2)d_t + 2(1-\nu_t/2)(1+\theta_t)d_t}}$$

holds with probability $1 - \sum_{t=1}^{T} \exp\big(-k_2 n^{\frac{(1-\nu_t/2)(1+\theta_t)d_t}{\alpha_t d_t + (\alpha_t+2)d_t/2 + (1-\nu_t/2)(1+\theta_t)d_t}}\big)$. When α_t can be selected arbitrarily large in which case the data are approximately separated near the optimal decision boundary,

the convergence rate of the value function is at most of order $O(n^{-1/3})$. In terms of risks, when α_t can be arbitrarily large and let ν_t go to 0, the risk constraint inequality indicates that the stagewise risk under the estimated rule can always be bounded by τ_t plus an error term of order up to $O(n^{-1/2})$. In terms of stage T, we note that the error bound is increasing exponentially with respect to the total number of stages. This result is similar to the risk bound of value function obtained in Q-learning (Murphy, 2005a) and O-learning (Zhao et al., 2015). As another note, the value function is Lipschitz continuous in terms of the model parameters for μ_t and $\mu_{t,j}$. Thus, when the models of μ_t and $\mu_{t,j}$ are parametric so the parameters are estimated at $O(n^{-1/2})$, the impact on the convergence rate of the regret will be of the same order and will not affect the minimum possible error rate of the value function and the adverse risk under the estimated rule. Similar arguments are given in Chen, Zeng and Wang (2021).

2.4 Simulation Studies

We demonstrate the performance of BR-DTRs via simulation studies in this section. We consider two settings both of which simulate the situation when adopting preferable treatment in the early stage would immensely affect the performance of possible treatments in later stages. Specifically, in both settings, we first generate an 8-dimensional baseline prognostic variable matrix X from independent uniform distribution $U[0,1]$. In the first setting, we consider a two-stage randomized trial where treatments A_1, A_2 are randomly assigned with an equal probability of 0.5. The stage-specific rewards and risks are defined by

$$Y_1 = 1 - X_1 + A_1(-X_1 - X_2 + 1) + \epsilon_{Y_1}, \quad R_1 = 2 + X_1 + A_1(-X_1/2 + X_2 + 1) + \epsilon_{R_1},$$
$$Y_2 = 1 - X_1 + A_2(Y_1 - 3X_1 + A_1 + 1) + \epsilon_{Y_2}, \quad R_2 = 1 + X_1 + A_2(Y_2/2 - X_1 + A_2/2 + 1) + \epsilon_{R_2},$$

where ϵ_{Y_1}, ϵ_{Y_2} are noises of reward outcomes generated from the independent standard normal distribution $N(0,1)$, and ϵ_{R_1}, ϵ_{R_2} are noises of adverse risks generated from the independent uniform distribution $U[-0.5, 0.5]$. In this setting, both Y_1, Y_2, R_1 and R_2 are the linear functions of $H_1 = X$ and $H_2 = (H_1, A_1, Y_1, R_1)$. In the second setting, Y_2 is a nonlinear function of H_2 and is generated according to

$$Y_1 = 1 + A_1(-X_1 - X_2/3 + 1.2) + \epsilon_{Y_1}, \quad R_1 = 1.5 + A_1(-X_1/3 + 1.5) + \epsilon_{R_1},$$
$$Y_2 = 1 + A_2(-X_1^2/2 - X_2^2/2 + 3A_1/2 + 1.5) + \epsilon_{Y_2}, \quad R_2 = 1 + A_2(2A_1 + 2) + \epsilon_{R_2},$$

and $(A_1, A_2, \epsilon_{Y_1}, \epsilon_{Y_2}, \epsilon_{R_1}, \epsilon_{R_2})$ are generated the same way as setting I. Note that for setting II, the optimal decision boundary in stage II is a circle with respect to (X_1, X_2).

For each simulation setting, we implement our proposed method with training data sample size n equal to 200 and 400, and η varies from 0.02 to 0.1 with an increment of 0.02. For the first simulation setting, we repeat the simulation for $\tau_1 = \tau_2 = 1.4$ and 1.5; for the second simulation setting, we repeat the simulation for $\tau_1 = \tau_2 = 1.3$ and 1.4. Both the linear kernel and Gaussian kernel are employed to compare their performance. The tuning parameter $C_n = (2n\lambda_{n,t})^{-1}$ will be selected by a 2-fold cross-validation procedure that maximizes the Lagrange dual function from a pre-specified grid of 2^{-10} to 2^{10}. To alleviate the computational burden, when using the Gaussian kernel we follow the idea of Wu, Zhang and Liu (2010) and fix $\sigma_{n,t}^{-1}$ to be 2median$\{\|H_i - H_j\| : A_i \neq A_j\}$ instead of picking $\sigma_{n,t}$ adaptively according to n and $\lambda_{n,t}$. In our simulations, all feature variables will be re-centered to mean 0 and rescaled into interval $[-1, 1]$. When solving the optimization problem, we choose the initial values for parameters either uniformly in a bounded interval or using the estimated parameters from the unconstrained problem. We recommend the latter approach as the performance is overall better than picking the initial point randomly. All quadratic programming programs in the DC procedure will be solved by R function *solve.QP()* from *quadprog* package (`https://cran.r-project.org/web/packages/quadprog/index.html`). As a comparison, we also implement the AOWL method proposed by Liu et al. (2018) as implemented in package *DTRlearn2* (`https://cran.r-project.org/web/packages/DTRlearn2/index.html`), which ignores the risk constraints. In addition, we also compare our method with the naive approach where in stage I, we simply use $Y_1 + Y_2$ as the outcome for estimation without adjusting for any delayed treatment effects even though the risk constraints are considered. To assess the performance of each method, we calculate the stage optimal estimated reward and risk on an independent testing dataset of size $N = 2 \times 10^4$. We repeat the analysis with 600 replicates.

Figure 2.1 displays the estimated reward and risk on the independent testing data for the first simulation setting under the different choices of training sample size, kernel basis and shifting parameter η for $\tau_1 = \tau_2 = 1.4$. From the plot, we notice that for the simple linear setting, under both linear and Gaussian kernel the median values of estimated reward/risk will be close to the theoretical reward/pre-specified risk constraints. This indicates that the proposed method can successfully maximize the reward while controlling the risks across both stages. In this setting, compared with the linear kernel, using the Gaussian kernel will significantly underestimate the risk on training data, leading to somewhat exceeding risk on the testing data. Also as expected, in this setting increasing sample size would improve the performance under both kernel choices.

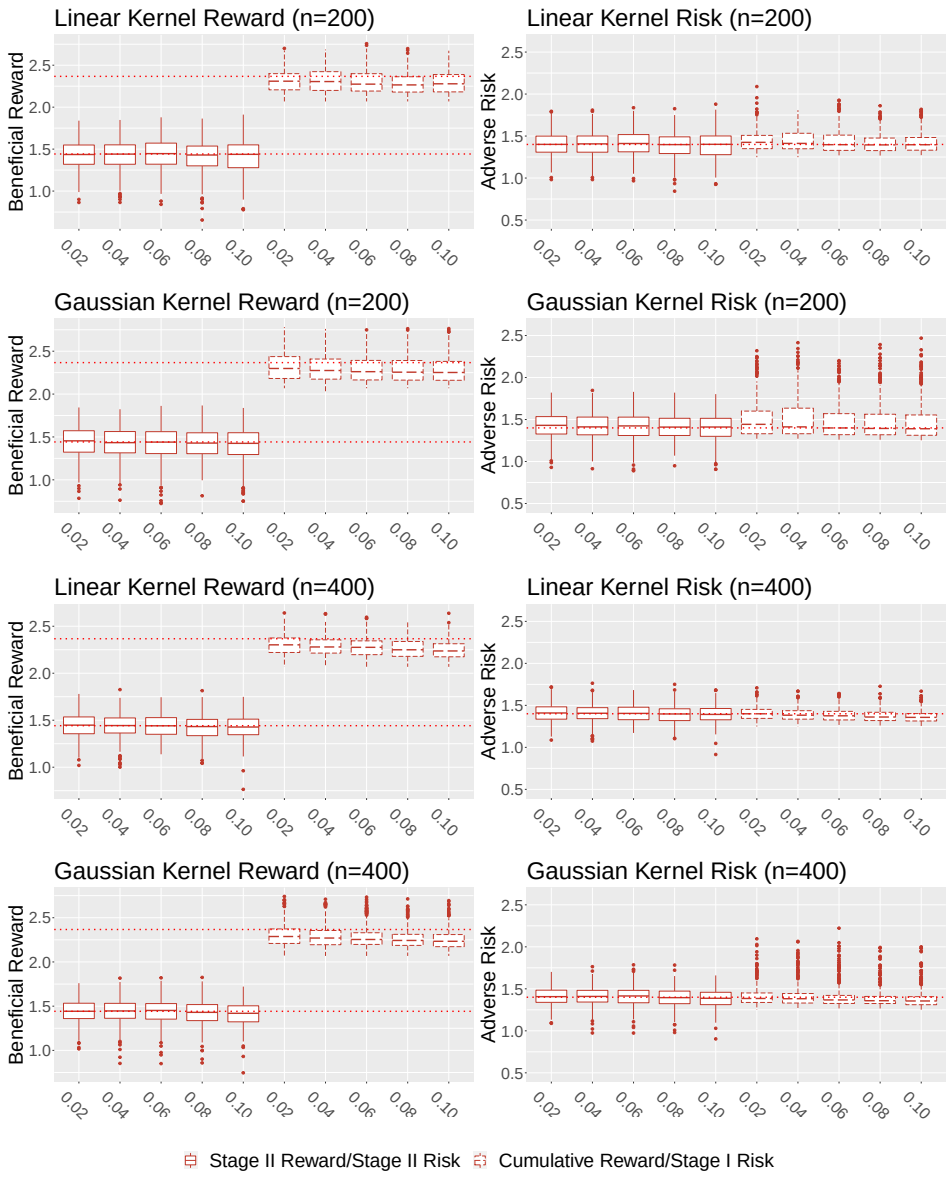

Figure 2.1: Estimated reward/risk on independent testing data set for simulation setting I, training sample size $n = \{200, 400\}$ and $\eta = \{0.02, 0.04, ..., 0.1\}$ (x-axis) under linear kernel or Gaussian kernel. The dashed line in reward plots refers to the theoretical optimal reward under given constraints. The dashed line in risk plots represents the risk constraint $\tau = 1.5$.

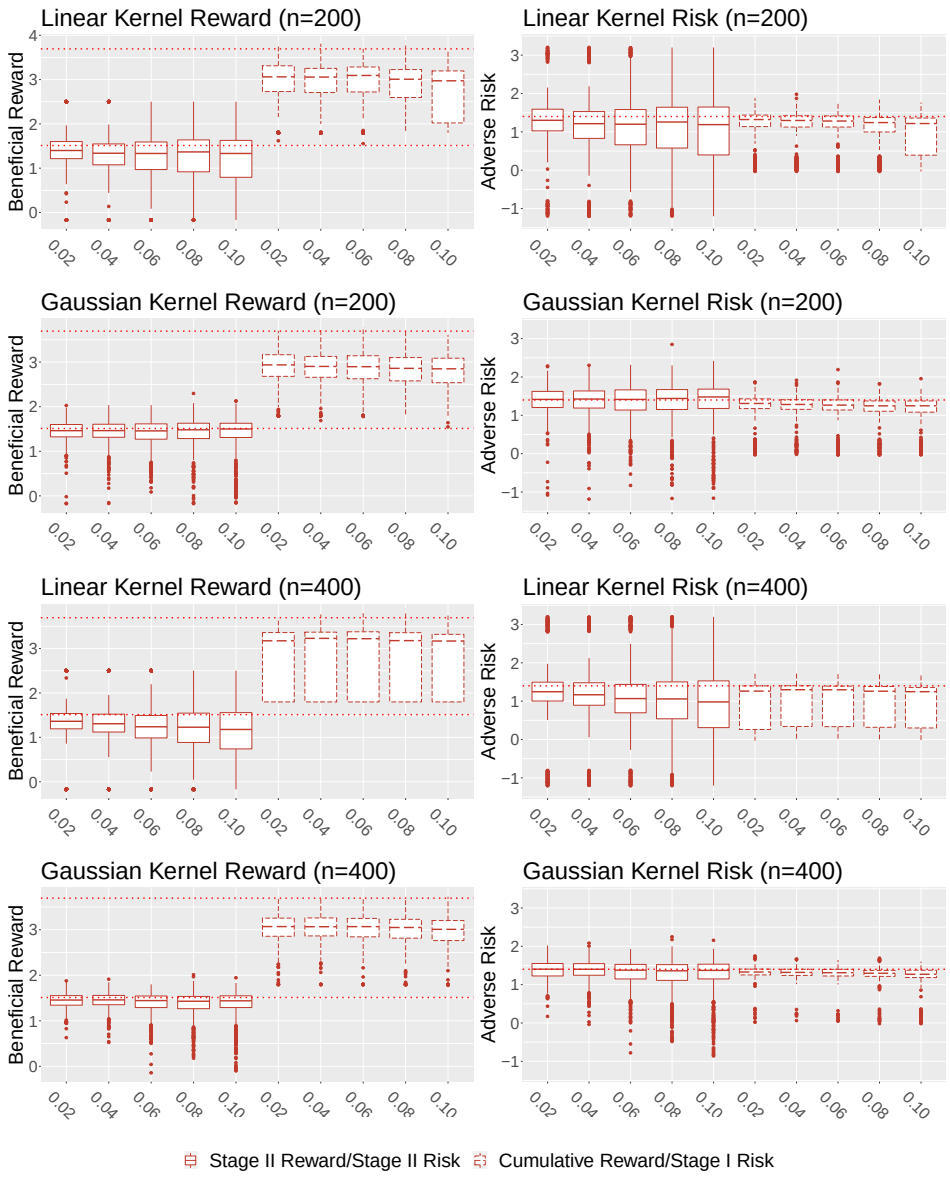

Figure 2.2: Estimated reward/risk on independent testing data set for simulation setting II, training sample size $n = \{200, 400\}$ and $\eta = \{0.02, 0.04, ..., 0.1\}$ (x-axis) under linear kernel or Gaussian kernel. The dashed line in reward plots refers to the theoretical optimal reward under given constraints. The dashed line in risk plots represents the risk constraint $\tau = 1.4$.

In terms of the shifting parameter η, in setting I there is no obvious preference of choosing a small value to a large value. The result from the second nonlinear simulation setting under $\tau_1 = \tau_2 = 1.4$ is presented in Figure 2.2. Under this more complicated setting and when both two stages' optimal decision boundaries are nonlinear, we notice that our method still yields a value close to the truth and the risks are reasonably controlled in both stages. The Gaussian kernel outperformed the linear kernel in both stages since using the linear kernel will misspecify the true model. When the sample size increased, the performance for the Gaussian kernel improved but it was not necessary for the linear kernel, likely due to the misspecification. We also observe that under the second simulation setting and when the Gaussian kernel is used, choosing a small shifting parameter η will achieve better performance on the testing data with much smaller variability. The results for $\tau = 1.5$ for setting I and $\tau = 1.3$ for setting II are similar to $\tau = 1.4$ already discussed. The same conclusions can be made and the results are presented in Section 2.10.

Finally, the results in Table 2.1 compare the performance of BR-DTRs to AOWL, which ignores the risk constraints, and the naive method, which considers the risk constraints but uses the immediate outcomes as the reward. Clearly, even though AOWL always gives a higher reward than BR-DTRs, the corresponding risks of applying the estimated treatment rules are much larger than the ones from BR-DTRs. In contrast, BR-DTRs can always give valid decision rules with risks close to pre-specified threshold values. When compared with the naive method, due to the nature of DTRs, the reward of the BR-DTRs method is always higher than the naive method.

2.5 Application to DURABLE Trial

We apply BR-DTRs to analyze the data from the DURABLE study (Fahrbach et al., 2008). The DURABLE study is a two-phase trial designed to compare the safety and efficacy of insulin glargine versus insulin lispro mix in addition to oral antihyperglycemic agents in T2D patients. During the first phase trial, patients were randomly assigned to the daily insulin glargine group or twice daily insulin lispro mix 75/25 (LMx2) group for 24 weeks. By the end of 24 weeks, patients who failed to reach an HbA1c level lower than 7.0% would enter the second phase intensification study and be randomly reassigned with either the basal-bolus therapy (BBT) or LMx2 for insulin glargine group, or basal-bolus therapy (BBT) or three times daily insulin lispro mix 50/50 (MMx3) therapy for LMx2 group. Any other patients who reached HbA1c 7.0% or lower would enter the maintenance study and keep the initial therapy for another 2 years.

Table 2.1: Estimated reward/risk on independent testing data for $\tau_1 = \tau_2 = 1.4$ and $n = 400$ under 3 different methods using linear/Gaussian kernel.

			Linear Kernel				Gaussian Kernel			
Setting	η	Method	Reward - II	Risk - II	Cumulative Reward	Risk - I	Reward - II	Risk - II	Cumulative Reward	Risk - I
Setting I	0.02	BR-DTRs	1.449(0.086)[1]	1.410(0.072)	2.306(0.077)	1.400(0.053)	1.447(0.087)	1.411(0.072)	2.287(0.080)	1.384(0.055)
	0.02	Naive	—	—	2.224(0.072)	1.377(0.062)	—	—	2.202(0.087)	1.368(0.073)
	0.04	BR-DTRs	1.441(0.083)	1.404(0.067)	2.279(0.071)	1.384(0.051)	1.447(0.083)	1.410(0.069)	2.270(0.076)	1.381(0.055)
	0.04	Naive	—	—	2.207(0.086)	1.359(0.064)	—	—	2.189(0.092)	1.355(0.074)
	0.06	BR-DTRs	1.442(0.089)	1.405(0.074)	2.276(0.071)	1.377(0.050)	1.451(0.090)	1.415(0.073)	2.256(0.064)	1.367(0.045)
	0.06	Naive	—	—	2.185(0.086)	1.348(0.063)	—	—	2.174(0.101)	1.349(0.075)
	0.08	BR-DTRs	1.431(0.086)	1.393(0.070)	2.249(0.078)	1.358(0.048)	1.430(0.091)	1.395(0.073)	2.242(0.064)	1.356(0.042)
	0.08	Naive	—	—	2.164(0.088)	1.322(0.066)	—	—	2.161(0.076)	1.325(0.066)
	0.1	BR-DTRs	1.428(0.081)	1.394(0.066)	2.237(0.066)	1.357(0.045)	1.422(0.091)	1.391(0.075)	2.234(0.069)	1.356(0.048)
	0.1	Naive	—	—	2.168(0.074)	1.321(0.051)	—	—	2.166(0.070)	1.329(0.049)
		AOWL	1.983(0.010)	2.149(0.044)	3.257(0.018)	2.678(0.096)	1.914(0.030)	2.099(0.083)	3.212(0.036)	2.584(0.218)
Setting II	0.02	BR-DTRs	1.362(0.173)	1.246(0.247)	3.174(0.283)	1.262(0.198)	1.456(0.106)	1.403(0.157)	3.069(0.192)	1.329(0.076)
	0.02	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.816(0.019)	0.184(0.018)
	0.04	BR-DTRs	1.306(0.202)	1.166(0.288)	3.228(0.188)	1.299(0.130)	1.459(0.102)	1.402(0.153)	3.066(0.196)	1.319(0.080)
	0.04	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.06	BR-DTRs	1.238(0.252)	1.067(0.371)	3.221(0.197)	1.297(0.126)	1.444(0.123)	1.377(0.189)	3.068(0.208)	1.311(0.086)
	0.06	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.08	BR-DTRs	1.228(0.329)	1.059(0.479)	3.178(0.229)	1.260(0.149)	1.430(0.129)	1.360(0.195)	3.049(0.204)	1.297(0.078)
	0.08	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.1	BR-DTRs	1.177(0.404)	0.980(0.576)	3.169(0.239)	1.247(0.152)	1.438(0.123)	1.371(0.179)	3.009(0.206)	1.271(0.094)
	0.1	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.001)	0.167(0.001)
		AOWL	2.440(0.064)	3.017(0.002)	5.188(0.000)	2.839(0.000)	2.424(0.080)	3.018(0.002)	5.188(0.000)	2.839(0.000)

[1] The estimated results are reported in *median(dev)* format. *median* denotes the median of expected risk/reward estimated via normalized estimator among 600 repeated analyses. *dev* denotes the median value of the absolute difference of estimated risk/reward against the *median*

In the DURABLE study, the major objective is lowering patients' endpoint blood glucose level measured in HbA1c level, and in this analysis, we use the reduction of HbA1c level at 48 weeks since 0 week as the reward outcome. The risk outcome is set to be hypoglycemia frequency encountered by patients, which reflects the potential risk induced by adopting assigned treatment. Since not all patients entered the intensification study with treatment reassignment, to implement BR-DTRs, we assume that for the patients who reached HbA1c level lower than 7.0% at the end of the first phase trial, their current treatment was already optimal and should not be adjusted. Under this assumption, the second stage analysis only involves patients who entered the intensification study; while only in the first stage will all patients be included in the analysis. In the first stage estimation, for the patients in the maintenance study, their future reward outcome (reduction of HbA1c) is assumed to be maintained. That is, in Stage I, the reward outcome becomes

$$Y' = \begin{cases} Y, & \text{if subject is from the maintenance study} \\ Y\frac{\mathbb{I}(A_2\widehat{f}_2(H_2)>0)}{0.5}, & \text{if subject is from the intensification study.} \end{cases}$$

Finally, the second stage risk outcome is the total frequency of hyperglycemia events during the intensification study (from 24 weeks to 48 weeks) and the first stage risk outcome is defined to be the total hypoglycemia events from week 0-24 for patients who entered intensification study, and the total hypoglycemia events from week 0-48 rescaled to 24 weeks for the remaining patients who entered maintenance study. In the analysis, we eventually apply the logarithm transformation to these counts to handle some extremely large counts in the data.

We consider 20 relevant covariates as the baseline predictors H_1, including HbA1c testing result, heart rate, systolic/diastolic blood pressures, body weight, body height, BMI and 7 points self-monitored blood glucose measured at baseline (week 0) along with patient's age, gender, duration of T2D and 3 indicator variables indicating whether patients were taking metformin, thiazolidinedione or sulfonylureas. The second stage predictors H_2 include all predictors in H_1, patient's treatment assignment, the cumulative number of hyperglycemia events during the first stage, along with heart rate, systolic/diastolic blood pressures, HbA1c and same 7 points self-monitored blood glucose measured at the initial time of the second stage (24 weeks). All covariates are centered at mean 0 and rescaled to be within $[-1, 1]$.

The final study cohort includes 579 patients from the intensification study and another 781 from the maintenance study. To compare the performance, we randomly sample 50% patients from the intensification

Table 2.2: Estimated reward/risk under different risk constraints for DURABLE study analysis

Risk Constraint		BR-DTRs			Naive		
τ_2	τ_1	Reward	Stage II Risk	Stage I Risk	Reward	Stage II Risk	Stage I Risk
0.334	0.893	1.471(0.072)	0.311(0.033)	0.844(0.044)	1.460(0.087)	0.311(0.033)	0.842(0.049)
	0.948	1.520(0.078)	0.311(0.033)	0.874(0.067)	1.499(0.091)	0.311(0.033)	0.868(0.066)
	1.005	1.547(0.089)	0.311(0.033)	0.929(0.102)	1.527(0.098)	0.311(0.033)	0.923(0.111)
∞	0.893	1.598(0.043)	0.347(0.028)	0.832(0.039)	1.604(0.048)	0.347(0.028)	0.840(0.040)
	0.948	1.605(0.053)	0.347(0.028)	0.832(0.040)	1.607(0.056)	0.347(0.028)	0.850(0.056)
	1.005	1.620(0.068)	0.347(0.028)	0.922(0.107)	1.625(0.062)	0.347(0.028)	0.888(0.103)
	∞	1.713(0.052)	0.347(0.025)	1.040(0.047)	-	-	-

study as the training sample for stage II and additional 50% patients from the maintenance study as the training sample for stage I. The remaining patients will be treated as the testing data to assess the performance of the estimated rules. We consider different risk constraints $\tau_2 = (0.334, \infty)$ and $\tau_1 = (0.893, 0.948, 1.005)$ where we rescale the risk to hypoglycemia events per 4 weeks. We note that 0.334 and 0.948 are the mean risks of stage II and stage I, respectively, and 1.005 is close to the median estimated risk on testing data under the unconstrained case. We repeat the analysis 100 times for random splitting of the training and testing data. In our method, we use the Gaussian kernel and choose $\eta = 0.02$, while tuning parameter C_n for each stage will be selected by two-fold cross-validation similar to the simulation studies. The bandwidth of the Gaussian kernel is also selected similar to the simulation studies.

All real data analysis results are displayed in Table 2.2. From Table 2.2 we first notice that in each stage, the median estimated risk on testing data is tightly controlled by the prespecified risk constraints. This demonstrates that BR-DTRs can also successfully control adverse risks in real applications. Under each risk constraint, the cumulative reward estimated by BR-DTRs is only slightly better or closed against the estimated reward using the naive method. One reason is that the majority of the patients in stage I would not enter the intensification study and, hence, have no delayed treatment effect at all.

Among all 7 constraint settings, the uncontrolled setting, as expected, produces the estimated rules with both the highest reward and risks, and the estimated reward decreases as the risk constraint of either stage is decreasing. Under the unconstrained estimated optimal rules, all patients are recommended to receive LMx2 in the first stage and later switch to MMx3 after 24 weeks if patients' HbA1c level is greater than 7% by the end of the first phase. As a comparison, when the risk constraint is imposed in stage II, the optimal rules will instead recommend all patients to receive BBT when patients failed to reach HbA1c lower than 7% in the second stage at a price of significantly lower reduction in HbA1c by the end of 48 months. Similar treatment

preference change happens in stage I as the optimal estimated rule becomes less favorable to LMx2 against insulin glargine when τ_1 decreases.

Comparing the reward and risks under the different choice of risk constraint, $\tau_1 = 1.005$ and $\tau_2 = \infty$ produces the second highest reward with moderate risk in the second stage and 10% lower risk in the first stage compared to the unconstrained setting. Under this suboptimal setting, the estimated rules recommend only 50.7% of patients start with LMx2 therapy and later switch to MMx3 therapy if patients fail to reach an HbA1c level of less than 7.0% by the end of the first phase of treatment. By checking the baseline covariates between the patients who received different treatment recommendations, under this estimated rule for the patients whose baseline HbA1c falls in the range $[7, 8)$, $[8, 9)$ and $[9, 10)$, the proportion of the patients who are recommended with LMx2 therapy drops from 62.7% to 56.3% and 46.3%; similarly, for the patients whose baseline BMI falls in the range $[28, 32)$, $[32, 34)$ to $[34, 36)$, the proportion of patients recommended with LMx2 also drops from 59.3% to 53.8% and 51.3%. The negative correlation between the increment of baseline HbA1c/BMI against the proportion of patients recommended with LMx2 as the first phase treatment indicates that the patients with a worse initial health condition are less likely to be recommended with LMx2 therapy as the initial treatment when the risk impact is considered. This is consistent with the fact that LMx2 is an intenser therapy compared with insulin glargine therapy and would cause more hypoglycemia events among unhealthier T2D patients. In particular, the suboptimal rules obtained from BR-DTRs meet the ADA guidance which suggests that intensive insulin therapy should be prescribed to patients according to patients' health condition to reduce potential hypoglycemia events. In conclusion, the real data application demonstrates that, by evaluating the impact of adverse risks along with beneficial reward, BR-DTRs can produce better personalized, more practically implementable treatment recommendations compared with standard O-learning which only takes beneficial reward into consideration.

2.6 Discussion

In this chapter, we introduced a new statistical framework BR-DTRs to estimate the optimal dynamic treatment rules under the stagewise risk constraints. The backward induction technique provided a natural numerical algorithm to solve for BR-DTRs efficiently through iteratively solving a series of standard quadratic programming problems. We also established a non-asymptotic risk bound for the value and stagewise risks

under the estimated decision function. The theoretical results, for the first time, provided the performance guarantee for the constrained support vector machine problem.

It is worth noting that even though in BR-DTRs we assumed treatments to be dichotomous and only one risk constraint is imposed at each stage, our method can also be extended to problems with more treatment options and risk constraints at each stage. One can achieve this by imposing multiple smooth risk constraints to multicategory learning algorithms, such as AD-learning proposed by Qi et al. (2020). However, verifying the Fisher consistency of generalized problems is not trivial. Moreover, even though throughout the chapter we defined R_t as the adverse risk of treatments, the interpretation of R_t can indeed be more general based on the context of the applications. For example, R_t can be the cost of resources used in treating patients, constraints due to patients' preference for treatment design, or fairness of treatment policy among races, genders, or different socioeconomic backgrounds. By allowing different constraints to be imposed in BR-DTRs, our method can also provide a unified framework that optimizes the cumulative benefit while meeting all the constraints from practice.

When proposing the BR-DTRs, our main motivation is to find the optimal decision rules that maximize population reward while satisfying stagewise risk constraints. For many applications, finding the most influential feature variables that decide the optimal decision rules under the consideration of adverse risks is of equal importance, and our proposed method can also be extended to address this by including feature selection during the estimation. For example, when RKHS is generated by the linear kernel, the optimal decision boundary is linear, and one can add a penalty term with a group structure to impose sparsity on feature variables. Lastly, our proposed method focuses on the scenario when controlling the short-term risks of the treatments is of interest with the finite time horizon. In other real applications, the long-term risk or disease burden may also need to be controlled for patients' benefit, and the time horizon can be infinite such as in mobile health. Due to the limitation of the backward induction technique, the BR-DRTs framework cannot be directly extended to deal with these problems. Further investigation is needed to address the infinite time horizon problem.

2.7 Details of DC Algorithm for Solving Single Stage BR-DTRs

In this section, we describe the DC algorithm for solving BR-DTRs at stage t. The algorithm was originally proposed in Wang, Fu and Zeng (2018). Given estimated rules $(\widehat{f}_{t+1}, ..., \widehat{f}_T)$, one can calculate $\widehat{Y}_{it}$

and $\widehat{A}_{it}$ from (2.5). Our goal is to solve the optimization problem

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathbb{R}^d,\beta_0\in\mathbb{R}} \quad & C_n\sum_{i=1}^{n}\frac{\widehat{Y}_{it}}{p(A_{it}|H_{it})}\phi\big(\widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\big)+\frac{1}{2}\boldsymbol{\beta}^T K_t\boldsymbol{\beta}\\ \text{subject to} \quad & \sum_{i=1}^{n}\frac{R_{it}}{p(A_{it}|H_{it})}\psi\big(A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0),\eta\big)\le n\tau_t,\end{aligned}$$

where $C_n=(2n\lambda_{n,t})^{-1}$. K_t is the n-by-n kernel matrix of stage t defined by $K_{ij}=K(H_{it},H_{jt})$ where $K(\cdot,\cdot)$ is the inner product equipped by RKHS $\mathcal{G}_t$ and $K_{i,t}$ is the i-th row of K_t.

Note that the shifted ramp loss can be decomposed as $\psi(x,\eta)=\eta^{-1}(x+\eta)_+-\eta^{-1}(x)_+$. By applying the DC algorithm, given $\boldsymbol{\beta}^{(s)}$ and $\beta_0^{(s)}$, we update $(\boldsymbol{\beta},\beta_0)$ by solving optimization problem

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_t},\beta_0\in\mathbb{R}} \quad & C_n\sum_{i=1}^{n}\frac{\widehat{Y}_{it}}{p(A_{it}|H_{it})}\phi\big(\widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\big)+\frac{1}{2}\boldsymbol{\beta}^T K_t\boldsymbol{\beta}\\ \text{subject to} \quad & \sum_{i=1}^{n}\frac{R_{it}}{p(A_{it}|H_{it})}\Big[\big\{A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)+\eta\big\}_+-C_{it}^{(s)}A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\Big]\le n\eta\tau_t,\end{aligned}$$

where $C_{it}^{(s)}=\mathbb{I}(A_{it}(K_{i,t}\boldsymbol{\beta}^{(s)}+\beta_0^{(s)})>0)$. Similar to standard SVM, we introduce slacking variables $\xi_i\ge 1-\widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)$, $\xi_i\ge 0$ to replace $\phi\big(\widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\big)$ in the objective function. Moreover, we introduce additional slacking variables $\zeta_i\ge A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)+\eta$, $\zeta_i\ge 0$ to replace $\big\{A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)+\eta\big\}_+$ in the risk constraint. After plugging the slacking variables, the optimization problem becomes

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_t},\beta_0\in\mathbb{R}} \quad & C_n\sum_{i=1}^{n}\frac{\widehat{Y}_{it}}{p(A_{it}|H_{it})}\xi_i+C_n\sum_{i=1}^{n}\frac{\zeta_i}{n}+\frac{1}{2}\boldsymbol{\beta}^T K_t\boldsymbol{\beta}\\ \text{subject to} \quad & \sum_{i=1}^{n}\frac{R_{it}}{p(A_{it}|H_{it})}\Big[\zeta_i-C_{it}^{(s)}A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\Big]\le n\eta\tau_t,\\ & 1-\widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)\le\xi_i,\ \ 0\le\xi_i,\\ & A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)+\eta\le\zeta_i,\ \ 0\le\zeta_i,\ \ \text{for } i=1,..,n.\end{aligned}\tag{2.8}$$

The additional term $C_n\sum_{i=1}^{n}\frac{\zeta_i}{n}$ in the objective function is to guarantee that the slacking variable ζ_i is equal to $\big\{A_{it}(K_{i,t}\boldsymbol{\beta}+\beta_0)+\eta\big\}_+$. For fixed tuning parameter C_n, this optimization problem will be equivalent to the original problem as the additional term will eventually vanish when the sample size n increases.

The Lagrange function of (2.8) is given by

$$\begin{aligned}L =& C_n\bigg(\sum_{i=1}^n \frac{\widehat{Y}_{it}}{p(A_{it}|H_{it})}\xi_i + \sum_{i=1}^n \frac{\zeta_i}{n}\bigg) + \frac{1}{2}\boldsymbol{\beta}^T K_t \boldsymbol{\beta} \\ & - \pi\bigg[n\eta\tau_t - \sum_{i=1}^n \frac{R_{it}}{p(A_{it}|H_{it})}\bigg(\zeta_i - \sum_{i=1}^n C_{it}^{(s)} A_{it}(K_{i,t}\boldsymbol{\beta} + \beta_0)\bigg)\bigg] \\ & - \sum_{i=1}^n \alpha_i\bigg[\xi_i - 1 + \widehat{A}_{it}(K_{i,t}\boldsymbol{\beta} + \beta_0)\bigg] - \sum_{i=1}^n \mu_i\xi_i - \sum_{i=1}^n \kappa_i\bigg[\zeta_i - \eta - A_{it}(K_{i,t}\boldsymbol{\beta} + \beta_0)\bigg] - \sum_{i=1}^n \rho_i\kappa_i.\end{aligned}$$

Taking derivatives w.r.t. ξ_i, ζ_i, $\boldsymbol{\beta}$ and β_0, one can obtain that the optimal Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n^T)$, $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_n)^T$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$, $\boldsymbol{\rho} = (\rho_1, ..., \rho_n)$ and π satisfy

$$\begin{aligned} C_n \boldsymbol{V}_{t,Y} - \boldsymbol{\alpha} - \boldsymbol{\mu} &= \mathbf{0}, \\ C_n \mathbf{1}/n + \pi \boldsymbol{V}_{t,R} - \boldsymbol{\kappa} - \boldsymbol{\rho} &= \mathbf{0}, \\ \boldsymbol{\beta} - \pi \boldsymbol{V}_{t,R,A,C}^{(s)} - \widehat{A}_t\boldsymbol{\alpha} + A_t\boldsymbol{\kappa} &= \mathbf{0}, \\ \pi \mathbf{1}^T \boldsymbol{V}_{t,R,A,C}^{(s)} + \mathbf{1}^T\widehat{A}_t\boldsymbol{\alpha} - \mathbf{1}^T A_t\boldsymbol{\kappa} &= 0,\end{aligned}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote n-by-1 vectors with all entries equal to 1 and 0 respectively,

$$\boldsymbol{V}_{t,Y} = \begin{bmatrix} \dfrac{\widehat{Y}_{1t}}{p(A_{1t}|H_{1t})} \\ \vdots \\ \dfrac{\widehat{Y}_{nt}}{p(A_{nt}|H_{nt})} \end{bmatrix}, \quad \boldsymbol{V}_{t,R} = \begin{bmatrix} \dfrac{R_{1t}}{p(A_{1t}|H_{1t})} \\ \vdots \\ \dfrac{R_{nt}}{p(A_{nt}|H_{nt})} \end{bmatrix}, \quad \boldsymbol{V}_{t,R,A,C}^{(s)} = \begin{bmatrix} \dfrac{R_{1t}}{p(A_{1t}|H_{1t})} A_{1t} C_{1t}^{(s)} \\ \vdots \\ \dfrac{R_{nt}}{p(A_{nt}|H_{nt})} A_{nt} C_{nt}^{(s)} \end{bmatrix}.$$

Here, we abuse the notation and define $\widehat{A}_t = \text{diag}\{(\widehat{A}_{1t}, ..., \widehat{A}_{nt})\}$ and $A_t = \text{diag}\{(A_{1t}, ..., A_{nt})\}$. Plugging the equations back to L and note that $\boldsymbol{\alpha} \geq 0$, $\boldsymbol{\kappa} \geq \mathbf{0}$, $\boldsymbol{\mu} \geq \mathbf{0}$, $\boldsymbol{\rho} \geq \mathbf{0}$ and $\pi \geq 0$, after some algebra one can obtain that the dual problem of (2.8) w.r.t. $\boldsymbol{\omega} = (\pi, \boldsymbol{\alpha}^T, \boldsymbol{\kappa}^T)^T$ is given by

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & \frac{1}{2}\boldsymbol{\omega}^T(H^T K_t H)\boldsymbol{\omega} - \boldsymbol{\omega}^T \boldsymbol{l}_{\eta,\tau_t} \\ \text{subject to} \quad & \boldsymbol{a} \leq W\boldsymbol{\omega} \leq \boldsymbol{b}, \quad \mathbf{0}_{(2n+1)\times 1} \leq \boldsymbol{\omega} \leq \boldsymbol{u}, \end{aligned}$$

where

$$H = \begin{bmatrix} \boldsymbol{V}_{t,R,A,C}^{(s)} & \widehat{A}_t & -A_t \end{bmatrix}, \quad W = \begin{bmatrix} \boldsymbol{V}_{t,R} & \mathbf{0}_{n\times n} & -I_n \\ \mathbf{1}^T \boldsymbol{V}_{t,R,A,C}^{(s)} & \mathbf{1}^T\widehat{A}_t & -\mathbf{1}^T A_t \end{bmatrix},$$

$\boldsymbol{l}_{\eta,\tau_t} = (-n\eta\tau_t, \mathbf{1}^T, \eta\mathbf{1}^T)^T$, $\boldsymbol{a} = (-C_n\mathbf{1}^T/n, 0)^T$, $\boldsymbol{b} = (\infty\mathbf{1}^T, 0)^T$ and $\boldsymbol{u} = (\infty, C_n\boldsymbol{V}_{t,Y}^T, \infty\mathbf{1}^T)^T$. Note that the optimization w.r.t. $\boldsymbol{\omega}$ is a standard quadratic optimization problem, which can be solved efficiently via gradient descent methods. Denote the optimal solution of previous optimization problem by $\widehat{\boldsymbol{\omega}}^{(s)}$, we update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(s+1)} = \widehat{\pi}^{(s)}\boldsymbol{V}_{t,R,A,C}^{(s)} + \widehat{A}_t\widehat{\boldsymbol{\alpha}}^{(s)} - A_t\widehat{\boldsymbol{\kappa}}^{(s)}.$$

The new $\beta_0^{(s+1)}$ can be determined via grid search such that the original objective function is maximized among values that satisfy the constraint given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s+1)}$, and we adopted this approach in our work. Alternatively, using Karush–Kuhn–Tucker conditions one can also determine $\beta_0^{(s+1)}$ by taking the average of constraints among all support vectors lie on the margin (Hastie, Tibshirani and Friedman, 2009, Chapter 12.2), i.e., $\beta_0^{(s+1)}$ is given by solving either

$$\sum_{\{i\in\{1,\ldots,n\}|\widehat{\alpha}_i^{(s)}>0,\widehat{\mu}_i^{(s)}>0\}} \left[1 - \widehat{A}_{it}(K_{i,t}\boldsymbol{\beta}^{(s+1)} + \beta_0)\right] = 0,$$

or

$$\sum_{\{i\in\{1,\ldots,n\}|\widehat{\kappa}_i^{(s)}>0,\widehat{\rho}_i^{(s)}>0\}} \left[\eta + A_{it}(K_{i,t}\boldsymbol{\beta}^{(s+1)} + \beta_0)\right] = 0,$$

or combine both. The DC iteration stops when the termination condition $\max(|\boldsymbol{\beta}^{(s+1)} - \boldsymbol{\beta}^{(s)}|_\infty, |\beta_0^{(s+1)} - \beta^{(s)}|) \leq \epsilon$ is satisfied. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_n)^T$ and $\widehat{\beta}_0$ denote the final solution returned by DC iteration, then the final estimated decision function at stage t is given by $\widehat{f}_t(\cdot) = \sum_{i=1}^n K(H_{it}, \cdot)\widehat{\beta}_i + \widehat{\beta}_0$.

2.8 Proof of Theorem 2.1

We summarize the additional notations used in the proofs below:

f_t^*	the true optimal decision function solving the BR-DTRs (2.3), and we use $f_{t,\tau}^*$ for f_t^* whenever τ is necessary in the context;
$\mathcal{V}_{t,\phi}(s,h)$	$-\{\phi(s)E[Q_t\|H_t = h, A_t = 1] + \phi(-s)E[Q_t\|H_t = h, A_t = -1]\}$;
$\mathcal{R}_{t,\psi}(s,\eta,h)$	$\psi(s,\eta)E[R_t\|H_t = h, A_t = 1] + \psi(-s,\eta)E[R_t\|H_t = h, A_t = -1]$.

When $T = 1$, we omit subscript t from all these notations.

2.8.1 Proof of Theorem 2.1 for $T = 1$

We consider $T = 1$. After dropping the stage subscript, both (2.1) and (2.2) problems are equivalent to solving

$$\begin{aligned} &\min_{f\in\mathcal{F}} E\left[\frac{YI(Af(H)<0)}{p(A|H)}\right] \\ &\text{subject to } E\left[\frac{RI(Af(H)>0)}{p(A|H)}\right] \le \tau, \end{aligned} \tag{2.9}$$

and its resulting decision is given by $\text{sign}(g^*)$. Without loss of generality, we assume that Y is non-negative; otherwise, we can change Y to $|Y|$ and A to $A * \text{sign}(Y)$.

We define $\mathcal{M} = \{h : \delta_Y(h)\delta_R(h) < 0\}$, i.e., the set of subjects where the beneficial treatment also reduces risk. Then according to Theorem 1 in Wang, Fu and Zeng (2018), for any $\tau \in (\tau_{\min}, \tau_{\max})$, the optimal f^* can be chosen as

$$g^*(h) = \begin{cases} \text{sign}(\delta_Y(h)), & \text{if } h \in \mathcal{M} \\ 1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) > \lambda^*, \delta_Y(h) > 0\} \cap \mathcal{M}^c \\ -1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) < \lambda^*, \delta_Y(h) > 0\} \cap \mathcal{M}^c \\ -1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) > \lambda^*, \delta_Y(h) < 0\} \cap \mathcal{M}^c \\ 1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) < \lambda^*, \delta_Y(h) < 0\} \cap \mathcal{M}^c, \end{cases}$$

where λ^* satisfies $E[\mathcal{R}(g^*, H)] = \tau$. Our surrogate problem to be solved is (2.3), which is

$$\begin{aligned} &\min_{f\in\mathcal{F}} E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] \\ &\text{subject to } E\left[\frac{R\psi(Af(H),\eta)}{p(A|H)}\right] \le \tau \end{aligned} \tag{2.10}$$

We let f^* denote the solution. Our following theorem (the same version for Theorem 2.1 for $T = 1$) gives an explicit expression for f^* so that the solution for the surrogate problem has the same sign as g^*.

Theorem 2.3 *For any fixed $\tau_{\min} < \tau < \tau_{\max}$, suppose that $P(\delta_Y(H)\delta_R(H) = 0) = 0$ and random variable $\delta_Y(H)/\delta_R(H)$ has distribution function with a continuous density function in the support of H. Then for*

any $\eta \in (0,1]$, $f^*(h)$ *can be taken as*

$$f^*(h) = \begin{cases} sign(\delta_Y(h)), & \text{if } h \in \mathcal{M} \\ 1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) > \lambda^*, \delta_Y(h) > 0\} \cap \mathcal{M}^c \\ -\eta, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) < \lambda^*, \delta_Y(h) > 0\} \cap \mathcal{M}^c \\ -1, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) > \lambda^*, \delta_Y(h) < 0\} \cap \mathcal{M}^c \\ \eta, & \text{if } h \in \{\delta_Y(h)/\delta_R(h) < \lambda^*, \delta_Y(h) < 0\} \cap \mathcal{M}^c, \end{cases} \tag{2.11}$$

where λ^* *is the same one in the definition of* g^*.

By comparing the expressions for g^* and f^*, we immediately conclude that they have the same signs so solving (2.10) leads to a Fisher consistent solution to the original problem in (2.9). The proof consists of several steps. For any decision function f, we say that f is feasible meaning that f satisfies the risk constraint in the surrogate problem (2.10), and for any two feasible functions, f_1 and f_2, "f_1 is non-inferior to f_2" means that the objective function in (2.10) is less than or equal to the one for f_2, and "f_1 is superior to f_2" if the objective function is strictly less than.

From now on, we assume $\eta \in (0,1]$ and $\tau \in (\tau_{\min}, \tau_{\max})$. By the definitions of $\mathcal{V}_\phi$ and $\mathcal{R}_\psi$, we note

$$E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] = -E[\mathcal{V}_\phi(f,H)],$$

$$E\left[\frac{R\psi(Af(H),\eta)}{p(A|H)}\right] = E[\mathcal{R}_\psi(f,\eta,H)].$$

Proof of Theorem 2.3:

Step 1. We show that the value for the optimal solution, f^*, can be restricted within $[-1,1]$. That is, the following lemma holds.

Lemma 2.1 *For any feasible decision function* $f(h)$*, define* $\widetilde{f}(h) = \min(\max(f(h),-1),1)$ *as the truncated* f *at -1 and 1. Then* $\widetilde{f}$ *is non-inferior to* f.

Proof: Note that $\psi(h,\eta) = \psi(1,\eta)$ for any $h > 1$ and $\psi(h,\eta) = \psi(-1,\eta)$ for any $h < -1$. Thus, it follows from $\eta \leq 1$ that $E[\mathcal{R}_\psi(\widetilde{f},\eta,H)] = E[\mathcal{R}_\psi(f,\eta,H)] \leq \tau$, so $\widetilde{f}$ is feasible. Moreover, it is easy to see that if

$f(h) > 1$, then $\widetilde{f}(h) = 1$ so

$$E\left[\frac{Y\phi(Af(H))}{p(A|X)}\middle|H = h\right] = E[Y|A = -1, H = h](1 + f(h))$$

$$\geq 2E[Y|A = -1, H = h] = E\left[\frac{Y\phi(A\widetilde{f}(H))}{p(A|X)}\middle|H = h\right].$$

Similarly, if $f(h) < -1$,

$$E\left[\frac{Y\phi(Af(H))}{p(A|X)}\middle|H = h\right] = E[Y|A = 1, H = h](1 - f(h))$$

$$\geq 2E[Y|A = -1, H = h] = E\left[\frac{Y\phi(A\widetilde{f}(H))}{p(A|X)}\middle|H = h\right].$$

Since $f(h) = \widetilde{f}(h)$, when $|f(h)| \leq 1$, we conclude

$$E\left[\frac{Y\phi(Af(H))}{p(A|X)}\right] \geq E\left[\frac{Y\phi(A\widetilde{f}(H))}{p(A|X)}\right].$$

Thus, Lemma 2.1 holds. □

Step 2. We characterize the expression of $f^*(h)$ for $h \in \mathcal{M}$, which is the region where the beneficial treatment also reduces the risk.

Lemma 2.2 *For any feasible function f with $|f| \leq 1$, we define*

$$\widetilde{f}(h) = f(h)I(h \in \mathcal{M}^c) + \textit{sign}(\delta_Y(h))I(h \in \mathcal{M}).$$

Then $\widetilde{f}$ is non-inferior to f.

Proof: For h with $\delta_Y(h) > 0$ and $\delta_R(h) < 0$, $\mathcal{R}_\psi(s, \eta, h)$ is minimized when $s \in [\eta, 1]$, while $\mathcal{V}_\phi(s, h)$ is maximized at $s = 1$. Since $\widetilde{f}(h) = 1$ $\mathcal{R}_\psi(\widetilde{f}(h), \eta, h) \leq \mathcal{R}_\psi(f(h), \eta, h)$ and $\mathcal{V}_\phi(\widetilde{f}(h), h) \geq \mathcal{V}_\phi(f(h), h)$. The same inequalities hold for h with $\delta_Y(h) < 0$ and $\delta_R(h) > 0$. In other words, they hold for any $h \in \mathcal{M}$.

Since $\widetilde{f}(h) = f(h)$ for $h \in \mathcal{M}^c$,

$$E[\mathcal{R}_\psi(f, \eta, H)] - E[\mathcal{R}_\psi(\widetilde{f}, \eta, H)] = E[(\mathcal{R}_\psi(f, \eta, H) - \mathcal{R}_\psi(\widetilde{f}, \eta, H))\mathbb{I}(H \in \mathcal{M})] \geq 0,$$

and similarly, $E[\mathcal{V}_\phi(f,H)] - E[\mathcal{V}_\phi(\widetilde{f},H)] \leq 0$. We conclude that $\widetilde{f}$ is non-inferior to f. $\square$

Step 3. From steps 1 and 2, we can restrict f to satisfy $|f| \leq 1$ and $f(h) = \text{sign}(\delta_Y(h))$ for $h \in \mathcal{M}$. Furthermore, since $\tau_{\max}$ is the risk under decision rule $\text{sign}(\delta_Y(h))$, $\tau < \tau_{\max}$ implies that

$$P(f(H) \neq \text{sign}(\delta_Y(H)), H \in \mathcal{M}^c) > 0.$$

In this step, we wish to show that the optimal solution should attain the risk bound, i.e., $E[\mathcal{R}_\psi(f,\eta,H)] = \tau$. Otherwise, assume for some feasible solution f such that $E[\mathcal{R}_\psi(f,\eta,H)] = \tau_0 < \tau$. Consider two sets

$$\mathcal{D}^+ = \{h \in \mathcal{H} : f(h) < 1, \delta_Y(h) > 0\} \cap \mathcal{M}^c$$

$$\mathcal{D}^- = \{h \in \mathcal{H} : f(h) > -1, \delta_Y(h) < 0\} \cap \mathcal{M}^c,$$

then $P(\mathcal{D}^+) + P(\mathcal{D}^-) > 0$. Without loss of generality, we assume that $P(\mathcal{D}^+) > 0$. We construct

$$\widetilde{f}(h) = \begin{cases} f(h), & \text{if } h \notin \mathcal{D}^+ \\ \min\left(f(h) + \frac{\eta(\tau-\tau_0)}{MP(\mathcal{D}^+)}, 1\right), & \text{if } h \in \mathcal{D}^+, \end{cases}$$

where M is the bound for R.

For $h \in \mathcal{D}^+$, $\mathcal{V}_\phi(\widetilde{f}(h),h) > \mathcal{V}_\phi(f(h),h)$ since $1 \geq \widetilde{f}(h) > f(h)$ and $\mathcal{V}_\phi(s,h)$ is an strictly increasing function of $s \in [-1,1]$ due to $\delta_Y(h) > 0$. We immediately conclude $E[\mathcal{V}_\phi(\widetilde{f},H)] > E[\mathcal{V}_\phi(f,H)]$. On the other hand, $\mathcal{R}_\psi(s,\eta,h)$ is a piecewise linear function of s with absolute value of slopes no larger than

$$\frac{\max(E[R|H=h,A=1], E[R|H=h,A=-1])}{\eta} \leq \frac{M}{\eta}.$$

Hence, it follows that

$$\begin{aligned} E[\mathcal{R}_\psi(\widetilde{f},\eta,H)] =& E[\mathcal{R}_\psi(\widetilde{f},\eta,H)] - E[\mathcal{R}_\psi(f,\eta,H)] + E[\mathcal{R}_\psi(f,\eta,H)] \\ \leq& E[(\mathcal{R}_\psi(\widetilde{f},\eta,H) - \mathcal{R}_\psi(f,\eta,H))\mathbb{I}(H \in \mathcal{D}^+)] + \tau_0 \\ \leq& \frac{M}{\eta}\frac{\eta(\tau-\tau_0)}{MP(\mathcal{D}^+)}P(\mathcal{D}^+) + \tau_0 = \tau. \end{aligned}$$

As a result, $\widetilde{f}$ is superior to f with a strictly larger objective function, a contradiction. In other words, the expected risk for the optimal solution should attain the bound.

With steps 1-3, we can restrict within the class

$$\mathcal{W} = \{f : |f| \leq 1, f(h) = \text{sign}(\delta_Y(h)) \text{ for } h \in \mathcal{M}, E[R_\psi(f, \eta, H)] = \tau\}$$

to find the optimal decision function.

Step 4. We derive the expression of the optimal function for f by considering solving a Lagrange multiplier for the problem (2.10):

$$\max_f -E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] - \nu\left(E\left[\frac{R\psi(Af(H))}{p(A|H)}\right] - \tau\right),$$

where ν is a constant to be determined by the constraint in $\mathcal{W}$. We maximize the above function by maximizing the conditional mean of the term in the expectation given $H = h$ for every h, which is given by

$$G(f) \equiv \mathcal{V}_\phi(f, h) - \nu R_\psi(f, \eta, h).$$

Note that $G(f)$ is now a function w.r.t. the value of f given fixed h. Since $f \in [-1, 1]$ and f in $\mathcal{W}$ is already given for $h \in \mathcal{M}$, it suffices to examine that for $h \in \mathcal{M}^c$. In addition, $G(f)$ is a piecewise linear function for $f \in [-1, -\eta]$, $(-\eta, 0]$, $(0, \eta]$ and $(\eta, 1]$. Thus, the maximizer can only be achieved at points $-1, -\eta, 0, \eta$ and 1. Note that R is assumed to be positive, $G'(0) = -\nu/\eta(E[R|H = h, A = 1] + E[R|H = h, A = -1]) < 0$ if $\nu > 0$, or > 0 if $\nu < 0$. For $\nu = 0$, $G(0) = -E[Y|H = h, A = 1] - E[Y|H = h, A = -1] = (G(1) + G(-1))/2$. Thus, the maximum for $G(f)$ can always be attained at f which is not zero. In other words, we only need to compare the values at $f \in \{-1, -\eta, \eta, 1\}$.

Simple calculation gives

$$G(-1) = -2E[Y|H = h, A = 1] - \nu E[R|H = h, A = -1],$$

$$G(-\eta) = -(1 + \eta)E[Y|H = h, A = 1] - (1 - \eta)E[Y|H = h, A = -1] - \lambda E[R|H = h, A = -1],$$

$$G(\eta) = -(1 - \eta)E[Y|H = h, A = 1] - (1 + \eta)E[Y|H = h, A = -1] - \nu E[R|H = h, A = 1],$$

and

$$G(1) = -2E[Y|H = h, A = -1] - \nu E[R|H = h, A = 1].$$

When $\delta_Y(h) > 0$ so $\delta_R(h)$ is also positive, it is straightforward to check $G(1) > G(\eta)$ and $G(-\eta) > G(-1)$. Note $G(1) - G(-\eta) = (1+\eta)\delta_Y(h) - \lambda\delta_R(h)$ so we immediately conclude that the optimal value for f should be 1, if $\delta_Y(h) > \lambda$, where $\lambda = \nu/(1+\eta)$, and it is $-\eta$ otherwise. When $\delta_Y(h) \leq 0$, we use the same arguments to obtain that the optimal value for f should be -1 if $\delta_Y(h) > \lambda$, and it is η otherwise. Therefore, the optimal function maximizing the Lagrange multiplier for any fixed ν (equivalently, λ) has the same expression as (2.11).

Next, we show that there is some positive λ^* such that

$$E[RI(Ag^*(H) > 0)/p(A|H)] = E[RI(Af^*(H) > 0)/p(A|H)] = E[\mathcal{R}_\psi(f^*, \eta, H)] = \tau.$$

The first equality follows from the fact that $\text{sign}(g^*) = \text{sign}(f^*)$, and the second equality follows from that $R_\psi(s, \eta, h)$ is constant for any $s \in [-1, -\eta]$ and $s \in [0, \eta]$. To prove the existence of λ^*, we notice

$$\begin{aligned}
\Gamma(\lambda) \equiv & E[RI(Af^*(H) > 0)/p(A|H)] \\
= & E[E[R|H, A = 1]\mathbb{I}(H \in \{\delta_Y(h) > 0\} \cap \mathcal{M})] \\
& + E[E[R|H, A = -1]\mathbb{I}(H \in \{\delta_Y(h) < 0\} \cap \mathcal{M})] \\
& + E[E[R|H, A = 1]\mathbb{I}(H \in \{\delta_Y(h)/\delta_R(h) > \lambda, \delta_Y(h) > 0\} \cap \mathcal{M}^c)] \\
& + E[E[R|H, A = -1]\mathbb{I}(H \in \{\delta_Y(h)/\delta_R(h) < \lambda, \delta_Y(h) > 0\} \cap \mathcal{M}^c)] \\
& + E[E[R|H, A = -1]\mathbb{I}(H \in \{\delta_Y(h)/\delta_R(h) > \lambda, \delta_Y(h) < 0\} \cap \mathcal{M}^c)] \\
& + E[E[R|H, A = 1]\mathbb{I}(H \in \{\delta_Y(h)/\delta_R(h) < \lambda, \delta_Y(h) < 0\} \cap \mathcal{M}^c)]
\end{aligned} \tag{2.12}$$

is a continuous function of λ since $\delta_Y(H)/\delta_R(H)$ has continuous density function. Furthermore, $\Gamma(\infty) = \tau_{\min}$, $\Gamma(0) = \tau_{\max}$. Thus, there exists some $\lambda^* > 0$ such that $\Gamma(\lambda^*) = \tau$.

Finally, for any f, based on steps 1-3, we have

$$-E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] \leq \max_{f \in \mathcal{W}} \left\{-E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right]\right\}.$$

On the other hand, for $f \in \mathcal{W}$, $E[R\psi(Af(H))/p(A|H)] = \tau$ and

$$-E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] - \lambda^*(1+\eta)\left(E\left[\frac{R\psi(Af(H))}{p(A|H)}\right] - \tau\right)$$

$$\leq -E\left[\frac{Y\phi(Af^*(H))}{p(A|H)}\right] - \lambda^*(1+\eta)\left(E\left[\frac{R\psi(Af^*(H))}{p(A|H)}\right] - \tau\right).$$

Therefore,

$$E\left[\frac{Y\phi(Af(H))}{p(A|H)}\right] \geq E\left[\frac{Y\phi(Af^*(H))}{p(A|H)}\right].$$

In other words, f^* given by (2.11) is the optimal solution to the problem (2.10). We thus complete the proof of Theorem 2.3.

2.8.2 Proof of Theorem 2.1 for $T \geq 2$

Start from stage T. For any given $f_1, ..., f_{T-1}$, we consider f_T maximizing

$$E\left[\frac{(\sum_{t=1}^{T} Y_t) I(A_T f_T(H_T) > 0)}{p(A_T|H_T)} \frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right]$$

subject to constraint

$$E\left[\frac{R_T I(A_T f_T(H_T) > 0)}{p(A_T|H_T)} \frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] \leq \tau_T.$$

Based on Theorem 1 in Wang, Fu and Zeng (2018), the optimal solution can be chosen as

$$\widetilde{g}_T^*(h) = \begin{cases} \text{sign}(\delta_{\widetilde{Y}}(h)), & \text{if } h \in \widetilde{\mathcal{M}} \\ 1, & \text{if } h \in \{\delta_{\widetilde{Y}}(h)/\delta_{\widetilde{R}}(h) > \widetilde{\lambda}, \delta_{\widetilde{Y}}(h) > 0\} \cap \widetilde{\mathcal{M}}^c \\ -1, & \text{if } h \in \{\delta_{\widetilde{Y}}(h)/\delta_{\widetilde{R}}(h) < \widetilde{\lambda}, \delta_{\widetilde{Y}}(h) > 0\} \cap \widetilde{\mathcal{M}}^c \\ -1, & \text{if } h \in \{\delta_{\widetilde{Y}}(h)/\delta_{\widetilde{R}}(h) > \widetilde{\lambda}, \delta_{\widetilde{Y}}(h) < 0\} \cap \widetilde{\mathcal{M}}^c \\ 1, & \text{if } h \in \{\delta_{\widetilde{Y}}(h)/\delta_{\widetilde{R}}(h) < \widetilde{\lambda}, \delta_{\widetilde{Y}}(h) < 0\} \cap \widetilde{\mathcal{M}}^c \end{cases}$$

where $\widetilde{\mathcal{M}} = \big\{h : \delta_{\widetilde{Y}}(h)\delta_{\widetilde{R}}(h) < 0\big\}$,

$$\delta_{\widetilde{Y}}(h) = (E[Q_T|H_T = h, A_T = 1] - E[Q_T|H_T = h, A_T = -1])\frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)},$$

$$\delta_{\widetilde{R}}(h) = (E[R_T|H_T = h, A_T = 1] - E[R_T|H_T = h, A_T = -1])\frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)},$$

and and $\widetilde{\lambda}$ satisfies

$$E\left[\frac{R_T I(A_T \widetilde{g}_T^*(H_T) > 0)}{p(A_T|H_T)}\frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] = \tau_T.$$

Note that for h in the support of H_t where $A_t f_t(H_t) \leq 0$ for any $t = 1, ..., T-1$, $\widetilde{g}_T^*(h)$ can be any arbitrary value since it does not affect the value and risk expectations. On the other hand, recall that $g_T^*(h)$ is the function maximizing

$$E\left[\frac{(\sum_{t=1}^{T} Y_t) I(A_T f_T(H_T) > 0)}{p(A_T|H_T)}\right]$$

subject to constraint

$$E\left[\frac{R_T I(A_T f_T(H_T) > 0)}{p(A_T|H_T)}\right] \leq \tau_T.$$

Based on Theorem 1 in Wang, Fu and Zeng (2018), g_T^* is given as

$$g_T^*(h) = \begin{cases} \text{sign}(\delta_{Q_T}(h)), & \text{if } h \in \mathcal{M} \\ 1, & \text{if } h \in \{\delta_{Q_T}(h)/\delta_{R_T}(h) > \lambda^*, \delta_{Q_T}(h) > 0\} \cap \mathcal{M}^c \\ -1, & \text{if } h \in \{\delta_{Q_T}(h)/\delta_{R_T}(h) < \lambda^*, \delta_{Q_T}(h) > 0\} \cap \mathcal{M}^c \\ -1, & \text{if } h \in \{\delta_{Q_T}(h)/\delta_{R_T}(h) > \lambda^*, \delta_{Q_T}(h) < 0\} \cap \mathcal{M}^c \\ 1, & \text{if } h \in \{\delta_{Q_T}(h)/\delta_{R_T}(h) < \lambda^*, \delta_{Q_T}(h) < 0\} \cap \mathcal{M}^c \end{cases}$$

where $\mathcal{M} = \{h : \delta_{Q_T}(h)\delta_{R_T}(h) < 0\}$, and λ^* satisfies

$$E\left[\frac{R_T I(A_T \widetilde{g}_T^*(H_T) > 0)}{p(A_T|H_T)}\frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] = \tau_T.$$

From the above two expressions, it is clear that on the set when $A_t f_t(H_t) > 0$ for all $t = 1, ..., T-1$, $\widetilde{g}_T^*(h)$ takes the same form as the solution as $g_T^*(h)$. Furthermore, due to Assumption 2.4,

$$E\left[\frac{R_T I(A_T f_T(H_T) > 0)}{p(A_T|H_T)} \frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] = E\left[\frac{R_T I(A_T f_T(H_T) > 0)}{p(A_T|H_T)}\right].$$

Thus, we conclude that $\widetilde{\lambda}$ can be chosen to be the same as λ^* so $\widetilde{g}_T^*(h)$ can be chosen to be exactly the same as $g_T^*(h)$. In other words,

$$\mathcal{V}(f_1, ..., f_{T-1}, g_T^*) \geq \mathcal{V}(f_1, ..., f_T)$$

and g_T^* satisfies

$$E\left[\frac{R_T I(A_T g_T^*(H_T) > 0)}{p(A_T|H_T)} \frac{\prod_{t=1}^{T-1} I(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] = \tau_T.$$

By Theorem 2.3, both g_T^* and f_T^* have the same signs. Therefore,

$$\mathcal{V}(f_1, ..., f_{T-1}, f_T^*) \geq \mathcal{V}(f_1, ..., f_T)$$

and f_T^* satisfies

$$E\left[\frac{R_T I(A_T f_T^*(H_T) > 0)}{p(A_T|H_T)}\right] = \tau_T.$$

Once f_T^* is determined, we consider the $T-1$ stage. Now the original problem (2.1) becomes

$$\begin{aligned} \max \mathcal{V}(f_1, ..., f_{T-1}, f_T^*) = \quad & E\left[\frac{(\sum_{t=1}^{T} Y_t) I(A_T f_{T,\eta}^*(H_T) > 0)}{p(A_T|H_T)} \frac{\prod_{t=1}^{T-1} \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] \\ \text{subject to} \quad & E\left[\frac{R_t \mathbb{I}(A_t f_t(H_t) > 0)}{p(A_t|H_t)}\right] \leq \tau_t, \qquad t = 1, ..., T-1. \end{aligned}$$

We repeat the same arguments as for stage T as before, to conclude

$$\mathcal{V}(f_1, ..., f_{T-1}^*, f_T^*) \geq \mathcal{V}(f_1, ..., f_{T-1}, f_T)$$

and f_{T-1}^* satisfies

$$E\left[\frac{R_{T-1} I(A_{T_1} f_{T-1}^*(H_{T-1}) > 0)}{p(A_{T_1}|H_{T-1})}\right] = \tau_{T-1}.$$

We continue this proof till $t = 1$ so conclude that $(f_1^*, ..., f_T^*)$ maximizes the multistage value and satisfies the constraints over all the stages. The above arguments also show that f_t^* has the same sign as g_t^*. Theorem 2.1 thus holds.

2.9 Proof of Theorem 2.2

Instead, we prove a more general version of Theorem 2.2.

Theorem 2.4 *In addition to the conditions in Theorem 2.1, suppose that Assumption 2.5 holds and H_t takes value from a compact subset of $\mathbb{R}^{d_t}$ for $t = 1, ..., T$. Let $(\tau_1, ..., \tau_T)$ and $(\delta_{0,1}, ..., \delta_{0,T})$ denote the constraints and corresponding constants in Assumption 2.5. Let $\delta_t > 0$, $1 \leq x_t$, $0 < \theta_{1,t}$, $0 < \theta_{2,t}$, $0 < \nu_{1,t} < 2$, $0 < \nu_{2,t} \leq 2$ for $t = 1, ..., T$. Give positive parameter $\lambda_{n,t} \to 0$ and $\sigma_{n,t} \to \infty$, and let*

$$\xi_{n,t}^{(1)} = c\Bigg(\frac{2M}{c_1}\sqrt{\frac{M}{c_1\lambda_{n,t}} + \sigma_{n_t}^{d_t}} + \lambda_{n,t}\left(\frac{M}{c_1\lambda_{n,t}} + \sigma_{n,t}^{d_t}\right)\Bigg) n^{-1/2}(\sigma_{n,t}^{(1-\nu_{1,t}/2)(1+\theta_{1,t})d_t/2} + 2\sqrt{2x_t} + 2x_t/\sqrt{n}),$$

$\xi_{n,t}^{(2)} = c(\lambda_{n_t}\sigma_{n_t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t})$ *and* $\xi_{n,t} = \xi_{n,t}^{(1)} + \xi_{n,t}^{(2)}$. *In addition, let*

$$\epsilon'_{n,t} = \delta_t + C_{1,t}\sigma_n^{-\alpha_t d_t}\eta_{n,t}^{-1} + C_{3,t}n^{-1/2}\sigma_{n,t}^{(1-\nu_{2,t}/2)(1+\theta_{2,t})d_t/2}\left(\frac{M}{c_1\lambda_{n,t}} + \sigma_{n,t}^{d_t}\right)^{\nu_{2,t}/4}\eta_{n,t}^{-\nu_{2,t}/2}$$

and

$$h_t(n, x_t) = 2\exp\left(-\frac{2n\delta_{0,t}^2 c_1^2}{M^2}\right) + 2\exp\left(-\frac{n\delta_t^2 c_1^2}{2M^2}\right) + \exp(-x_t).$$

Then for any $n \geq 1$ and $(\lambda_{n,t}, \sigma_{n,t}, \eta_{n,t})$ such that

$$C_{1,t}\sigma_n^{-\alpha_t d_t}\eta_{n,t}^{-1} \leq \delta_{0,t},$$

$$C_{2,t}\sigma_{n,t}^{(1-\nu_{1,t}/2)(1+\theta_{1,t})d_t} \leq 1,$$

$\epsilon'_{n,t} < 2\delta_{0,t}$, *and* $x_t \geq 1$, *with probability at least* $1 - \sum_{t=1}^T h_t(n, x_t)$, *we have*

$$|\mathcal{V}(\widehat{f}_1, ..., \widehat{f}_T) - \mathcal{V}(f_1^*, ..., f_T^*)| \leq \sum_{t=1}^T (c_1/5)^{1-t}(\xi_{n,t} + (T - t + 1)M\eta_{n,t} + 2c\epsilon'_{n,t}). \tag{2.13}$$

Moreover, with probability at least $1 - h_t(n, x_t)$ *the risk induced by* $\widehat{f}_t$ *satisfies*

$$E\left[\frac{R_t \mathbb{I}(A_t \widehat{f}_t(H_t) > 0)}{p(A_t|H_t)}\right] \leq \tau_t + \delta_t + C_{3,t}\sigma_{n,t}^{(1-\nu_{2,t}/2)(1+\theta_{2,t})d_t/2}\left(\frac{M}{c_1\lambda_{n,t}} + \sigma_{n,t}^{d_t}\right)^{\nu_{2,t}/4}\eta_{n,t}^{-\nu_{2,t}/2}. \quad (2.14)$$

Here, c *in front of* $\xi_{n,t}^{(1)}$ *is a positive constant only depends on* $(\nu_{1,t}, \theta_{1,t}, d_t)$*,* c *in front of* $\xi_{n,t}^{(2)}$ *is a positive constant only depends on* (α_t, d_t, K_t, M) *and* c *of* $\epsilon'_{n,t}$ *is a positive constant only depends on* $(\tau_t, \delta_{0,t})$*.* $C_{1,t}$ *is a positive constant depend on* (α_t, d_t, K_t, M)*,* $C_{2,t}$ *is a positive constant depends on* $(\nu_{1,t}, \theta_{1,t}, d_t)$*,* $C_{3,t}$ *a positive constant depends on* $(\nu_{2,t}, \theta_{2,t}, d_t, c_1, M)$*.*

Theorem 2.2 can be obtained from Theorem 2.4 by setting $\theta_t = \theta_{1,t} = \theta_{2,t}$, $\nu_t = \nu_{1,t} = \nu_{2,t}$ and $x_t = \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t}$. $C_i = \sup_t C_{i,t}$ for $i = 1, 2, 3$. We first prove Theorem 2.4 for $T = 1$ and then extend the result to $T \geq 2$.

2.9.1 Proof of Theorem 2.2 (Theorem 2.4) for $T = 1$

Since $T = 1$, we omit the subscript for the stage in this subsection so all the notations are the same as in Section 2.8. Since τ is necessary in the proof, we use f_τ^* to refer to f^* that solves (2.10) corresponding to τ and $\eta = \eta_n$.

2.9.1.1 An excessive risk inequality

In this section, we prove some preliminary lemmas for (2.10). Lemma 2.3 shows that the regret from the optimal decision function solving the original problem (2.9) is bounded by the regret from the one solving the surrogate problem (2.10), plus an additional biased term of order $O(\eta_n)$. Lemma 2.4 shows that the optimal value using the surrogate loss is Lipschitz continuous with respect to τ.

Lemma 2.3 *Under the condition of Theorem 2.3, for any* $f : \mathcal{H} \to \mathbb{R}$ *and any* $\eta_n \in (0, 1]$*, we have*

$$\mathcal{V}(f_\tau^*) - \mathcal{V}(f) \leq E[\mathcal{V}_\phi(f_\tau^*, H)] - E[\mathcal{V}_\phi(f, H)] + M\eta_n.$$

Proof: Theorem 2.3 shows that f_τ^* must have expression (2.11) almost surely. Let $\widetilde{\mathcal{V}}(f, h) = I(f(h) > 0)E[Y|H = h, A = 1] + I(f(h) \leq 0)E[Y|H = h, A = -1]$. For any $h \in \{\delta_Y(h) > 0\}$, we consider the following 6 scenarios:

1. When $h \in \mathcal{M}$, $f_\tau^*(h) = 1$ and $f(h) > 0$, we have $\widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) = 0$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} (1 - f(h))\delta_Y(h), & f(h) \leq 1 \\ (f(h) - 1)m_Y(h, -1), & f(h) > 1, \end{cases}$$

2. When $h \in \mathcal{M}$, $f_\tau^*(h) = 1$ and $f(h) \leq 0$, we have $\widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) = \delta_Y(h)$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} (1 - f(h))\delta_Y(h), & f(h) \geq -1 \\ 2\delta_Y(h) + (-f(h) - 1)m_Y(h, 1), & f(h) < -1, \end{cases}$$

3. When $h \in \mathcal{M}^c$, $f_\tau^*(h) = 1$ and $f(h) > 0$, we have $\widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) = 0$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} (1 - f(h))\delta_Y(h), & f(h) \leq 1 \\ (f(h) - 1)m_Y(h, -1), & f(h) > 1, \end{cases}$$

   in which case $\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) \geq \widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h)$.

4. When $h \in \mathcal{M}^c$, $f_\tau^*(h) = 1$ and $f(h) \leq 0$, we have $\mathcal{V}(f_\tau^*, h) - \mathcal{V}(f, h) = \delta_Y(h)$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} (1 - f(h))\delta_Y(h), & f(h) \geq -1 \\ 2\delta_Y(h) + (-f(h) - 1)m_Y(h, 1), & f(h) < -1, \end{cases}$$

   in which case $\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) \geq \mathcal{V}(f_\tau^*, h) - \mathcal{V}(f, h)$.

5. When $h \in \mathcal{M}^c$, $f_\tau^*(h) = -\eta_n$ and $f(h) > 0$, we have $\widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) = -\delta_Y(h)$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} -f(h)\delta_Y(h) - \eta_n\delta_Y(h), & f(h) \leq 1 \\ -\delta_Y(h) - \eta_n\delta_Y(h) + (f(h) - 1)m_Y(h, -1), & f(h) > 1. \end{cases}$$

   Thus, $\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) \geq -\delta_Y(h) - \eta_n\delta_Y(h) = \widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) - \eta_n\delta_Y(h)$.

6. When $h \in \mathcal{M}^c$, $f_\tau^*(h) = -\eta_n$ and $f(h) \leq 0$, we have $\mathcal{V}(f_\tau^*, h) - \mathcal{V}(f, h) = 0$ and

$$\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) = \begin{cases} -f(h)\delta_Y(h) - \eta_n\delta_Y(h), & f(h) \geq -1 \\ (f(h)-1)m_Y(h,1) + (1-\eta_n)\delta_Y(h), & f(h) < -1. \end{cases}$$

Thus, we obtain $\mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) \geq -\eta_n\delta_Y(h) = \widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) - \eta_n\delta_Y(h)$.

Hence, by combining all these cases, we conclude that

$$\widetilde{\mathcal{V}}(f_\tau^*, h) - \widetilde{\mathcal{V}}(f, h) \leq \mathcal{V}_\phi(f_\tau^*, h) - \mathcal{V}_\phi(f, h) + M\eta_n$$

for any $\eta_n \in (0, 1]$ and any decision function f. The same argument holds for any h such that $\delta_Y(h) < 0$. Consequently, since $\mathcal{V}(f) = E[\widetilde{\mathcal{V}}(f, H)]$,we have

$$\mathcal{V}(f_\tau^*) - \mathcal{V}(f) \leq E[\mathcal{V}_\phi(f_\tau^*, H)] - E[\mathcal{V}_\phi(f, H)] + M\eta_n.$$

□

Lemma 2.4 *For any $\delta > 0$ and τ such that $[\tau - 2\delta, \tau + 2\delta] \subseteq (\tau_{\min}, \tau_{\max})$, $E[\mathcal{V}_\phi(f_\tau^*, H)]$, as a function of τ, is Lipschitz continuous at τ.*

Proof: Let $\tau_1 = \tau$ and τ_2 be any number in $[\tau - 2\delta, \tau + 2\delta]$. Without loss of generality, we assume $\tau_2 < \tau_1$. We also let f_1^* and f_2^* be the optimal decision functions solving (2.10) for τ_1 and τ_2, respectively, and their corresponding λ^*'s values are denoted as λ_1 and λ_2. According to (2.11), it is easy to verify that

$$\begin{aligned} & E[\mathcal{V}_\phi(f_1^*, H)] - E[\mathcal{V}_\phi(f_2^*, H)] \\ =& E\left[(1+\eta_n)\delta_Y(H)\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(\delta_Y(H) > 0)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ & - E\left[(1+\eta_n)\delta_Y(H)\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(\delta_Y(H) < 0)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ =& (1+\eta_n)E\left[|\delta_Y(H)|\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(H \in \mathcal{M}^c)\right]. \end{aligned}$$

On the other hand

$$\begin{aligned}\tau_1 - \tau_2 =& E[\mathcal{R}_\psi(f_1^*, \eta_n, H)\mathbb{I}(H \in \mathcal{M}^c)] - E[\mathcal{R}_\psi(f_2^*, \eta_n, H)\mathbb{I}(H \in \mathcal{M}^c)] \\ =& E\left[\delta_R(H)\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(\delta_Y(H) > 0)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ & - E\left[\delta_R(H)\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(\delta_Y(H) < 0)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ =& E\left[|\delta_R(H)|\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(H \in \mathcal{M}^c)\right].\end{aligned}$$

The above two equations imply that

$$\begin{aligned}& E[\mathcal{V}_\phi(f_1^*, H)] - E[\mathcal{V}_\phi(f_2^*, H)] \\ =& (1+\eta_n)E\left[\frac{|\delta_Y(H)|}{|\delta_R(H)|}|\delta_R(H)|\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ \leq& 2\lambda_2 E\left[|\delta_R(H)|\mathbb{I}\left(\lambda_1 \leq \frac{\delta_Y(H)}{\delta_R(H)} \leq \lambda_2\right)\mathbb{I}(H \in \mathcal{M}^c)\right] \\ \leq& 2\lambda_2(\tau_1 - \tau_2).\end{aligned}$$

The lemma holds since λ_2 is no larger than λ^*-value corresponding to $\tau - 2\delta$. □

2.9.1.2 Approximation bias in RKHS

In this section, we prove a series of lemmas to quantify the approximation bias of $\widehat{f}$, where $\widehat{f}$ denotes the solution of single stage empirical problem

$$\begin{aligned}&\arg\min_{f \in \mathcal{G}} \ \frac{1}{n}\sum_{i=1}^n Y_i \frac{\phi(A_i f(H_i))}{p(A_i|H_i)} + \lambda_n \|f\|_{\mathcal{G}}^2 \\ &\text{subject to } \frac{1}{n}\sum_{i=1}^n R_i \frac{\psi(A_i f(H_i), \eta_n)}{p(A_i|H_i)} \leq \tau,\end{aligned} \tag{2.15}$$

resulted from restricting $\widehat{f}$ to be a function in Gaussian RKHS $\mathcal{G}$.

The section is organized as follows: Lemma 2.5 provides an approximation of f_τ^* using functions in Gaussian RKHS; in Lemma 2.6, we quantify the difference of risk under shifted ramp loss between f_τ^* and its approximation in $\mathcal{G}$; in Lemma 2.7, we show that $\|\widehat{f}\|_{\mathcal{G}}$ is bounded with high probability; in Lemma 2.8, we show that $\mathcal{A}_n(\tau)$ changes continuously w.r.t. τ and the approximation bias is given Lemma 2.9.

For convenience, we define

$$\mathfrak{L}_\phi(f) = Y\frac{\phi(Af(H))}{p(A|H)}, \quad \mathbb{P}_n[\mathfrak{L}_\phi(f)] = \frac{1}{n}\sum_{i=1}^n Y_i\frac{\phi(A_i f(H_i))}{p(A_i|H_i)},$$

$$\mathfrak{R}_\psi(f,\eta_n) = R\frac{\psi(Af(H),\eta_n)}{p(A|H)}, \quad \mathbb{P}_n[\mathfrak{R}_\psi(f,\eta_n)] = \frac{1}{n}\sum_{i=1}^n R_i\frac{\psi(A_i f(H_i),\eta_n)}{p(A_i|H_i)},$$

where $\mathbb{P}_n$ denotes the empirical distribution. Let $\mathcal{G} = \mathcal{G}(\sigma_n)$ denote the Gaussian RKHS with bandwidth σ_n^{-1}, we define

$$\mathcal{A}(\tau) = \left\{ f \in \mathcal{G} \middle| E[\mathfrak{R}_\psi(f,\eta_n)] \leq \tau \right\},$$

$$\mathcal{A}_n(\tau) = \left\{ f \in \mathcal{G} \middle| \mathbb{P}_n[\mathfrak{R}_\psi(f,\eta_n)] \leq \tau \right\},$$

where $\mathcal{A}_n(\tau)$ is equivalent to the definition of the feasible region of the empirical problem with $T = 1$. We also define $\bar{\mathcal{H}} = 3\mathcal{H}$,

$$\bar{\delta}_Y(h) = \begin{cases} \delta_Y(h), & \text{if } |h| \leq 1 \\ \delta_Y(h/|h|), & \text{if } |h| > 1, \end{cases} \quad \bar{\delta}_R(h) = \begin{cases} \delta_R(h), & \text{if } |h| \leq 1 \\ \delta_R(h/|h|), & \text{if } |h| > 1, \end{cases}$$

and recall that in Assumption 2.5 we defined

$$\bar{H}_{a,b} = \left\{ h \in \bar{\mathcal{H}} : a\bar{\delta}_Y(h) > 0, b(\bar{\delta}_Y(h) - \lambda^*\bar{\delta}_R(h)) > 0 \right\}$$

and when $T = 1$ $\Delta_\tau(h) = \sum_{a,b\in\{-1,1\}} \text{dist}(h, \bar{\mathcal{H}}/\bar{H}_{a,b}) I(h \in \bar{H}_{a,b})$, where $a, b \in \{-1, 1\}$ and λ^* is the value in f_τ^* so function Δ depends on τ, and

$$\bar{f}_\tau(h) = \begin{cases} 1, & \text{if } h \in \bar{H}_{1,1} \\ \eta_n, & \text{if } h \in \bar{H}_{-1,1} \\ -1, & \text{if } h \in \bar{H}_{-1,-1} \\ -\eta_n, & \text{if } h \in \bar{H}_{1,-1} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $\bar{f}_\tau$ can be viewed as an extension of f_τ^* from of $\mathcal{H}$ to $\bar{\mathcal{H}}$. Our first lemma is to determine the pointwise approximation bias of f_τ^* using the RKHS. Note that we assumed $\mathcal{H}$ is a compact subset of $\mathcal{G}$, without loss of generality, from now on we assume that $\mathcal{H} \subseteq \mathcal{B}_\mathcal{G}$ where $\mathcal{B}_\mathcal{G}$ denotes the unit ball in $\mathcal{G}$.

Lemma 2.5 *Let $\check{f}_\tau = (\sigma_n^2/\pi)^{d/4} \bar{f}_\tau$ and define linear operator*

$$V_\sigma \check{f}(x) = \frac{(2\sigma)^{d/2}}{\pi^{d/4}} \int_{\mathbb{R}^d} e^{-2\sigma^2 \|x-y\|_2^2} \check{f}(y) dy.$$

Then, we have

$$\|V_{\sigma_n} \check{f}_\tau\|_\mathcal{G}^2 \leq c\sigma_n^d, \tag{2.16}$$

and

$$|V_{\sigma_n} \check{f}_\tau(h) - f_\tau^*(h)| \leq 8e^{-\sigma_n^2 \Delta_\tau(h)^2/2d}. \tag{2.17}$$

holds for all $h \in \mathcal{H}$, where $\Delta_\tau(h)$ is defined in Assumption 2.5 and c denotes a constant which depends on the dimension of feature space $\mathcal{H}$.

Remark 1 *Note that $V_\sigma \check{f}_\tau$ is an approximation of f_τ^* in $\mathcal{G}$. Thus, Lemma 2.5 quantifies the distance between the true optimal decision function and its approximation at each point h.*

Proof: Since $\mathcal{H} \subset \mathcal{B}_\mathcal{G}$ and $\check{f}_\tau = (\sigma_n^2/\pi)^{d/4} \bar{f}_\tau$, we can easily obtain that the L_2 norm of $\check{f}_\tau$ satisfies

$$\|\check{f}_\tau\|_2^2 \leq \text{Vol}(d)^2 \left(\frac{81}{\pi}\right)^{d/2} \sigma_n^d = c\sigma_n^d,$$

where $\text{Vol}(d)$ is the volume of $\mathcal{B}_\mathcal{G}$ (see equation (25) from Steinwart and Scovel (2007)) so c is a positive constant depends only on d. Moreover, it has been shown in Steinwart, Hush and Scovel (2006) that $V_\sigma : L^2(\mathbb{R}^d) \to \mathcal{G}(\sigma)$ is an isometric isomorphism and the inequality above implies $\|V_{\sigma_n} \check{f}_\tau\|_\mathcal{G}^2 = \|\check{f}_\tau\|_2^2 \leq c\sigma_n^d$.

We now start proving (2.17). By the construction of $\bar{f}_\tau$, it is straightforward to see that $\bar{f}_\tau(h) = f_\tau^*(h)$ for all $h \in \mathcal{H}$. Note for any $h \in H_{1,1}$ we have

$$\begin{aligned} V_{\sigma_n} \check{f}_\tau(h) =& \left(\frac{2\sigma_n^2}{\pi}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\sigma_n^2 \|h-y\|^2} \bar{f}_\tau(y) dy \\ =& \left(\frac{2\sigma_n^2}{\pi}\right)^{d/2} \left[\int_{B(h,\Delta_\tau(h))} e^{-2\sigma_n^2 \|h-y\|^2} \bar{f}_\tau(y) dy + \int_{\mathbb{R}^d / B(h,\Delta_\tau(h))} e^{-2\sigma_n^2 \|h-y\|^2} \bar{f}_\tau(y) dy \right], \end{aligned}$$

where $B(h,r)$ is the ball of radius r centering at h under Euclidean norm. By Lemma 4.1 in Steinwart and Scovel (2007), the construction of $\bar{f}_\tau$ guarantees that $B(h,\Delta_\tau(h)) \subseteq \bar{H}_{1,1}$ for all $h \in H_{1,1}$. It then follows that for any $h \in H_{1,1}$

$$\begin{aligned}
|V_{\sigma_n}\check{f}_\tau - f^*_\tau(h)| =& |V_{\sigma_n}\check{f}_\tau(h) - \bar{f}_\tau(h)| \\
=& \left| V_{\sigma_n}\check{f}_\tau(h) - \left(\frac{2\sigma_n^2}{\pi}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\sigma_n^2\|h-y\|^2} dy \right| \\
=& \left| \left(\frac{2\sigma_n^2}{\pi}\right)^{d/2} \int_{\mathbb{R}^d/B(h,\Delta_\tau(h))} e^{-2\sigma_n^2\|h-y\|^2} [\bar{f}_\tau(y) - 1] dy \right| \\
\leq& 2P(|U| \geq \Delta_\tau(h)),
\end{aligned}$$

where the last step uses the fact that $|\bar{f}_\tau - 1|_\infty \leq 2$, and U follows the spherical Gaussian distribution on $\mathbb{R}^d$ with parameter σ_n. Following inequality (3.5) from Ledoux and Talagrand (1991), we have

$$P(|U| \geq \Delta_\tau(h)) \leq 4e^{-\sigma_n^2\Delta_\tau^2(h)/2d}.$$

Similarly, we can obtain the same bound for $h \in \bar{H}_{-1,1}, \bar{H}_{1,-1}$ and $H_{-1,-1}$. As a conclusion, we have

$$|V_{\sigma_n}\check{f}_\tau(h) - f^*_\tau(h)| \leq 8e^{-\sigma_n^2\Delta_\tau^2(h)/2d}$$

for any $h \in \mathcal{H}$. □

In the next lemma, we show that under Assumption 2.5, the difference of the risk under shifted ramp loss between $f^*_{\tau_t}$ and its approximation $V_{\sigma_n}\check{f}_{\tau_1}$ is uniformly bounded by $O(\sigma_n^{-\alpha d}\eta_n^{-1})$ for any $\tau_1 \in [\tau - 2\delta_0, \tau + 2\delta_0]$; moreover when n is sufficiently large, $V_{\sigma_n}\check{f}_{\tau-2\delta_0}$ will belong to the empirical feasible region $\mathcal{A}_n(\tau)$ with high probability.

Lemma 2.6 *For any $\tau_1 \in [\tau - 2\delta_0, \tau + 2\delta_0]$,*

$$|E[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] - E[\mathfrak{R}_\psi(f^*_{\tau_1}, \eta_n)]| \leq c\sigma_n^{-\alpha d}\eta_n^{-1}, \tag{2.18}$$

where c is a constant depending on (α, d, K, M). Moreover, for any σ_n and η_n such that $c\sigma_n^{-\alpha d}\eta_n^{-1} \leq \delta_0$, with probability $1 - 2\exp\left(\frac{-2n\delta_0^2 c_1^2}{M^2}\right)$, we have $V_{\sigma_n}\check{f}_{\tau-2\delta_0} \in \mathcal{A}_n(\tau)$.

Proof: First note that for any measurable function $f_1, f_2 : \mathcal{H} \to \mathbb{R}$, we always have

$$\begin{aligned}
&|E[\mathfrak{R}_\psi(f_1, \eta_n)] - E[\mathfrak{R}_\psi(f_2, \eta_n)] \\
=&E\Big[E[R|H, A = 1][\psi(f_1(H), \eta_n) - \psi(f_2(H), \eta_n)] \\
&+ E[R|H, A = -1][\psi(-f_1(H), \eta_n) - \psi(-f_2(H), \eta_n)]\Big] \\
\leq& 2M\eta_n^{-1}E[|f_1(H) - f_2(H)|].
\end{aligned}$$

Using result (2.17) in Lemma 2.5 and Assumption 2.5, we can obtain

$$\begin{aligned}
|E[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] - E[\mathfrak{R}_\psi(f^*_{\tau_1}, \eta_n)]| \leq& \eta_n^{-1} 16ME[e^{-\sigma_n^2\Delta_\tau(h)^2/2d}] \\
\leq& 16MK(2d)^{\alpha d/2}\sigma_n^{-\alpha d}\eta_n^{-1} \\
=& c\sigma_n^{-\alpha d}\eta_n^{-1}.
\end{aligned}$$

To prove the remaining part of the lemma, suppose $\tau_1 = \tau - 2\delta_0$. We note that $\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)$ is bounded by M/c_1. Based on Hoeffding's inequality, we can obtain

$$P\Big[|\mathbb{P}_n[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] - E[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)]| \geq \delta_0\Big] \leq 2\exp\left(\frac{-2n\delta_0^2c_1^2}{M^2}\right). \tag{2.19}$$

According to (2.18) and the choice of (σ_n, η_n), we have

$$\begin{aligned}
&|E[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] - (\tau - 2\delta_0)| \\
=&|E[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] - E[\mathfrak{R}_\psi(f^*_1, \eta_n)]| \\
\leq& c\sigma_n^{-\alpha d}\eta_n^{-1} \leq \delta_0.
\end{aligned} \tag{2.20}$$

Combining (2.19) and (2.20), we obtain

$$P\Big[\mathbb{P}_n[\mathfrak{R}_\psi(V_{\sigma_n}\check{f}_{\tau_1}, \eta_n)] \geq \tau\Big] \leq 2\exp\left(\frac{-2n\delta_0^2c_1^2}{M^2}\right),$$

which implies that $V_{\sigma_n}\check{f}_{\tau_1} \in \mathcal{A}_n(\tau)$ with probability at least $1 - 2\exp\left(\frac{-2n\delta_0^2c_1^2}{M^2}\right)$. □

In Lemma 2.7, we show that $\|\widehat{f}\|_{\mathcal{G}}$ is bounded with high probability.

Lemma 2.7 *$\widehat{f}_\tau$ satisfies*

$$P\bigg(\|\widehat{f}_\tau\|_{\mathcal{G}}^2 \leq c\bigg(\frac{M}{c_1\lambda_n}+\sigma_n^d\bigg)\bigg) \geq 1-2\exp\bigg(\frac{-2n\delta_0^2c_1^2}{M^2}\bigg), \tag{2.21}$$

for any choice of $c\sigma_n^{-\alpha d}\eta_n^{-1} \leq \delta_0$. Here, the constant c in front of σ_n^d only depends on dimension d and the constant in front of $\sigma_n^{-\alpha d}\eta_n^{-1}$ is equal to the constants of the same term in Lemma 2.6.

Proof: From the last claim of Lemma 2.6, we have $V_{\sigma_n}\check{f}_{\tau-2\delta_0} \in \mathcal{A}_n(\tau)$ holds with probability at least $1-2\exp\big(\frac{-2n\delta_0^2c_1^2}{M^2}\big)$. Using and (2.16) of Lemma 2.5, under the choice of (σ_n,η_n) we have

$$\lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 \leq \mathbb{P}_n[\mathfrak{L}_\phi(\widehat{f})]+\lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 \leq \mathbb{P}_n[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-2\delta_0})]+\lambda_n\|V_{\sigma_n}\check{f}_{\tau-2\delta_0}\|_{\mathcal{G}}^2 \leq c\bigg(\frac{M}{c_1}+\lambda_n\sigma_n^d\bigg),$$

which gives (2.21). □

Lemma 2.7 implies that, instead of $\mathcal{A}(\tau)$ and $\mathcal{A}_n(\tau)$, we can concentrate on the sets given by

$$\mathcal{A}(\tau,\mathcal{C}_n)=\bigg\{f\in\mathcal{G}\bigg|\|f\|_{\mathcal{G}}\leq\mathcal{C}_n, E[\mathfrak{R}_\psi(f,\eta_n)]\leq\tau\bigg\},$$

$$\mathcal{A}_n(\tau,\mathcal{C}_n)=\bigg\{f\in\mathcal{G}\bigg|\|f\|_{\mathcal{G}}\leq\mathcal{C}_n, \mathbb{P}_n[\mathfrak{R}_\psi(f,\eta_n)]\leq\tau\bigg\},$$

where $\mathcal{C}_n = c\sqrt{\frac{M}{c_1\lambda_n}+\sigma_n^d}$. This is because $\widehat{f}$ belongs to them with a high probability.

We further study the relationships among $\mathcal{A}(\tau,\mathcal{C}_n)$ and $\mathcal{A}_n(\tau,\mathcal{C}_n)$. The proof will use a general covering number property for Gaussian RKHS from Steinwart and Scovel (2007), which is stated as Proposition 2.1 in Section 2.9.1.4.

Lemma 2.8 *For any $\delta>0$ with probability at least $1-\exp\big(-\frac{n\delta^2c_1^2}{2M^2}\big)$, we have*

$$\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)\subset\mathcal{A}_n(\tau,\mathcal{C}_n)\subset\mathcal{A}(\tau+\epsilon_n,\mathcal{C}_n), \tag{2.22}$$

where

$$\epsilon_n = c\sigma_n^{(1-\nu_2/2)(1+\theta_2)d/2}\bigg(\frac{M}{c_1\lambda_n}+\sigma_n^d\bigg)^{\nu_2/4}\eta_n^{-\nu_2/2}+\delta$$

for $0<\nu_2\leq 2$ and $\theta_2>0$. Moreover, let

$$\epsilon_n' = \epsilon_n + c\sigma_n^{-\alpha d}\eta_n^{-1},$$

then for any λ_n and σ_n such that $\epsilon_n' \leq 2\delta_0$, we have $V_{\sigma_n}\check{f}_{\tau-\epsilon_n'} \in \mathcal{A}(\tau-\epsilon_n, \mathcal{C}_n)$ and

$$|E[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-\epsilon_n'})] - E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon_n'})]| \leq c\sigma_n^{-\alpha d}.$$

Here, the constants in front of σ_n^{-ad} and $\sigma_n^{-ad}\eta_n^{-1}$ are equal to the constants in Lemma 2.7. c in front of ϵ_n is a constant only dependent on $(M, c_1, \nu_2, \theta_2, d)$.

Proof: To prove (2.22), it is sufficient to show that with probability $1-\exp\big(-\frac{n\delta^2 c_1^2}{2M^2}\big)$ we have

$$\sup_{f\in\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n)} |\mathbb{P}_n[f] - E[f]| \leq \epsilon_n, \tag{2.23}$$

where $\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n) = \{\mathfrak{R}_\psi(f,\eta_n)|f\in\mathcal{B}_\mathcal{G}(\mathcal{C}_n)\}$ and $\mathcal{B}_\mathcal{G}(\mathcal{C}_n)$ denotes the closed ball in $\mathcal{G}$ with radius $\mathcal{C}_n$. By Theorem 4.10 from Wainwright (2019), we have that

$$\sup_{f\in\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n)} |\mathbb{P}_n[f] - E[f]| \leq 2\text{Rad}_n(\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n)) + \delta$$

holds with probability $1-\exp\big(-\frac{n\delta^2c_1^2}{2M^2}\big)$, where $\text{Rad}_n(\mathcal{F})$ is the Rademacher complexity of some functional set $\mathcal{F}$ defined as

$$\text{Rad}_n(\mathcal{F}) = E_X E_\epsilon \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|, \quad \epsilon_i \sim i.i.d.\ P(\epsilon_i = \pm 1) = 0.5.$$

Following the proof in Example 5.24 from Wainwright (2019), by Dudley's entropy integral the Rademacher complexity is upper bound by

$$\begin{aligned}
\text{Rad}_n(\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n)) \leq & E\left[\frac{24}{\sqrt{n}}\int_0^{2\frac{M}{c_1}} \sqrt{\log\mathcal{N}(\mathfrak{R}_{\psi,\eta_n}\circ\mathcal{B}_\mathcal{G}(\mathcal{C}_n), \epsilon, L_2(\mathbb{P}_n))}d\epsilon\right] \\
\overset{(i)}{\leq} & E\left[\frac{24}{\sqrt{n}}\int_0^{2\frac{M}{c_1}} \sqrt{\log\mathcal{N}\left(\mathcal{B}_\mathcal{G}, \frac{\eta_n c_1}{M\mathcal{C}_n}\epsilon, L_2(\mathbb{P}_n)\right)}d\epsilon\right] \\
\overset{(ii)}{\leq} & c\sigma_n^{(1-\nu_2/2)(1+\theta_2)d/2}\left(\frac{M}{c_1\lambda_n} + c_d^2\sigma_n^d\right)^{\nu_2/4}\eta_n^{-\nu_2/2},
\end{aligned} \tag{2.24}$$

where to obtain (i) we have used the fact that $\mathfrak{R}_{\psi,\eta_n}$ is a Lipschitz function of f with Lipschitz constant $\frac{M}{c_1\eta_n}$, and in (ii) we used the covering number property of $\mathcal{B}_\mathcal{G}$ stated in Proposition 2.1.

For the second part of the lemma, $V_{\sigma_n}\check{f}_{\tau-\epsilon'_n} \in \mathcal{A}(\tau-\epsilon_n, \mathcal{C}_n)$ is a direct conclusion of (2.18) from Lemma 2.6 since

$$\begin{aligned}
&E[\mathfrak{R}(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}, \eta_n)] \\
\leq &|E[\mathfrak{R}(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}, \eta_n)] - E[\mathfrak{R}(V_{\sigma_n}f^*_{\tau-\epsilon'_n}, \eta_n)]| + |E[\mathfrak{R}(V_{\sigma_n}f^*_{\tau-\epsilon'_n}, \eta_n)]| \\
\leq &\tau - \epsilon'_n + c\sigma_n^{-\alpha d} = \tau - \epsilon_n.
\end{aligned}$$

Note that for any $f_1, f_2 : \mathcal{H} \to [-1, 1]$ we always have

$$\begin{aligned}
&|E[\mathfrak{L}_\phi(f_1)] - E[\mathfrak{L}_\phi(f_2)]| \\
=&E\Big[E[Y|H, A=1][\phi(f_1(H)) - \phi(f_2(H))] + E[Y|H, A=-1][\phi(-f_1(H)) - \phi(-f_2(H))]\Big] \\
=&E[\delta_Y(H)[f_1(H) - f_2(H)]] \\
\leq &E[|\delta_Y(H)||f_1(H) - f_2(H)|].
\end{aligned}$$

Hence, using (2.17) in Lemma 2.5 and Assumption 2.5 we have

$$\begin{aligned}
|E[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n})] - &E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon'_n})]| \leq \\
&8E[|\delta_Y(H)|e^{-\sigma_n^2\Delta^2_{\tau-\epsilon'_n}(h)/2d}] \leq 8MK(2d)^{\alpha d/2}\sigma_n^{-\alpha d} = c\sigma_n^{-\alpha d}.
\end{aligned}$$

This completes the proof for the lemma. □

As a corollary of Lemma 2.8, we can establish risk error bound (2.14) stated in Theorem 2.4 for $T = 1$. We stated this results as Corollary 2.1 below

Corollary 2.1 *Suppose (σ_n, η_n) satisfy the requirement in Lemma 2.7, then for any $0 < \nu_2 \leq 2$, $\theta_2 > 0$ and $\delta > 0$ with probability at least $1 - 2\exp\big(\frac{-2n\delta_0^2 c_1^2}{M^2}\big) - 2\exp\big(-\frac{n\delta^2 c_1^2}{2M^2}\big)$ we have*

$$E\left[\frac{R\mathbb{I}(A\widehat{f}(H_t) > 0)}{p(A|H)}\right] \leq \tau + \delta + cn^{-1/2}\sigma_n^{(1-\nu_2/2)(1+\theta_2)d/2}\left(\frac{M}{c_1\lambda_n} + \sigma_n^{d_t}\right)^{\nu_2/4}\eta_n^{-\nu 2/2}.$$

Here, c is a constant only depends on $(M, c_1, \nu_2, \theta_2, d)$.

Proof: Lemma 2.7 implies that $\widehat{f}$ is bounded by $\mathcal{C}_n$ with probability at least $1 - 2\exp\big(\frac{-2n\delta_0^2 c_1^2}{M^2}\big)$. Moreover, the concentration inequality (2.23) of Lemma 2.8 implies that

$$E[\mathfrak{R}_\psi(f,\eta_n)] - \mathbb{P}_n[\mathfrak{R}_\psi(f,\eta_n)] \leq \delta + cn^{-1/2}\sigma_n^{(1-\nu_2/2)(1+\theta_2)d/2}\left(\frac{M}{c_1\lambda_n} + \sigma_n^{d_t}\right)^{\nu_2/4}\eta_n^{-\nu 2/2}$$

holds with probability at least $1 - 2\exp\big(-\frac{n\delta^2 c_1^2}{2M^2}\big)$ for any $\delta > 0$ and $f \in \mathcal{B}_{\mathcal{G}}(\mathcal{C}_n)$. The result holds since $\mathbb{P}_n[\mathfrak{R}_\psi(\widehat{f},\eta_n)] \leq \tau$ by definition and note that

$$E\left[\frac{R\mathbb{I}(A\widehat{f}(H) > 0)}{p(A|H)}\right] \leq E[\mathfrak{R}_\psi(\widehat{f},\eta_n)].$$

□

Lemma 2.8 indicates that $V_{\sigma_n}\check{f}_{\tau-\epsilon'_n} \in \mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n) \subseteq \mathcal{A}_n(\tau,\mathcal{C}_n)$ holds with high probability. In Lemma 2.9, we will show that $V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}$ can be used to quantify the approximation bias caused by RKHS.

Lemma 2.9 *Under the condition of Lemma 2.8, we have*

$$\inf_{f\in\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2 - E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon'_n})]) \leq \xi_n^{(2)}.$$

Proof: Let $\check{f}_{\tau-\epsilon'_n} = (\sigma_n^2/\pi)^{d/4}\bar{f}_{\tau-\epsilon'_n}$, then from Lemma 2.8 we have

$$|E[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n})] - E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon'_n})]| \leq c\sigma_n^{-\alpha d},$$

and $V_{\sigma_n}\check{f}_{\tau-\epsilon'_n} \in \mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)$. Moreover, (2.16) from Lemma 2.5 gives that $\|V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}\|\|_{\mathcal{G}}^2 \leq c\sigma_n^d$. Hence, we have

$$\begin{aligned}
&\inf_{f\in\mathcal{A}(\tau-\epsilon_n)}[E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2 - E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon'_n})]] \\
=&\inf_{f\in\mathcal{A}(\tau-\epsilon_n)}[E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2 - E[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n})] - \lambda_n\|V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}\|_{\mathcal{G}}^2] + \lambda_n\|V_{\sigma_n}\check{f}_{\tau-\epsilon'_n}\|_{\mathcal{G}}^2 \\
&+ E[\mathfrak{L}_\phi(V_{\sigma_n}\check{f}_{\tau-\epsilon'_n})] - E[\mathfrak{L}_\phi(f^*_{\tau-\epsilon'_n})] \\
\leq& c(\lambda_n\sigma_n^d + \sigma_n^{-\alpha d}) \equiv \xi_n^{(2)}.
\end{aligned}$$

□

2.9.1.3 Completing the proof of Theorem 2.2 (Theorem 2.4) for $T = 1$

We first establish the error bound for excessive risk (2.13). Since the Fisher consistency of Theorem 2.3 indicates $\mathcal{V}(g^*) = \mathcal{V}(f_\tau^*)$ and using the excessive risk inequality in Lemma 2.3 we have

$$\mathcal{V}(f_\tau^*) - \mathcal{V}(\widehat{f}) \leq E[\mathcal{V}_\phi(f_\tau^*, H)] - E[\mathcal{V}_\phi(\widehat{f}, H)] + M\eta_n,$$

the proof is completed if we can show

$$E[\mathcal{V}_\phi(f_\tau^*, H)] - E[\mathcal{V}_\phi(\widehat{f}, H)] = E[\mathfrak{L}_\phi(\widehat{f})] - E[\mathfrak{L}_\phi(f_\tau^*)] \leq \xi_n + 2\lambda_0\epsilon_n' = \xi_n + c\epsilon_n' \tag{2.25}$$

holds with probability at least $1 - h(n, x)$, where λ_0 denotes the λ^*-value for $(\tau - 2\delta_0)$ which is a constant only depends on (τ, δ_0).

According to Lemma 2.7, we have shown that $\|\widehat{f}\|_{\mathcal{G}}$ is bounded by $\mathcal{C}_n = c\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d}$ with probability least $1 - 2\exp\left(\frac{-2n\delta_0^2 c_1^2}{M^2}\right)$. Hence, similar to proof of Corollary 2.1, we can restrict to set $\mathcal{B}_{\mathcal{G}}(\mathcal{C}_n)$, and replace $\mathcal{A}(\tau)$ and $\mathcal{A}_n(\tau)$ by $\mathcal{A}(\tau, \mathcal{C}_n)$ and $\mathcal{A}_n(\tau, \mathcal{C}_n)$ correspondingly.

To prove (2.25), we note that the left-hand side of the inequality can be composed as

$$\begin{aligned}
&E[\mathfrak{L}_\phi(\widehat{f})] - E[\mathfrak{L}_\phi(f_\tau^*)] \\
\leq &E[\mathfrak{L}_\phi(\widehat{f})] + \lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 - E[\mathfrak{L}_\phi(f_\tau^*)] \\
\leq &E[\mathfrak{L}_\phi(\widehat{f})] + \lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 - \inf_{f\in\mathcal{A}_n(\tau,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2) \\
&+ \inf_{f\in\mathcal{A}_n(\tau,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2) - E[\mathfrak{L}_\phi(f_\tau^*)] \\
\leq &\underbrace{E[\mathfrak{L}_\phi(\widehat{f})] + \lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 - \inf_{f\in\mathcal{A}_n(\tau,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2)}_{(I)} \\
&+ \underbrace{\inf_{f\in\mathcal{A}_n(\tau,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2) - \inf_{f\in\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2)}_{(II)} \\
&+ \underbrace{\inf_{f\in\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2) - E[\mathcal{L}_\phi(f_{\tau-\epsilon_n'}^*)]}_{(III)} + \underbrace{E[\mathfrak{L}_\phi(f_{\tau-\epsilon_n'}^*)] - E[\mathfrak{L}_\phi(f_\tau^*)]}_{(IV)}.
\end{aligned} \tag{2.26}$$

Using the inclusion result from Lemma 2.8, we have

$$\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)\subseteq\mathcal{A}_n(\tau,\mathcal{C}_n)\subseteq\mathcal{A}(\tau+\epsilon_n,\mathcal{C}_n)$$

holds with probability no more than $2\exp\big(-\frac{n\delta_0^2c_1^2}{2M^2}\big)$, and $\mathcal{A}(\tau-\epsilon_n,\mathcal{C}_n)\subseteq\mathcal{A}_n(\tau,\mathcal{C}_n)$ implies $(II)<0$. In addition, $\mathcal{A}_n(\tau,\mathcal{C}_n)\subseteq\mathcal{A}(\tau+\epsilon_n,\mathcal{C}_n)$ implies that

$$\begin{aligned}
&E[\mathfrak{L}_\phi(\widehat{f})]-E[\mathfrak{L}_\phi(f_\tau^*)]\\
\geq&E[\mathfrak{L}_\phi(\widehat{f})]-E[\mathfrak{L}_\phi(f_{\tau+\epsilon_n'}^*)]+E[\mathfrak{L}_\phi(f_{\tau+\epsilon_n'}^*)]-E[\mathfrak{L}_\phi(f_\tau^*)]\\
\geq&E[\mathfrak{L}_\phi(f_{\tau+\epsilon_n'}^*)]-E[\mathfrak{L}_\phi(f_\tau^*)].
\end{aligned}$$

which provides a lower bound for the difference of the surrogate reward between $\widehat{f}$ and f_τ^*. Using the approximation bias obtained in Lemma 2.9, (III) is bounded by $\xi_n^{(2)}$. For term (IV), using the Lipschitz continuity property of the value function obtained in Lemma 2.4 we have

$$|E[\mathfrak{L}_\phi(f_{\tau-\epsilon_n'}^*)]-E[\mathfrak{L}_\phi(f_\tau^*)]|\leq 2\lambda_0\epsilon_n'.$$

Hence, it remains to derive the bound for (I). To this end, we define

$$\widetilde{f}_\tau=\underset{f\in\mathcal{A}(\tau)}{arg\min}\,E[\mathfrak{L}_\phi(f)]+\lambda_n\|f\|_{\mathcal{G}}^2, \tag{2.27}$$

and apply Proposition 2.2 for $\mathcal{L}(f)=\mathfrak{L}_\phi(f)+\lambda_n\|f\|_{\mathcal{G}}^2$,

$$\mathcal{W}_{(X_1,\ldots,X_n)}=\{\mathcal{L}(f)-\mathcal{L}(\widetilde{f}_\tau)|f\in\mathcal{A}_n(\tau,\mathcal{C}_n)\},$$

and

$$\mathcal{W}=\{\mathcal{L}(f)-\mathcal{L}(\widetilde{f}_\tau)|f\in\mathcal{B}_{\mathcal{G}}(\mathcal{C}_n)\}.$$

By similar argument used in Lemma 2.7, we can show that $\|\widetilde{f}\|_{\mathcal{G}}\leq\mathcal{C}_n$ for any choice of $c\sigma_n^{-\alpha d}\eta_n^{-1}\leq\delta_0$ and consequently we can replace $\mathcal{A}(\tau)$ by $\mathcal{A}(\tau,\mathcal{C}_n)$ in (2.27).

Proposition 2.2 requires $\|w\|_\infty$ is uniformly bounded by some constant B for any $w\in\mathcal{W}$ and the ϵ-covering number of $\mathcal{N}(B^{-1}\mathcal{W},\epsilon,L_2(\mathbb{P}_n))$ is uniformly bounded with polynomial order of ϵ^{-1}. To verify

the first condition, using the property of Gaussian RKHS we have $\|f\|_\infty \leq \|f\|_\mathcal{G}$ for all $f \in \mathcal{B}_\mathcal{G}(\mathcal{C}_n)$ and it follows that

$$\begin{aligned}\|\mathcal{L}(f) - \mathcal{L}(\widetilde{f}_\tau)\|_\infty &\leq \frac{M}{c_1}\|f - \widetilde{f}_\tau\|_\infty + \lambda_n\|\widetilde{f}_\tau\|_\mathcal{G}^2 + \lambda_n\|f\|_\mathcal{G}^2 \\ &\leq \frac{2cM}{c_1}\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d} + 2c^2\lambda_n\left(\frac{M}{c_1\lambda_n} + \sigma_n^d\right) = B,\end{aligned}$$

which gives the choice of B. Moreover, from the sub-additivity of the entropy we have

$$\begin{aligned}\log\mathcal{N}(B^{-1}\mathcal{W}, 2\epsilon, L_2(\mathbb{P}_n)) \leq &\underbrace{\log\mathcal{N}(B^{-1}\{\mathfrak{L}_\phi(f) : f \in \mathcal{B}_\mathcal{G}(\mathcal{C}_n), \epsilon, L_2(\mathbb{P}_n)\})}_{(V)} \\ &+ \underbrace{\log\mathcal{N}(B^{-1}\{\lambda_n\|f\|_\mathcal{G}^2 : f \in \mathcal{B}_\mathcal{G}(\mathcal{C}_n), \epsilon, L_2(\mathbb{P}_n)\})}_{(VI)}.\end{aligned}$$

For (V) we have

$$\begin{aligned}(V) &\leq \log\mathcal{N}(\mathcal{B}_\mathcal{G}(\mathcal{C}_n), \frac{c_1B\epsilon}{M}, L_2(\mathbb{P}_n)) \\ &= \log\mathcal{N}(\mathcal{B}_\mathcal{G}, \frac{c_1B\epsilon}{M}\left(c\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d}\right)^{-1}, L_2(\mathbb{P}_n)) \\ &\leq \log\mathcal{N}(\mathcal{B}_\mathcal{G}, 2\epsilon, L_2(\mathbb{P}_n)),\end{aligned}$$

since $\mathfrak{L}$ is a 1-Lipschitz function of f and $\frac{c_1B\epsilon}{M}\left(c\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d}\right)^{-1} \geq 2$ for sufficient small λ_n and large σ_n. For (VI) we have

$$(VI) \leq \log\left(c\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d}\Big/(B\epsilon)\right) = \log\left(c\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d}\Big/B\right) - \log\epsilon \leq -\log\epsilon.$$

Combining the upper bound of (V) and (VI) and using the covering number property for $\log\mathcal{N}(\mathcal{B}_\mathcal{G}, \epsilon, L_2(\mathbb{P}_n))$ given in Proposition 2.1, we have

$$\begin{aligned}\sup_{\mathbb{P}_n}\log\mathcal{N}(B^{-1}\mathcal{W}, 2\epsilon, L_2(\mathbb{P}_n)) \leq &\sup_{\mathbb{P}_n}\log\mathcal{N}(\mathcal{B}_\mathcal{G}, 2\epsilon, L_2(\mathbb{P}_n)) - \log\epsilon \\ \leq &c\epsilon^{-\nu_1},\end{aligned}$$

for any $\sigma_n > 0$, $0 < \nu_1 < 2$, $\theta_1 > 0$, $\epsilon > 0$ and some positive constant c which only depends on (ν_1, θ_1, d). This implies that

$$\sup_{\mathbb{P}_n} \log \mathcal{N}(B^{-1}\mathcal{W}, \epsilon, L_2(\mathbb{P}_n)) \leq c\sigma_n^{(1-\nu_1/2)(1+\theta_1)d}\epsilon^{-\nu_1}.$$

Therefore, let $B = \frac{2cM}{c_1}\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d} + 2c^2\lambda_n\left(\frac{M}{c_1\lambda_n} + \sigma_n^d\right)$ and $l = c\sigma_n^{(1-\nu_1/2)(1+\theta_1)d}$, Proposition 2.2 implies that

$$P^*([E[\mathfrak{L}_\phi(\widehat{f})] + \lambda_n\|\widehat{f}\|_{\mathcal{G}}^2 - \inf_{f\in\mathcal{A}_n(\tau,\mathcal{C}_n)}(E[\mathfrak{L}_\phi(f)] + \lambda_n\|f\|_{\mathcal{G}}^2)] \geq \xi_n^{(1)}) \leq e^{-x},$$

where

$$\xi_n^{(1)} = c\left(\frac{2M}{c_1}\sqrt{\frac{M}{c_1\lambda_n} + \sigma_n^d} + 2\lambda_n\left(\frac{M}{c_1\lambda_n} + \sigma_n^d\right)\right)n^{-1/2}(\sigma_n^{(1-\nu_1/2)(1+\theta_1)d/2} + 2\sqrt{2x} + 2x/\sqrt{n})$$

for some positive c only depends on (ν_1, θ_1, d). The risk inequality (2.14) is guaranteed by Corollary 2.1 and this completes the proof for $T = 1$.

2.9.1.4 Statement of Propositions

In this section, we give the complete statement of all general propositions used for establishing Theorem 2.4. The first proposition states that the ϵ-covering number of $\mathcal{B}_\mathcal{G}$ under $L_2(\mathbb{P}_n)$ is uniformly with polynomial order in terms of σ_n and ϵ. This result was first established as Theorem 2.1 in Steinwart, Hush and Scovel (2006).

Proposition 2.1 *(Steinwart and Scovel (2007, Theorem 2.1)) For any $\epsilon > 0$, we have*

$$\sup_{\mathbb{P}_n} \log \mathcal{N}(\mathcal{B}_\mathcal{G}, \epsilon, L_2(\mathbb{P}_n)) \leq c\sigma_n^{(1-\nu/2)(1+\theta)d}\epsilon^{-\nu}$$

for any $0 < \nu \leq 2$ and $\theta > 0$. Here, $\mathcal{B}_\mathcal{G}$ is the closed unit ball in $\mathcal{G}$ w.r.t. $\|\cdot\|_\mathcal{G}$ and $\mathcal{N}(\cdot, \epsilon, L_2(\mathbb{P}_n))$ is the covering number of ϵ-ball w.r.t. empirical $L_2(\mathbb{P}_n)$ norm

$$\|f\|_{L_2(\mathbb{P}_n)} = \left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right)^{1/2}.$$

c is a constant only depends on (ν, θ, d).

Proposition 2.2 quantifies the stochastic error of $\widehat{f}$. The proof of Proposition 2.2 relies on first verifying Proposition 2.3 which is a weaker generalization of Theorem 5.6 of Steinwart and Scovel (2007) to allow the range of $\widehat{f}$ to be a subset depending on the sample. We give the proof of Proposition 2.3 after proving Proposition 2.2, and two preliminary results used to establish Proposition 2.3 are stated as Proposition 2.4 and Proposition 2.5 at the end of this section.

To state Proposition 2.2, we use $\mathcal{L}$ to denote any loss function from space $\mathcal{Z} \times \mathcal{F}$ to $\mathbb{R}$ and P to denote the probability distribution on $\mathcal{Z}$. In this section, we abuse the notation and use f^* and $\widehat{f}$ to denote the true and empirical minimizer under loss function $\mathcal{L}$, i.e.,

$$f^* = \arg\min_{f \in \mathcal{F}} E[\mathcal{L}(f)], \quad \widehat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{P}_n[\mathcal{L}(f)].$$

Proposition 2.2 *(Bounding Stochastic Error) Let $\mathcal{L}$ be a loss function from space $\mathcal{Z} \times \mathcal{F}$ to $\mathbb{R}$. Let P be a probability measure on $\mathcal{Z}$ and $\mathcal{F}$ be a set of bounded measurable functions from $\mathcal{Z}$ to $\mathbb{R}$. Suppose that we have a set of functional sets $\{F_{(z_1,...,z_n)}\}_{(z_1,...,z_n)\in\mathcal{Z}^n}$ and for any index $(z_1, ..., z_n)$ we have $\mathcal{F}_{(z_1,...,z_n)} \subset \mathcal{F}$. Let*

$$\mathcal{W} = \{\mathcal{L}(f) - \mathcal{L}(f^*) | f \in \mathcal{F}\}.$$

If there exists $B > 0$ such that $\|w\|_\infty \leq B$ for all $w \in \mathcal{W}$ and $\mathcal{W}$ is separable w.r.t. $\|\cdot\|_\infty$. Moreover, there are constants $n \geq l \geq 1$ and $0 < p < 2$ such that

$$\sup_{\mathbb{P}_n} \log \mathcal{N}(B^{-1}\mathcal{W}, \epsilon, L_2(\mathbb{P}_n)) \leq l\epsilon^{-p}$$

for any $\epsilon > 0$. Then there exists $c > 0$ depending only on p such that for any $n \geq 1$, $h \geq 1$ we have

$$P^*(E[\mathcal{L}(\widehat{f})] > E[\mathcal{L}(f^*)] + c\zeta_n(l, B, h)) \leq e^{-h},$$

where

$$\zeta_n(l, B, h) = 6cB\left(\frac{l}{n}\right)^{1/2} + 2\sqrt{2}B\sqrt{\frac{h}{n}} + 2B\frac{h}{n}.$$

Proof: Applying Proposition 2.3 to set $\mathcal{W} = \mathcal{F}$ and

$$\mathcal{W}_{(z_1,...,z_n)} = \{\mathcal{L}(f) - \mathcal{L}(f^*) | f \in \mathcal{F}_{(z_1,...,z_n)}\},$$

we have

$$\begin{aligned} P^*(E[\mathcal{L}(\widehat{f})] >& E[\mathcal{L}(f^*)] + c\zeta_n) \\ &\leq P^*(\exists w \in \mathcal{W}_{(Z_1,...,Z_n)} \; with \; \mathbb{P}_n(w) \leq 0 \; we \; have \; E(w) \geq \zeta_n) \\ &\leq e^{-h}, \end{aligned}$$

where $\zeta_n = 3E \sup_{s\in\mathcal{S}} \mathbb{P}_n(s) + 2\sqrt{2}B\sqrt{\frac{h}{n}} + 2B\frac{h}{n}$ and $\mathcal{S}$ are defined in statement of Proposition 2.3. Hence, the result is proved if we can show that under additional covering number assumption, ζ_n is upper bounded by $\zeta_n(l, B, h)$. To give an upper bound of $E \sup_{s\in\mathcal{S}} \mathbb{P}_n(s)$, it is worth noticing that by the definition of $\mathcal{S}$ in Proposition 2.3 we have

$$E \sup_{s\in\mathcal{S}} \mathbb{P}_n(s) \leq E \sup_{w\in\mathcal{W}, E(w^2)\leq B^2} |E(w) - \mathbb{P}_n(w)| = \omega_n(\mathcal{W}, B^2),$$

where $\omega_n(\mathcal{W}, \xi)$ is the modulus of the continuity of $\mathcal{W}$. Define the local Rademacher complexity of $\mathcal{W}$ to be

$$\text{Rad}_n(\mathcal{W}, \xi) = E_Z \, E_\epsilon \sup_{w\in\mathcal{W}, E(w^2)\leq\xi} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i) \right|,$$

where $\{\epsilon_i\}$ are n i.i.d. Rademacher random variables. According to van der Vaart and Wellner (1996), we have

$$\omega_n(\mathcal{W}, \xi) \leq 2\text{Rad}_n(\mathcal{W}, \xi).$$

Using the property that $\forall r > 0$

$$\text{Rad}_n(r\mathcal{W}, \xi) = r\text{Rad}_n(\mathcal{W}, r^{-2}\xi)$$

and applying Proposition 5.5 of Steinwart and Scovel (2007), which is stated as Proposition 2.4, under the assumption on the covering number of $\mathcal{W}$, we have

$$E \sup_{s\in\mathcal{S}} \mathbb{P}_n(s) \leq \omega_n(\mathcal{W}, B^2) \leq 2\text{Rad}_n(\mathcal{W}, B^2) \leq 2B\text{Rad}_n(B^{-1}\mathcal{W}, 1) \leq 2cB\left(\frac{l}{n}\right)^{1/2}.$$

□

Proposition 2.3 *Let P be a probability measure on $\mathcal{Z}$ and $\mathcal{W}$ be a set of bounded measurable functions from $\mathcal{Z}$ to $\mathbb{R}$. Suppose that we have a set of functional sets $\{\mathcal{W}_{(z_1,...,z_n)}\}_{(z_1,...,z_n)\in\mathcal{Z}^n}$ and for any index $(z_1, ..., z_n)$*

we have $\mathcal{W}_{(z_1,...,z_n)} \subset \mathcal{W}$. *Let* $(Z_1, ..., Z_n)$ *be* n *i.i.d. sample drawn from* P. *Suppose that* $\mathcal{W}$ *is separable w.r.t.* $\|\cdot\|_\infty$ *and* $\|w\|_\infty \leq B < \infty$ *for all* $w \in \mathcal{W}$. *Let* $\mathcal{S} = \{E(w) - w : w \in \mathcal{W}\}$. *Then for all* $n \geq 1$, $h \geq 1$ *and*

$$\zeta_n = 3E \sup_{s\in\mathcal{S}} \mathbb{P}_n(s) + 2\sqrt{2}B\sqrt{\frac{h}{n}} + 2B\frac{h}{n},$$

we have

$$P^*(for\ all\ w \in \mathcal{W}_{(Z_1,...,Z_n)}\ with\ \mathbb{P}_n(w) \leq 0\ we\ have\ E(w) \leq \zeta_n) \geq 1 - e^{-h}.$$

Proof: Let us define $\mathcal{S}_{(z_1,...,z_n)} = \{E(w) - w : w \in \mathcal{W}_{(z_1,...,z_n)}\}$, by the assumption of $\mathcal{W}$ it is obvious that $\mathcal{S}_{(z_1,...,z_n)} \subset \mathcal{S}$ with $E(s) = 0$, $\|s\|_\infty \leq 2B$ and $E(s^2) \leq 4B^2$ for all $s \in \mathcal{S}$. Moreover, it is also easy to verify that $\mathcal{S}$ is separable w.r.t. $\|\cdot\|_\infty$ given $\mathcal{W}$ is separable w.r.t. $\|\cdot\|_\infty$. Note that

$$\begin{aligned}
&P^*(\exists w \in \mathcal{W}_{(Z_1,...,Z_n)}\ with\ \mathbb{P}_n(w) \leq 0\ we\ have\ E(w) \geq \zeta_n) \\
\leq &P^*(\exists w \in \mathcal{W}_{(Z_1,...,Z_n)}\ with\ E(w) - \mathbb{P}_n(w) \geq \zeta_n) \\
\leq &P^n(\sup_{s\in\mathcal{S}_{(Z_1,...,Z_n)}} \mathbb{P}_n(s) \geq \zeta_n) \\
\leq &P^n(\sup_{s\in\mathcal{S}} \mathbb{P}_n(s) \geq \zeta_n).
\end{aligned}$$

Using Theorem 5.3 from Steinwart and Scovel (2007), which is stated as Proposition 2.5, with $b = 2B$ and $\iota = 4B^2$, we have

$$P^n(\sup_{s\in\mathcal{S}} \mathbb{P}_n(s) \geq \zeta_n) \leq P^n\left(\sup_{s\in\mathcal{S}} \mathbb{P}_n(s) \geq 3E\sup_{s\in\mathcal{S}} \mathbb{P}_n(s) + 2\sqrt{2}B\sqrt{\frac{h}{n}} + 2B\frac{h}{n}\right) \leq e^{-h}.$$

□

Proposition 2.4 *(Steinwart and Scovel (2007, Proposition 5.5)) Let* $\mathcal{W}$ *be a class of measurable functions from* Z *to* $[-1, 1]$ *which is separable w.r.t.* $\|\cdot\|_\infty$ *and let* P *be a probability measure on* Z. *Assume that there are constants* $q > 0$ *and* $0 < p < 2$ *with* $\sup_{\mathbb{P}_n} \log N(\mathcal{W}, \epsilon, L_2(\mathbb{P}_n)) \leq q\epsilon^{-p}$ *for all* $\epsilon > 0$. *Then there exists a constant* c *depending only on* p *such that for all* $n \geq 1$ *and all* $\epsilon > 0$ *we have*

$$Rad_n(\mathcal{W}, \epsilon) \leq c\left\{\epsilon^{1/2-p/4}\left(\frac{q}{n}\right)^{1/2}, \left(\frac{q}{n}\right)^{2/(2+p)}\right\}.$$

Proposition 2.5 *(Steinwart and Scovel (2007, Theorem 5.3)) Let P be a probability measure on Z and $\mathcal{W}$ be a set of bounded measurable functions from Z to $\mathbb{R}$ which is separable w.r.t. $\|\cdot\|_\infty$ and satisfies $E(w) = 0$ for all $w \in \mathcal{W}$. Furthermore, let $b > 0$ and $\iota \geq 0$ be constants with $\|w\|_\infty \leq b$ and $E(w^2) \leq \iota$ for all $w \in \mathcal{W}$. Then for all $x \geq 1$ and all $n \geq 1$ we have*

$$P^n\left(\sup_{w\in\mathcal{W}} \mathbb{P}_n(w) > 3E \sup_{w\in\mathcal{W}} \mathbb{P}_n(w) + \sqrt{\frac{2x\iota}{n}} + \frac{bx}{n}\right) \leq e^{-x}.$$

2.9.2 Proof of Theorem 2.2 (Theorem 2.4) for $T \geq 2$

We first prove (2.13) is Theorem 2.4. To this end, we define

$$\mathfrak{L}_{\phi,t}(f_t; f_{t+1}, ..., f_T) = E\left[\frac{(Y_t + U_{t+1}(H_{t+1}; f_{t+1}, ..., f_T))}{p(A_t|H_t)}\phi(A_t f_t(H_t))\right],$$

and

$$\widetilde{V}_t = \sup_{f_t \in \mathcal{A}_t(\tau_t)} \mathcal{V}_t(f_t, \widehat{f}_{t+1}, ..., \widehat{f}_T), \tag{2.28}$$

where

$$\mathcal{V}_t(g_t, ..., g_T) = E\left[\frac{(\sum_{j=t}^T Y_j)\prod_{j=t}^T \mathbb{I}(A_j g_j(H_j) > 0)}{\prod_{j=t}^T p(A_j|H_j)}\right].$$

Note that the Fisher consistency in Theorem 2.1 indicates that $\mathcal{V}_t(g_t^*, ..., g_T^*) = \mathcal{V}_t(f_t^*, ..., f_T^*)$, and it is equivalent to derive an upper bound for $\mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T)$.

First, note that by repeating the same argument for $T = 1$, we can show $\|\widehat{f}_t\|_{\mathcal{G}_t}$ is bounded by $\mathcal{C}_{n,t} = c\sqrt{\frac{(T-1+t)M}{c_1\lambda_{n,t}} + \sigma_{n,t}^{d_t}}$ with probability at least $1 - 2\exp\left(-\frac{2n\delta_{0,t}^2 c_1^2}{(T-t+1)^2M^2}\right)$ for any $t = 1, .., T$. Hence, we can replace $\mathcal{A}_t(\tau_t)$ in (2.28) by $\mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})$ and obtain

$$\begin{aligned}
&\mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T)\\
=&\mathcal{V}_t(f_t^*, ..., f_T^*) - \widetilde{V}_t + \widetilde{V}_t - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T)\\
\leq&\underbrace{\mathcal{V}_t(f_t^*, ..., f_T^*) - \widetilde{V}_t}_{(I)} + \underbrace{\mathfrak{L}_{\phi,t}(\widehat{f}_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) - \inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T)}_{(II)} + (T-t+1)M\eta_{n,t},
\end{aligned}$$

where

$$\mathcal{A}_t(\tau_t, \mathcal{C}_{n,t}) = \left\{f \in \mathcal{G}_t \middle| \|f\|_{\mathcal{G}_t} \leq \mathcal{C}_{n,t}, E\left[\frac{R_t\psi(A_t f(H_t), \eta_{n,t})}{p(A_t|H_t)}\right] \leq \tau_t\right\},$$

and to obtain the last inequality we have used the fact that $|Q_t|_\infty \leq (T-t+1)M$ and the excessive risk inequality of Lemma 2.3 to replace the difference under 0-1 loss in terms of $\mathcal{V}_t$ by the difference under hinge loss in terms of $\mathfrak{L}_{\phi,t}$.

For (I), we have

$$\begin{aligned}
(I) \leq & \mathcal{V}_t(f_t^*, ..., f_T^*) - \sup_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} [\mathcal{V}_t(f_t, \widehat{f}_{t+1}, ..., \widehat{f}_T) - \mathcal{V}_t(f_t, f_{t+1}^* ..., f_T^*) + \mathcal{V}_t(f_t, f_{t+1}^* ..., f_T^*)] \\
\leq & \mathcal{V}_t(f_t^*, ..., f_T^*) - \sup_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathcal{V}_t(f_t, f_{t+1}^*, ..., f_t^*) + c_1^{-1} |\mathcal{V}_{t+1}(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
= & c_1^{-1} |\mathcal{V}_{t+1}(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| + (T-t+1) M \eta_{n,t}, \\
& + \underbrace{\inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; f_{t+1}^*, ..., f_T^*) - \mathfrak{L}_{\phi,t}(f_t^*; f_{t+1}^*, ..., f_T^*)}_{(III)}
\end{aligned}$$

where again we have used the fact that $|Q_t|_\infty \leq (T-t+1)M$ and the excessive risk inequality of Lemma 2.3. To bound the last term in (I), let $f_{t,\tau'}^*$ denotes the solution of (2.3) of t by replace the risk constraint from τ_t to τ'. Then the second part of Lemma 2.8 indicates that $V_{\sigma_{n,t}} \check{f}_{t,\tau_t - \epsilon'_{n,t}} \in \mathcal{A}_t(\tau_t - \epsilon_{n,t}, \mathcal{C}_{n,t}) \subseteq \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})$, where $\check{f}_{t,\tau_t - \epsilon'_{n,t}} = (\sigma_{n,t}^2/\pi) \bar{f}_{t,\tau_t - \epsilon'_{n,t}}$, and

$$|\mathfrak{L}_{\phi,t}(V_{\sigma_{n,t}} \check{f}_{t,\tau_t - \epsilon'_{n,t}}; f_{t+1}^*, ..., f_T^*) - \mathfrak{L}_{\phi,t}(f_{t,\tau_t - \epsilon'_{n,t}}^*; f_{t+1}^*, ..., f_T^*)| \leq c \sigma_{n,t}^{-\alpha_t d_t}.$$

Therefore, we have

$$\begin{aligned}
(III) \leq & \mathfrak{L}_{\phi,t}(V_{\sigma_{n,t}} \check{f}_{t,\tau_t - \epsilon'_{n,t}}; f_{t+1}^*, ..., f_T^*) - \mathfrak{L}_{\phi,t}(f_t^*; f_{t+1}^*, ..., f_T^*) \\
\leq & |\mathfrak{L}_{\phi,t}(V_{\sigma_{n,t}} \check{f}_{t,\tau_t - \epsilon'_{n,t}}; f_{t+1}^*, ..., f_T^*) - \mathfrak{L}_{\phi,t}(f_{t,\tau_t - \epsilon'_{n,t}}^*; f_{t+1}^*, ..., f_T^*)| \\
& + \mathfrak{L}_{\phi,t}(f_{t,\tau_t - \epsilon'_{n,t}}^*; f_{t+1}^*, ..., f_T^*) - \mathfrak{L}_{\phi,t}(f_t^*; f_{t+1}^*, ..., f_T^*) \\
\leq & c(\sigma_{n,t}^{-\alpha_t d_t} + \epsilon'_{n,t}) \leq O(\epsilon'_{n,t})
\end{aligned}$$

where in the second inequality from the bottom we used the Lipschitz continuity of the value function in Lemma 2.5 and by definition $f_t^* = f_{t,\tau_t}^*$.

For (II), we have

$$
\begin{aligned}
(II) \leq & \mathfrak{L}_{\phi,t}(\widehat{f}_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|\widehat{f}_t\|^2_{\mathcal{G}_t} - \inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) \\
= & \Bigg[\mathfrak{L}_{\phi,t}(\widehat{f}_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|\widehat{f}_t\|^2_{\mathcal{G}_t} \\
& - \inf_{f_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t})} \left(\mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) \Bigg] \\
& + \left[\inf_{f_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t})} \left(\mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) \right. \\
& \left. - \inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) \right],
\end{aligned} \tag{2.29}
$$

where

$$
\mathcal{A}_{t,n}(\tau, \mathcal{C}_{n,t}) = \left\{ f \in \mathcal{G}_t \middle| \|f\|_{\mathcal{G}_t} \leq \mathcal{C}_{n,t}, \frac{1}{n}\sum_{i=1}^{n} \frac{R_{it}\psi(A_{it}f(H_{it}), \eta_{n,t})}{p(A_{it}|H_{it})} \leq \tau_t \right\}.
$$

The first term on the right-hand side of the inequality (2.29) can be bounded by

$$
\begin{aligned}
& \mathfrak{L}_{\phi,t}(\widehat{f}_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|\widehat{f}_t\|^2_{\mathcal{G}_t} - \inf_{f_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t})} \left(\mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) \\
\leq & 2c_1^{-1} |\mathcal{V}_{t+1}(f^*_{t+1}, ..., f^*_T) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
& + \underbrace{\left[\mathfrak{L}_{\phi,t}(\widehat{f}_t; f^*_{t+1}, ..., f^*_T) + \lambda_{n,t}\|\widehat{f}_t\|^2_{\mathcal{G}_t} - \inf_{f_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t})} \left(\mathfrak{L}_{\phi,t}(f_t; f^*_{t+1}, ..., f^*_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) \right]}_{(IV)},
\end{aligned}
$$

and (IV) is equal to stochastic error term (I) in the proof of $T = 1$ with Y being replaced by Q_t. Note that $|Q_t| \leq (T - t + 1)M$ and consequently (IV) can be bounded using exactly the same argument for term (I), which turns out to have order $O(\xi^{(1)}_{n,t})$ with probability at least $1 - \exp(-x_t)$. For the second term of (2.29), we have

$$
\begin{aligned}
& \inf_{f_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t})} \left(\mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) - \inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; \widehat{f}_{t+1}, ..., \widehat{f}_T) \\
\leq & 2c_1^{-1} |\mathcal{V}_{t+1}(f^*_{t+1}, ..., f^*_T) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
& + \underbrace{\left[\inf_{f_t \in \mathcal{A}_{t,n}(\tau_t)} \left(\mathfrak{L}_{\phi,t}(f_t; f^*_{t+1}, ..., f^*_T) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t} \right) - \inf_{f_t \in \mathcal{A}_t(\tau_t, \mathcal{C}_{n,t})} \mathfrak{L}_{\phi,t}(f_t; f^*_{t+1,\tau_{t+1}}, ..., f^*_T) \right]}_{(V)}.
\end{aligned}
$$

Note that (V) is the approximation bias term in (2.29) also with Y being replaced by Q_t. Hence, following the same argument for $T = 1$, (V) can be decomposed to terms $(II) - (IV)$ in (2.29) and bounded separately, which turns out to have order $O(\xi_{n,t}^{(2)}) + O(\epsilon'_{n,t})$ in total with probability at least $1 - 2\exp\big(-\frac{n\delta_t^2 c_1^2}{2(T-t+1)^2 M^2}\big)$. Combing these results, we conclude that with probability at least $1 - h_n(t, x_t)$,

$$\begin{aligned}
\mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T) \leq & 5c_1^{-1}|\mathcal{V}_{t+1}(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
& + c\eta_{n,t}^{-1} + \underbrace{c(\epsilon'_{n,t})}_{(III)} + \underbrace{c(\xi_{n,t}^{(1)} + \xi_{n,t}^{(2)} + \epsilon'_{n,t})}_{(IV)+(V)} \\
\leq & 5c_1^{-1}|\mathcal{V}_{t+1}(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
& + c(\xi_{n,t} + \epsilon'_{n,t} + \eta_{n,t}^{-1})
\end{aligned} \tag{2.30}$$

for some constant c.

On the other hand, according to Lemma 2.8, similar to the prove when $T = 1$ we can show that $\widehat{f}_t \in \mathcal{A}_{t,n}(\tau_t, \mathcal{C}_{n,t}) \subseteq \mathcal{A}_t(\tau_t + \epsilon'_{n,t}, \mathcal{C}_{n,t})$ with probability at least $1 - 2\exp(\frac{n\delta_t^2 c_1^2}{2(T-t+1)^2 M^2})$. Therefore, we have

$$\begin{aligned}
& \mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T) \\
\geq & \mathcal{V}_t(f_t^*, ..., f_T^*) - \sup_{f_t \in \mathcal{A}_t(\tau_t + \epsilon'_{n,t}, \mathcal{C}_{n,t})} \mathcal{V}_t(f_t, \widehat{f}_{t+1}, ..., \widehat{f}_T) \\
\geq & \mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(f_{t,\tau_t+\epsilon'_{n,t}}^*, f_{t+1}^*, ..., f_T^*) \\
& - c_1^{-1}|\mathcal{V}_t(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
\geq & c\epsilon'_{n,t} - c_1^{-1}|\mathcal{V}_t(f_{t+1,\tau_{t+1}}^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_{t+1}, ..., \widehat{f}_T)|.
\end{aligned} \tag{2.31}$$

Finally, by combining (2.30) and (2.31), we obtain that with probability at least $1 - h_n(t, x_t)$,

$$\begin{aligned}
|\mathcal{V}_t(f_t^*, ..., f_T^*) - \mathcal{V}_t(\widehat{f}_t, ..., \widehat{f}_T)| \leq & 5c_1^{-1}|\mathcal{V}_{t+1}(f_{t+1}^*, ..., f_T^*) - \mathcal{V}_{t+1}(\widehat{f}_{t+1}, ..., \widehat{f}_T)| \\
& + c(\xi_{n,t} + \epsilon'_{n,t} + \eta_{n,t}^{-1}).
\end{aligned}$$

Hence, (2.13) in Theorem 2.4 follows by induction starting from $t = T$ to 1. The error bound of risk (2.14) can be established by repeating the same argument in Corollary 2.1 for each stage. This completes the proof of Theorem 2.4.

2.10 Additional Results for Simulation Setting I and II

This section reports the additional simulation results for setting I with $\tau = 1.5$ and setting II with $\tau = 1.3$. See Figures 2.3 and 2.4 and Table 2.4.

Linear 1.5

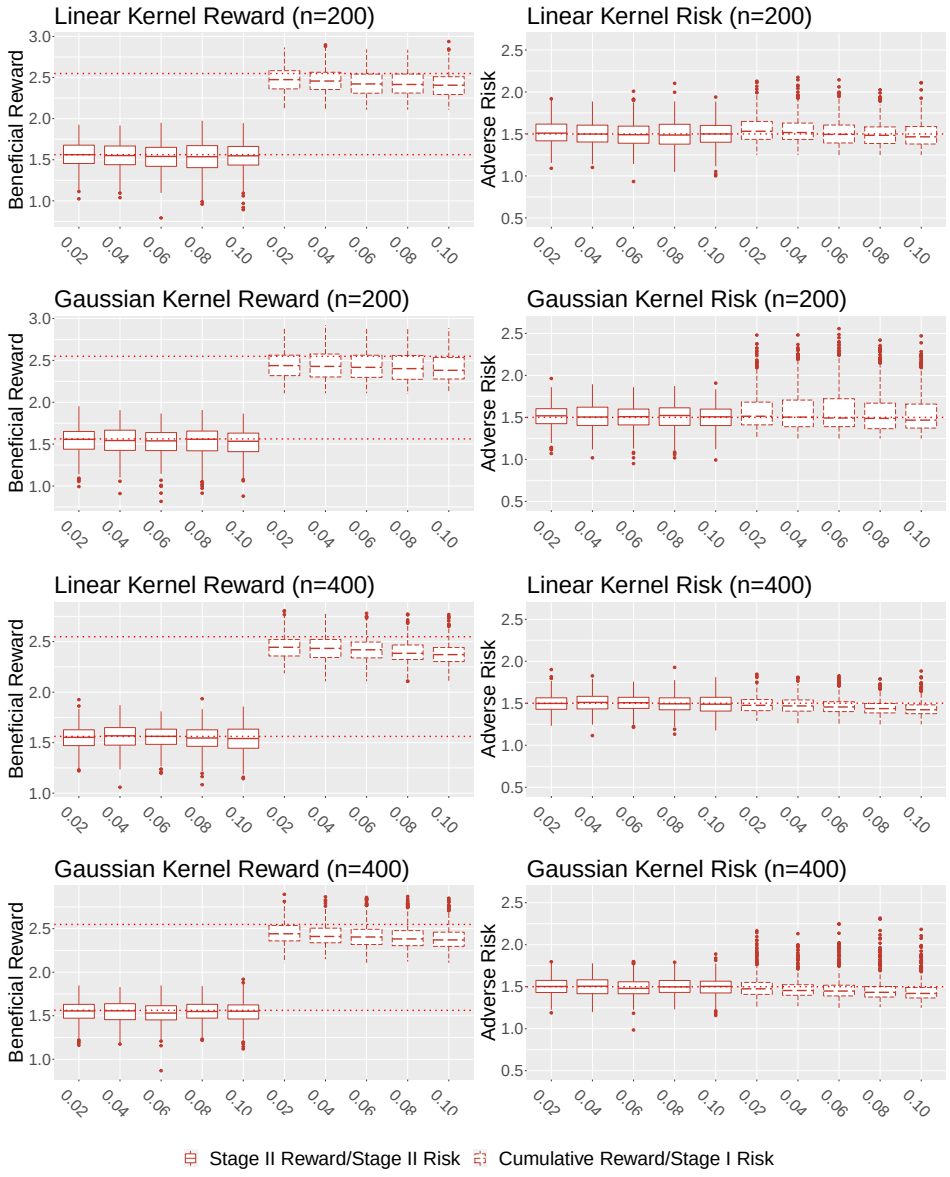

Figure 2.3: Estimated reward/risk on independent testing data set for simulation setting I, training sample size $n = \{200, 400\}$, $\eta = \{0.02, 0.04, ..., 0.1\}$ under linear kernel or Gaussian kernel. The dashed line in reward plots refers to the theoretical optimal reward under given constraints. The dashed line in risk plots represents the risk constraint $\tau = 1.5$.

Nonlinear 1.3

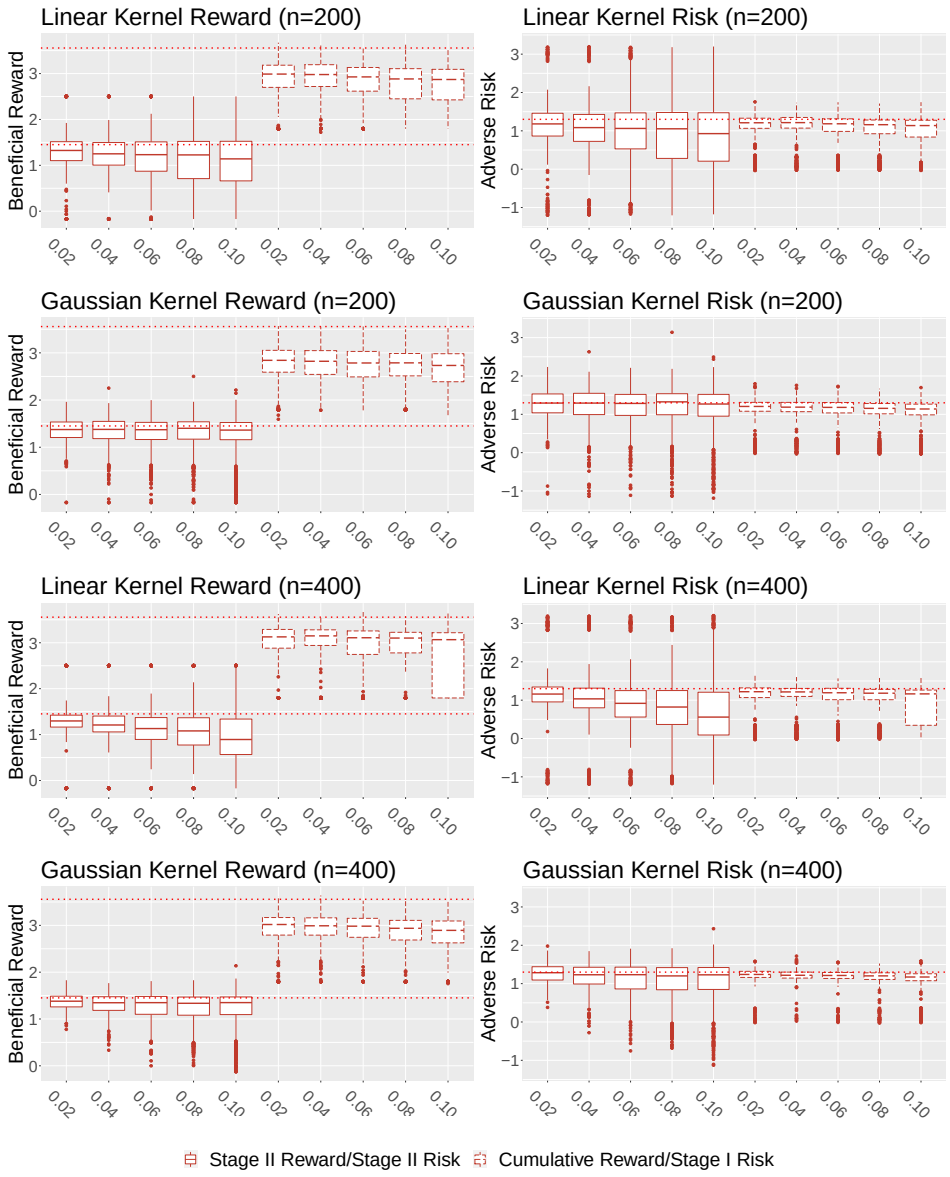

Figure 2.4: Estimated reward/risk on independent testing data set for simulation setting II, training sample size $n = \{200, 400\}$ and $\eta = \{0.02, 0.04, ..., 0.1\}$ (x-axis) under linear kernel or Gaussian kernel. The dashed line in reward plots refers to the theoretical optimal reward under given constraints. The dashed line in risk plots represents the risk constraint $\tau = 1.3$.

Table 2.4: Estimated reward/risk on independent testing data for setting I - $\tau_1 = \tau_2 = 1.5$, setting II - $\tau_1 = \tau_2 = 1.3$ and $n = 400$ under 3 different methods using linear/Gaussian kernel.

			Linear Kernel				Gaussian Kernel			
Setting	η	Method	Reward - II	Risk - II	Reward - I	Risk - I	Reward - II	Risk - II	Reward - I	Risk - I
Setting I	0.02	BR-DTRs	1.557(0.080)	1.502(0.070)	2.443(0.082)	1.475(0.066)	1.553(0.078)	1.502(0.069)	2.441(0.085)	1.474(0.069)
	0.02	Naive	—	—	2.339(0.096)	1.460(0.083)	—	—	2.339(0.099)	1.456(0.077)
	0.04	BR-DTRs	1.568(0.082)	1.510(0.072)	2.429(0.087)	1.467(0.066)	1.558(0.091)	1.506(0.081)	2.408(0.082)	1.453(0.062)
	0.04	Naive	—	—	2.319(0.098)	1.430(0.088)	—	—	2.297(0.107)	1.425(0.089)
	0.06	BR-DTRs	1.563(0.075)	1.507(0.066)	2.420(0.081)	1.455(0.059)	1.535(0.081)	1.484(0.070)	2.404(0.087)	1.449(0.065)
	0.06	Naive	—	—	2.306(0.100)	1.416(0.084)	—	—	2.297(0.110)	1.412(0.095)
	0.08	BR-DTRs	1.546(0.082)	1.493(0.071)	2.387(0.075)	1.439(0.059)	1.549(0.082)	1.498(0.074)	2.383(0.078)	1.432(0.060)
	0.08	Naive	—	—	2.274(0.112)	1.397(0.083)	—	—	2.273(0.097)	1.391(0.077)
	0.1	BR-DTRs	1.544(0.093)	1.489(0.082)	2.371(0.068)	1.421(0.054)	1.542(0.082)	1.493(0.072)	2.375(0.082)	1.422(0.060)
	0.1	Naive	—	—	2.273(0.093)	1.383(0.080)	—	—	2.272(0.096)	1.384(0.079)
		AOWL	1.983(0.010)	2.149(0.044)	3.257(0.018)	2.678(0.096)	1.914(0.030)	2.099(0.083)	3.212(0.036)	2.584(0.218)
Setting II	0.02	BR-DTRs	1.297(0.132)	1.159(0.196)	3.128(0.178)	1.221(0.116)	1.380(0.114)	1.285(0.169)	3.019(0.184)	1.240(0.079)
	0.02	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.04	BR-DTRs	1.209(0.168)	1.034(0.252)	3.150(0.157)	1.215(0.099)	1.347(0.139)	1.231(0.206)	2.991(0.192)	1.218(0.077)
	0.04	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.06	BR-DTRs	1.131(0.238)	0.917(0.351)	3.109(0.179)	1.193(0.129)	1.349(0.174)	1.233(0.253)	2.983(0.194)	1.213(0.078)
	0.06	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.08	BR-DTRs	1.080(0.299)	0.821(0.440)	3.102(0.172)	1.182(0.115)	1.335(0.168)	1.202(0.247)	2.937(0.199)	1.202(0.088)
	0.08	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
	0.1	Naive	—	—	1.797(0.000)	0.166(0.000)	—	—	1.797(0.000)	0.166(0.000)
		AOWL	2.440(0.064)	3.017(0.002)	5.188(0.000)	2.839(0.000)	2.424(0.080)	3.018(0.002)	5.188(0.000)	2.839(0.000)

CHAPTER 3: CONTROLLING CUMULATIVE ADVERSE RISK IN LEARNING OPTIMAL DYNAMIC TREATMENT REGIMENS

3.1 Introduction

In Chapter 2, we propose a new method to handle optimal DTRs problems for treating chronic diseases where aggressive treatment with a better reward is often accompanied by higher toxicity. The problem can be formulated as an optimal DTRs learning problem with stagewise risk restriction and a backward induction-based method, BR-DTRs, is proposed to learn optimal DTRs under the stagewise restriction. However, in practice, more aggressive treatments may lead to a higher efficacy but are also more likely to induce elevated risk in the long term. For example, when treating type 2 diabetes (T2D), American Diabetes Association (ADA) recommends intensified insulin therapy when patients fail to reach a safe hemoglobin A1c level after receiving first and second-line medications (American Diabetes Association, 2022*a*). However, several studies have shown that insulin therapies are commonly associated with long-term weight gain, which can potentially increase the risk of cardiovascular diseases (Wing et al., 2011; Apovian, Okemah and O'Neil, 2019). Thus, weight gain is recommended to be controlled under 5% for T2D patients (Park et al., 2022). Other examples include aggressive therapies for cancer (e.g., radiotherapy) or kidney failure (e.g., dialysis), which may treat diseases effectively but often lead to poor quality of life and a high economic burden for patients.

Most methods in personalized medicine literature that consider benefit-risk trade-offs are restricted to a single-stage decision problem. One class of methods (Lee et al., 2015; Butler et al., 2018) prespecify a utility function to combine benefit and risk outcomes into a single composite outcome, and the optimal decision is obtained by maximizing the utility function. A major limitation of these methods is that it is often difficult to reach a consensus on how to prespecify the composite outcome, especially when the benefit and risk outcomes are measured on very different scales. Recent work in reinforcement learning (e.g., Bhatnagar and Lakshmanan, 2012; Mahdavi, Jin and Yang, 2012; Chow et al., 2017; Yu et al., 2019; Cao, Zhang and Poor, 2021; Ding et al., 2021; Badanidiyuru, Kleinberg and Slivkins, 2018; Cayci, Eryilmaz and Srikant, 2020) have considered learning optimal policy under safety/budget constraints. However, these methods rely on

the Markovian decision process assumption (MDP) and require parametric models for the unknown policy, which do not hold for general DTRs problems.

When the cumulative risk needs to be considered in a DTR problem, the most important challenge is that due to delayed effects, treatments at one stage may affect both the benefit and risk outcomes in any of the future stages. Therefore, the estimation of the optimal treatment rule at any stage must take into account its cumulative impact on future stages. However, commonly used backward algorithms such as Q-learning or O-learning require the future stage rules to be already estimated optimally. These methods are no longer applicable because the cumulative risk control depends on not only the future stage rules but also the treatment decision, which is yet to be estimated at the current stage.

To respond to the real demand from clinical application, in this chapter, we propose a new statistical learning framework, namely, multistage cumulative benefit-risk (CBR) framework, to estimate the optimal DTRs that maximize the expected benefit (or reward) outcome but, at the same time, control the expected cumulative risk below a pre-specified threshold. We propose two methods to solve CBR. First, we introduce a Lagrange function and obtain its solution via solving an unconstrained DTR problem using a backward algorithm based on Q-learning or O-learning. Second and more interestingly, we propose a new procedure under multistage ramp loss (MRL) to estimate the DTRs simultaneously across all stages. The MRL can be viewed as an extension of the univariate ramp loss to a multivariate setting.

Our work presented in this chapter contains several novel contributions. First, converting the constrained estimation for DTRs to the unconstrained problem enables us to adopt the backward algorithm from the existing methods to estimate the optimal DTRs, and we prove that the latter leads to the optimal DTRs that satisfy the cumulative risk control. Second, in addition to the backward induction algorithm, we also propose a simultaneous learning method based on MRL, for which the estimation of one decision function is contingent on other decisions at later stages so that we can estimate the treatment rules using all data at the same time. Third, we show that the non-asymptotic convergence rates of the expected reward and risk under the estimated rules can be derived from the unconstrained DTRs associated with the Lagrange function, which provides the finite sample performance guarantee. We also show that using MRL is guaranteed to yield Fisher consistent rules for any unconstrained DTRs problem, and consequently, using the multistage ramp loss along with the proposed estimation procedure will yield the true optimal DTRs.

The remaining chapter is organized as follows. In Section 3.2, we formally introduce the CBR problem along with assumptions. We then describe a general framework to solve CBR after converting the problem

to an unconstrained one. In the same section, we present a backward algorithm based on Q-learning and O-learning, and the new MRL approach to obtain the solutions using empirical data. In Section 3.3, we obtain the non-asymptotic convergence rates for both the expected reward and risk under the estimated rules. In Section 3.4, we present results from simulation studies to examine the performance of the proposed approaches. In Section 3.5, we apply the proposed methods to estimate the optimal DTRs using a two-stage trial for treating T2D patients. In Section 3.6, we discuss possible future extensions based on our work. The DC algorithm of conducting MRL is presented in Section 3.7 and the proofs of the main results are given in Section 3.8 and Section 3.9.

3.2 Method

3.2.1 Problem setup and assumptions

Consider a T-stage decision problem, where T is finite and often small in clinical settings. We use Y to denote the total reward at the end of stage T and R to denote the cumulative risk at stage T, both assumed to be bounded by a constant M. We consider a sequence of dichotomous treatments over T stages and let $A_t \in \{-1, +1\}$ denote the observed treatment at stage t. Additionally, we let H_t denote all observed feature variables prior to stage t, including the treatments or any immediate outcomes in previous stages. Thus, $H_1 \subset H_2 \subset \cdots \subset H_T$. We assume that data are from a sequential multiple assignment randomized trial (SMART) (Murphy, 2005a), so the observed data for n independent subjects consist of $H_{i1}, A_{i1}, H_{i2}, \cdots, A_{iT}, H_{iT}, Y_i$ and R_i for $i = 1, ..., n$. Like previous chapters, we define a DTR to be any function from the space:

$$\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_T \to \{-1, +1\}^T, \quad \text{where } \mathcal{D}_t : \mathcal{H}_t \to \{-1, +1\}.$$

To control the cumulative risk, we formulate the CBR problem as seeking the optimal rule $\mathcal{D}^* = (\mathcal{D}_1^*, ..., \mathcal{D}_T^*)$ that solves the optimization problem

$$\max_{\mathcal{D}} \ E^{\mathcal{D}}[Y],$$
$$\text{subject to } \ E^{\mathcal{D}}[R] \leq \tau$$

for a prespecified risk constraint τ. Here, $E^{\mathcal{D}}[\cdot]$ denotes the expectation when A_t are forced to be $\mathcal{D}_t(H_t)$ for $t = 1, ..., T$. In other words, the optimal treatment rule yields the maximal reward at stage T among all feasible rules whose cumulative risk is no greater than the risk threshold τ.

Like the standard DTRs and BR-DTRs, to ensure that $E^{\mathcal{D}}[\cdot]$ is estimable given the observed data, we also require several assumptions.

Assumption 3.1 *Stable Unit Treatment Value (SUTV): A subject's cumulative potential outcome is not influenced by other subjects' treatments allocation, i.e.,*

$$(Y, R) = (Y(\bar{a}_T), R(\bar{a}_T)), \text{ if } \bar{A}_T = \bar{a}_T.$$

Assumption 3.2 *No Unmeasured Confounders (NUC): For any $t \in \{1, ..., T\}$ and $\bar{a}_T \in \{-1, +1\}^T$,*

$$A_t \perp\!\!\!\perp (H_{t+1}(\bar{a}_t), ..., H_T(\bar{a}_{T-1}), Y(\bar{a}_T), R(\bar{a}_T)) \big| H_t.$$

Assumption 3.3 *(Positivity) For any $t = 1, ..., T$, there exists universal constants $0 < c_1 \leq c_2 < 1$ such that*

$$c_1 \leq p(A_t = 1|H_t) \leq c_2 \quad \textit{for } H_t \textit{ a.s.}$$

Assumption 3.1 and 3.2 are SUTV and NUC assumptions under the cumulative risk control framework and Assumption 3.3 is a restatement of the Positivity assumption. In particular, again under Assumption 3.1 to 3.3, Qian and Murphy (2011) showed that the original problem can be reformulated as

$$\max_{\mathcal{D}} \; E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right], \quad \text{subject to } E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right] \leq \tau. \tag{3.1}$$

Finally, assuming that the decision rules are determined as the signs of some decision functions $(f_1, ..., f_T)$, i.e., $\mathcal{D}_t(H_t) = \text{sign}(f_t(H_t))$, then (3.1) becomes

$$\max_{(f_1,...,f_T) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_T} \; E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right], \quad \text{subject to } E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] \leq \tau, \tag{3.2}$$

where $\mathcal{F}_t$ denotes the set of all measurable functions from $\mathcal{H}_t$ to $\mathbb{R}$

3.2.2 A general procedure for solving CBR problem

To solve CBR problems, we consider the Lagrange function of (3.1), or equivalently, (3.2). For any $\kappa \in [0, \infty]$, the Lagrange function of (3.1) with multiplier κ is given by

$$E\left[\{Y - \kappa(R - \tau)\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right].$$

Letting $\gamma = \kappa/(1+\kappa) \in [0, 1]$, we aim to solve the following problem for each γ:

$$\mathcal{D}^*_\gamma = (\mathcal{D}^*_{1,\gamma}, ..., \mathcal{D}^*_{T,\gamma}) = \arg\max_{\mathcal{D}} E\left[\{(1-\gamma)Y - \gamma R\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \tag{3.3}$$

where we omit the constant τ which will not affect the solution when γ is fixed.

Let $\mathfrak{Y}(\gamma)$ and $\mathfrak{R}(\gamma)$ denote the expected reward and risk associated with the optimal decision rules of (3.3), i.e.,

$$\mathfrak{Y}(\gamma) = E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma}(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right],\ \ \mathfrak{R}(\gamma) = E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma}(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right].$$

To ensure that there exists a non-trivial solution to the above problem, we also require the following regular assumption:

Assumption 3.4 *$\mathfrak{R}(\gamma)$ is a continuous function for $\gamma \in [0, 1]$ and $\mathfrak{R}(1) < \tau \leq \mathfrak{R}(0)$.*

As a note, the restriction $\mathfrak{R}(1) < \tau$ in Assumption 3.4 ensures that there exists at least one feasible DTRs that satisfy the risk constraint, and $\tau \leq \mathfrak{R}(0)$ is to exclude the trivial case when the cumulative risk for the optimal DTR without the constraint is not larger than τ. The continuity assumption in Assumption 3.4 implies that there exists some γ^*, which may not be unique, satisfying $\mathfrak{R}(\gamma^*) = \tau$. For any such γ^*, our following lemma shows that $\mathcal{D}_{\gamma^*}$ is indeed the optimal DTRs.

Lemma 3.1 *Under Assumption 3.1 to 3.4, both $\mathfrak{Y}(\gamma)$ and $\mathfrak{R}(\gamma)$ are non-increasing function of γ. Furthermore, $E^{\mathcal{D}_{\gamma^*}}[Y] \geq E^{\mathcal{D}}[Y]$ for any DTRs, $\mathcal{D}$, satisfying $E^{\mathcal{D}}[R] \leq \tau$.*

Lemma 3.1 indicates that solving unconstrained problem (3.3) associated with γ^* produces a solution of the CBR problem. The proof of Lemma 3.1 is given in the Section 3.8. In addition, the continuity of $\mathfrak{R}(\gamma)$ implies that searching for γ^* can be carried out using the bisection procedure starting from $\gamma_{\min} = 0$ and

$\gamma_{\max} = 1$ until reaching the termination condition $|\gamma_{\min} - \gamma_{\max}| \leq \epsilon$ for some convergence threshold ϵ. The complete numerical algorithm based on bisection search is provided in the Section 3.7.

As an important remark, although the lemma implies that the optimal DTRs are associated with a linear combination of Y and R, it should be noted that the coefficient in this linear combination, i.e, γ^*, is data-driven and depends on the DTRs. Therefore, this problem is fundamentally different from learning the optimal DTRs based on a utility function where the linear combination needs to be pre-specified.

3.2.3 Backward algorithm for maximizing the Lagrange function

Since (3.3) is an unconstrained problem for estimating DTRs for fixed γ, many existing methods such as Q-learning and OWL can be used to learn the optimal DTRs using a backward procedure, after treating $(1-\gamma)Y - \gamma R$ as the reward outcome. Specifically, we define Q-function in turn for $t = T, T-1, ..., 1$ as

$$Q_{t,\gamma}(h_t, a_t) = E[\max_{a_{t+1} \in \{-1,+1\}} Q_{t+1,\gamma}(H_{t+1}, a_{t+1}) | H_t = h_t, A_t = a_t]$$

with $Q_{T+1,\gamma} = (1-\gamma)Y - \gamma R$. Then the optimal solution for $\mathcal{D}^*_\gamma$ is

$$\mathcal{D}^*_{t,\gamma}(h_t) = \text{sign}(Q_{t,\gamma}(h_t, 1) - Q_{t,\gamma}(h_t, -1)), \quad t = 1, ..., T.$$

A backward Q-learning estimates the conditional expectation in the definitions of $Q_{t,\gamma}$ using regression models, in turn from $t = T$ to $t = 1$, then the estimated DTRs are obtained by plugging the estimated Q-functions into the above expression (Qian and Murphy, 2011).

A more robust procedure without fitting regression models, namely backward OWL, uses weighted support vector machines to directly optimize the objective function at each stage (Zhao et al., 2015). Specifically, let $(g^*_{1,\gamma}, ..., g^*_{T,\gamma})$ denotes the optimal decision functions corresponding to the outcome $O_\gamma = (1-\gamma)Y - \gamma R$, then Zhao et al. (2015) indicates that $\{g^*_{t,\gamma}\}_{t=1}^T$ can be sequentially estimated via

$$g^*_{t,\gamma} = \arg\max_{f \in \mathcal{F}_t} E\left[O_\gamma \frac{\mathbb{I}(A_t f_t(H_t) > 0) \prod_{s=t+1}^T \mathbb{I}(A_s g^*_{s,\gamma}(H_s) > 0)}{\prod_{s=t}^T p(A_s | H_s)} \right] \tag{3.4}$$

for $t = T, \cdots, 1$ in a backward order. In other words, the optimal decision function at stage t can be obtained by maximizing the expected cumulative reward up to stage t among patients whose future observed treatments follow the optimal treatments. Using empirical data, an estimator of $g^*_{t,\gamma}$ can be obtained via solving the

empirical version of (3.4) in a backward order, and by replacing the zero-one function, $\mathbb{I}(A_t f_t(H_t) > 0)$, with some surrogate function. In particular, Zhao et al. (2012) adopted the hinge loss defined as $\phi(x) = (1-x)^+$ and sequentially solved the following problem

$$\widehat{f}_{t,\gamma} = \arg\min_{f_t \in \mathcal{G}_t} \frac{1}{n} \sum_{i=1}^{n} \{(1-\gamma)Y_i - \gamma R_i\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_{is} \widehat{f}_{s,\gamma}(H_{is}) > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} \phi(A_{it} f_t(H_{it})) + \lambda_{n,t} \|f_t\|_{\mathcal{G}_t}^2, \quad (3.5)$$

where $\mathcal{G}_t$ is a subspace of $\mathcal{F}_t$. The last term, $\lambda_{n,t}\|f_t\|_{\mathcal{G}_t}^2$, is a regularization term to mitigate overfitting. When $\{\mathcal{G}_t\}_{t=1}^T$ are reproducing kernel Hilbert space (RKHS), the optimization problem (3.5) can be reformulated as a weighted support vector machine problem (SVM) (Cortes and Vapnik, 1995), which can be efficiently solved using standard optimization software. Typical choices of RKHS include the space generated by a linear kernel or a Gaussian kernel with inner product $\langle H_{it}, H_{jt} \rangle = e^{-\sigma^2 \|H_{it} - H_{jt}\|_2^2}$ for bandwidth σ^{-1}. Given observed data, the tuning parameter $\{\lambda_{n,t}\}_{t=1}^T$ and $\{\sigma_{n,t}\}_{t=1}^T$ can be selected via cross-validation.

As introduced in Section 1.1, to further improve the performance of OWL, Liu et al. (2018) proposed the augmented OWL (AOWL) by predicting the expected Q-function of subjects whose observed treatment assignments do not follow the optimal estimated rules and incorporating such predictions to calculate pseudo-outcomes through a doubly robust construction. In our subsequent numerical studies, we use both OWL and AOWL for this backward algorithm to solve the Lagrange function in (3.3) and use O-learning to refer to either OWL or AOWL when the context is clear.

3.2.4 Simultaneous algorithm for maximizing the Lagrange function

One disadvantage of O-learning is that the estimation of the early stage can only utilize the information from patients whose observed treatment assignments follow the optimal rules as shown in (3.4). Moreover, for backward induction methods such as Q-learning and O-learning, the estimation error from later stages due to either model misspecification or overfitting will be accumulated and always present in early-stage estimation. To overcome these disadvantages, in this section, we propose a simultaneous algorithm based on multi-stage ramp loss (MRL) described as follows.

Our key idea is to replace the multivariate zero-one indicator function in (3.3) with a continuous surrogate function to be directly optimized without any backward algorithm. Specifically, define $\psi(\cdot)$ as a piecewise linear function given by $\psi(x) = \max(\min(x, 1), 0)$, then we consider solving the following surrogate

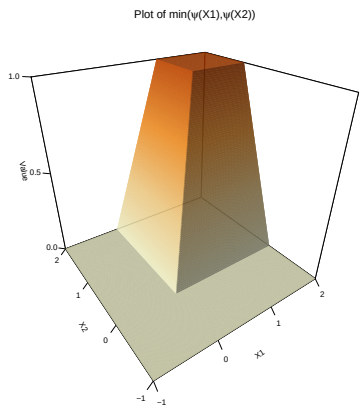

Figure 3.1: 3D plot of multivariate ramp loss, $\min(\psi(x_1), \psi(x_2))$.

problem to substitute (3.3):

$$\max_{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T} E\left[O_\gamma^+ \frac{\min(\psi(A_1 f_1(H_1)),...,\psi(A_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] + E\left[\sum_{a_t\in\{-1,+1\},\ a_t\neq A_t} O_\gamma^- \frac{\min(\psi(a_1 f_1(H_1)),...,\psi(a_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right], \tag{3.6}$$

Here, O_γ^+ and O_γ^- denote the positive and negative part of O_γ respectively, i.e., $O_\gamma^+ = \max(O_\gamma, 0)$ and $O_\gamma^- = \max(-O_\gamma, 0)$. When O_γ is non-negative, the optimization problem (3.6) can be viewed as a minimization problem with loss function

$$L(f) = -E[\min(\psi(A_1 f_1(H_1)),...,\psi(A_T f_T(H_T)))] + 1.$$

Figure 3.1 presents a three-dimensional visualization of this loss function (i.e., $T = 2$). In other words, L can be considered as a multivariate extension of the univariate ramp loss function proposed by Huang, Shi and Suykens (2014). Numerically, MRL can be more robust against extreme errors in f_t's than O-learning because MRL is bounded between 0 and 1 and is closer to the 0-1 loss compared with the hinge loss used in O-learning. Note that the expression (3.6) does not require the decision function f_{t_1} to be estimated before or after another decision function f_{t_2}. This implies that MRL indeed solves the optimal decision rules simultaneously so that all patients' information will be used during the estimation, and updating the decision functions in early stages will also update the decision functions of later stages. The second augmentation term of (3.6) changes the negative response variable to a positive value by reverting the observed treatments to any

other treatment sequences. This expression ensures that the weights in each term are always non-negative even if O_γ is negative. The following lemma ensures that MRL is a valid surrogate problem for (3.3).

Lemma 3.2 *If* $(f_1^*, ..., f_T^*)$ *is a solution to (3.6), then* $(sign(f_1^*), ..., sign(f_T^*))$ *maximizes (3.3).*

The proof of Lemma 3.2 is provided in the Section 3.9. Using the empirical data, we propose to solve

$$\max_{(f_1,...,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \frac{1}{n}\sum_{i=1}^{n} O_{i,\gamma}^{+}\frac{\min(\psi(A_{i1}f_1(H_{i1})), ..., \psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} + \frac{1}{n}\sum_{i=1}^{n}\sum_{a_t\in\{-1,1\}, a_t\neq A_{it}} O_{i,\gamma}^{-}\frac{\min(\psi(a_1f_1(H_{i1})), ..., \psi(a_Tf_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} - \sum_{t=1}^{T}\lambda_{n,t}\|f_t\|_{\mathcal{G}_t}^2, \tag{3.7}$$

where $O_{i,\gamma} = (1-\gamma)Y_i - \gamma R_i$. Again, we introduce a regularization term $\sum_{t=1}^{T}\lambda_{n,t}\|f_t\|_{\mathcal{G}_t}^2$ to prevent overfitting.

As a remark, note that in (3.3), the optimal solution is not affected after we subtract any function of H_1 from response variable O_γ. Similar to the augmentation technique used in AOWL, we can replace $O_{i,\gamma}$ by $\widehat{O}_{i,\gamma} = O_{i,\gamma} - \widehat{m}(H_{i1})$, where $\widehat{m}(H_1)$ is an estimator of the conditional expectation of O_γ given baseline feature variables H_1. The refined empirical problem then becomes

$$\max_{(f_1,...,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \frac{1}{n}\sum_{i=1}^{n} \widehat{O}_{i,\gamma}^{+}\frac{\min(\psi(A_{i1}f_1(H_{i1})), ..., \psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} + \frac{1}{n}\sum_{i=1}^{n}\sum_{a_t\in\{-1,1\}, a_t\neq A_{it}} \widehat{O}_{i,\gamma}^{-}\frac{\min(\psi(a_1f_1(H_{i1})), ..., \psi(a_Tf_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} - \sum_{t=1}^{T}\lambda_{n,t}\|f_t\|_{\mathcal{G}_t}^2. \tag{3.8}$$

When context is clear, we will use $(\widehat{f}_{1,\gamma}, ..., \widehat{f}_{T,\gamma})$ to denote the solution of (3.7) and (3.8).

Computationally, the objective function of (3.8) can be further decomposed as the difference between two convex functions. Therefore, one can adopt the difference of convex (DC) algorithm (Tao and An, 1997) to solve (3.8) iteratively. When $\{\mathcal{G}_t\}_{t=1}^{T}$ are RKHS, in each iteration of the DC algorithm, the optimization problem can be further reduced to a quadratic programming problem so can be efficiently solved using existing software. The details are given in Section 3.7.

3.2.5 Estimating γ^* using the risk control

Finally, to determine the estimate for γ^*, since the empirical estimator of the risk, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} R_i \frac{\prod_{t=1}^{T} \mathbb{I}(A_{it}\widehat{f}_{t,\gamma}(H_{it}) > 0)}{\prod_{t=1}^{T} p(A_{it}|H_{it})},$$

is not continuous in γ, a small change of γ may lead to a significant risk control violation. Thus, we propose to estimate γ^* based on a smooth approximation to the above function. Specifically, we obtain γ^*'s estimator, denoted by $\widehat{\gamma}$, via bisection method to solve equation

$$\frac{1}{n}\sum_{i=1}^{n} R_i \frac{\min(\psi(A_{i1}\widehat{f}_{1,\widehat{\gamma}}(H_{i1})/\eta), ..., \psi(A_{iT}\widehat{f}_{T,\widehat{\gamma}}(H_{iT})/\eta))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} = \tau. \tag{3.9}$$

Here, $\eta \in (0, 1]$ is a small shifting parameter to be chosen data dependently.

3.3 Theoretical Results

In this section, we present the theoretical results for the expected reward and risk under the estimated DTRs. Recall that $(g^*_{1,\gamma}, ..., g^*_{T,\gamma})$ are the optimal decision functions of unconstrained problem (3.3) and let $(g^*_1, ..., g^*_T)$ denote the optimal decision function of original CBR problem (3.1), then Lemma 3.1 indicates $(g^*_1, ..., g^*_T)$ can be selected as $(g^*_{1,\gamma^*}, ..., g^*_{T,\gamma^*})$. We wish to obtain a non-asymptotic lower bound for

$$\mathcal{V}(\widehat{f}_{1,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}}) - \mathcal{V}(g^*_1, ..., g^*_T) = E\left[Y\frac{\prod_{t=1}^{T} \mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] - E\left[Y\frac{\prod_{t=1}^{T} \mathbb{I}(A_t g^*_t(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \tag{3.10}$$

and an upper bound for

$$E\left[R\frac{\prod_{t=1}^{T} \mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] - \tau, \tag{3.11}$$

where $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^{T}$ are either from the O-learning algorithm or the simultaneous learning algorithm. We assume $\{\mathcal{G}_t\}_{t=1}^{T}$ to be Gaussian RKHS with bandwidth $\sigma_{n,t}^{-1}$.

We need additional assumptions to characterize the complexity of true optimal decision functions of each unconstrained DTRs under different multipliers γ. Similar to Chapter 2, for any given t and γ, we define

$$\mathcal{H}_{t,\gamma,1} = \{h_t \in \mathcal{H}_t | g^*_{t,\gamma}(h_t) > 0\}, \;\; \mathcal{H}_{t,\gamma,-1} = \{h_t \in \mathcal{H}_t | g^*_{t,\gamma}(h_t) < 0\},$$

and the Δ-function to be

$$\Delta_{t,\gamma}(h_t) = d(h_t, \mathcal{H}_{t,\gamma,1})I(h_t \in \mathcal{H}_{t,\gamma,-1}) + d(h_t, \mathcal{H}_{t,\gamma,-1})I(h_t \in \mathcal{H}_{t,\gamma,1}),$$

where $d(x, \mathcal{S})$ denote the Euclidean distant from point x to set $\mathcal{S}$. We assume the following conditions for $t = 1, ..., T$.

Assumption 3.5 *For any $\gamma \in [0, 1]$, there exist universal positive constants $\{\alpha_t\}_{t=1}^T$ and $K > 0$ such that*

$$\int_{\mathcal{H}_t} e^{-\frac{\Delta_{t,\gamma}^2(h)}{s}} P_t(dh) \leq K s^{\alpha_t d_t/2}$$

holds for $t = 1, ..., T$. Here, d_t denotes the dimension of $\mathcal{H}_t$ and P_t denotes the density function of the random variable H_t.

As a note, Assumption 3.5 is the general version of the geometric noise exponent assumption under the framework of CBR and will hold for arbitrary α_t like Assumption 2.5 when data is sparse near the decision rule. Our next assumption concerns the discrimination property of Q-function between the two treatments, which is sufficient to establish the convergence rate for the risk control.

Assumption 3.6 *For any $\gamma \in [0, 1]$, $t = 1, ..., T$, set $D_t \subseteq \mathcal{H}_t$ and $\eta_1 > 0$,*

$$E[|Q_{t,\gamma}(H_t, A_t = 1) - Q_{t,\gamma}(H_t, A_t = -1)|\mathbb{I}(H_t \in D_t)] \leq \eta_1$$

implies that $P(H_t \in D_t) \leq K_1\eta_1$ for some fixed positive constant K_1.

The following theorem gives the non-asymptotic convergence rates for the estimated DTRs using the O-learning algorithm.

Theorem 3.1 *Under Assumption 3.1 to 3.6, for any $0 < \theta_t$, $0 < \theta_t'$, $0 < \nu_t < 2$, $0 < \nu_t' < 2$ for $t = 1, ..., T$, assume that $\lambda_{n,t} \to 0$, $\sigma_{n,t} \to \infty$ and $\lambda_{n,t}\sigma_{n,t}^{d_t} \to 0$ where d_t denotes the dimension of $\mathcal{H}_t$. For sufficient small $\delta \geq 0$, let*

$$\epsilon_n = \sum_{t=1}^{T} c_1^{-(1-t)} C_{1,t} \left(\frac{1}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} + \lambda_{n,t}\sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \right)$$

and

$$\xi_n = Tc_1^{-2T}\sum_{t=1}^{T} C_{2,t}\frac{1}{\sqrt{n}}\sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2}\lambda_{n,t}^{-\nu_t'/4}\epsilon_n^{-\nu_t'/2}.$$

Then for $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^{T}$ *estimated from the O-learning approaches, we have*

$$\mathcal{V}(\widehat{f}_{1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}}) - \mathcal{V}(g_1^*,...,g_T^*) \geq -C_4\{C_3Tc_1^{-T}(\delta+\epsilon_n)+\xi_n\}$$

and

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \tau \leq C_4\{C_3Tc_1^{-T}(\delta+\epsilon_n)+\xi_n\}$$

holds with probability at least $1-\sum_{t=1}^{T}e^{-c_t'\delta^2 n} - e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$*, where* $\widehat{\gamma}$ *is determined via (3.9) with* $\eta=\epsilon_n/M$*. Here,* $C_{1,t}$ *denotes a positive constant depending on parameters* $(\nu_t,\theta_t,d_t,c_1,M,K)$*,* $C_{2,t}$ *denotes a positive constant depending on parameters* $(\nu_t',\theta_t',d_t,c_1,M)$*,* c_t' *denotes a positive constant depending on parameters* $(\nu_t,\theta_t,d_t,c_1,M)$ *for* $t=1,...,T$*,* C_3 *is a positive constant depending on* (M,K_1) *and* C_4 *is a positive constant depending on risk constraint* τ*.*

Let $\nu_t \to 0$, $\nu_t' \to 0$, $\theta_t \to 0$, $\theta_t' \to 0$ and assume that parameter α_t in Assumption 3.5 can be arbitrarily large for any $t=1,...,T$, then Theorem 3.1 indicates that the left-hand side of (3.10) and (3.11) can be both lower and upper bounded by a term of order as close as $O(n^{-1/2})$. Hence, Theorem 3.1 shows that under the ideal case, the beneficial reward under the estimated rules will be expected as high as the reward under optimal decision rules up to a small loss of order $O(n^{-1/2})$, with an induced adverse risk no exceeding than τ plus an error term also up to order $O(n^{-1/2})$.

Similar to O-learning, we can obtain the non-asymptotic convergence rate for the DTRs in the MRL approach using the Gaussian kernel. When $\widehat{\gamma}$ is determined via (3.9), a slightly different discrimination assumption is needed to quantify the impact of using $\widehat{\gamma}$ as an approximation or multiplier γ^* associated with τ. To this end, we assume

Assumption 3.7 *For any* $\gamma\in[0,1]$*,* $D_t \subseteq \mathcal{H}_t\times\{-1,+1\}$*,* $t\in\{1,...,T\}$ *and* $\eta_2>0$*, we assume that*

$$E\left[\frac{\mathbb{I}((H_t,A_t)\in D_t)}{\prod_{s=1}^{t}p(A_s|H_s)}U_{t+1}(H_{t+1};g_{t+1,\gamma}^*,...,g_{T,\gamma}^*;\gamma)\right] \leq \eta_2$$

implies $P((H_t, A_t) \in D_t) \leq K_2\eta_2$ *for some fixed positive constant* K_2*. Here,*

$$U_t(H_t; f_t, ..., f_T; \gamma) = E\Big[O_\gamma^+ \frac{\prod_{s=t}^T \mathbb{I}(A_s f_s(H_s) > 0)}{\prod_{s=t}^T p(A_s|H_s)} + \sum_{a_s \in \{-1,+1\},\ a_s \neq A_s} O_\gamma^- \frac{\prod_{s=t}^T \mathbb{I}(a_s f_s(H_s) > 0)}{\prod_{s=t}^T p(A_s|H_s)} \Big| H_t\Big].$$

The following theorem gives the theoretical results for the MRL approach.

Theorem 3.2 *Under Assumption 3.1 to 3.5 and 3.7, for any* $0 < \theta_t$, $0 < \theta_t'$, $0 < \nu_t \leq 2$, $0 < \nu_t' \leq 2$ *for* $t = 1, ..., T$*, assume that* $\lambda_{n,t} \to 0$, $\sigma_{n,t} \to \infty$ *and* $\lambda_{n,t}\sigma_{n,t}^{d_t} \to 0$ *where* d_t *denotes the dimension of* $\mathcal{H}_t$*. For sufficient small* $\delta \geq 0$*, let*

$$\epsilon_n = c_1^{-T} \sum_{t=1}^T C_{1,t} \bigg(\frac{1}{\sqrt{n}} \Big(\sqrt{T} + (T^2 c_1^{-3T})^{\nu_t/4} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} \Big) + \lambda_{n,t}\sigma_{n,t}^{d_t} + c_1^{-T}\sigma_{n,t}^{-\alpha_t d_t} \bigg)$$

and

$$\xi_n = T c_1^{-2T} \sum_{t=1}^T C_{2,t} \bigg(\frac{1}{\sqrt{n}} c_1^{-T\nu_t'/4} \sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2} \lambda_{n,t}^{-\nu_t'/4} \bigg).$$

Then for $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^T$ *estimated from the MRL approach, we have*

$$\mathcal{V}(\widehat{f}_{1,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}}) - \mathcal{V}(g_1^*, ..., g_T^*) \geq -C_4\{C_3 T c_1^{-T}(\delta + \epsilon_n) + \xi_n\}$$

and

$$E\bigg[R \frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg] - \tau \leq C_4\{C_3 T c_1^{-T}(\delta + \epsilon_n) + \xi_n\}$$

holds with probability at least $1 - 3e^{-\frac{1}{2}c_1^{2T} M^{-2}\delta^2 n}$ *where* $\widehat{\gamma}$ *is determined via (3.9) with* $\eta = \epsilon_n$*. Here,* $C_{1,t}$ *denotes a positive constant depending on parameters* $(\nu_t, \theta_t, d_t, c_1, M, K)$*,* $C_{2,t}$ *denotes a positive constant depending on parameters* $(\nu_t', \theta_t', d_t, c_1, M)$*,* C_3 *is a positive constant depending on* (M, K_2) *and* C_4 *is a positive constant depending on* τ*.*

Similar to before, let $\nu_t \to 0$, $\nu_t' \to 0$, $\theta_t \to 0$, $\theta_t' \to 0$ and assume that parameter α_t in Assumption 3.5 can be arbitrarily large for any $t = 1, ..., T$, then Theorem 3.2 implies that the right-hand side of (3.10) and (3.11) can also be lower and upper bounded by a term of order as close as $O(n^{-1/2})$, the same order as for the estimated DTRs using O-learning approach.

The main challenge to establish Theorem 3.1 and Theorem 3.2 is to show that the estimated multiplier $\widehat{\gamma}$ determined via (3.9) satisfies with a high probability that the expected risk under $(\widehat{f}_{1,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}})$ is close to

τ. This can be guaranteed by showing that each unconstrained DTRs problem (3.3) can be uniformly well estimated for fixed $\gamma \in [0, 1]$ under Assumption 3.5. The proof of both theorems is given in Section 3.9.

3.4 Simulation Studies

In the first simulation setting, we consider a 2-stage SMART. We first generate 7 baseline feature variables independently from Unif[-1,1], denoted as $(X_1, ..., X_7)$. To mimic the patient's evolving health status, we also generate a time-dependent covariate at the two stages, denoted as $(X_{8,1}, X_{8,2})$, using $X_{8,1} = \omega_0 + \omega_1, X_{8,2} = \omega_0 + \omega_2$, where ω_0, ω_1 and ω_2 are independently from Unif$[-0.5, 0.5]$. Treatments at the two stages, A_1 and A_2, take values 1 or -1 with equal probability. Finally, the cumulative reward variable Y and risk variable R are obtained using the following models:

$$Y = 1 - X_1 + X_2 + A_1(X_1 + 0.25X_{8,1} + 0.5) + A_2(X_{8,2} + A_1 + 0.25) + \epsilon_Y,$$

$$R = 2 + X_1 + X_2 + A_1(X_1 - X_2 + 0.5) + A_2(0.5X_1 + 0.5X_3 - X_{8,2} + 1) + \epsilon_R,$$

where $\epsilon_Y = \epsilon_0 + \epsilon_1, \epsilon_R = \epsilon_0 + \epsilon_2$ with ϵ_0 from $N(0, 1)$ truncated at ± 0.25 and ϵ_1 and ϵ_2 both from $N(0, 1)$ truncated at ± 0.5. In the second simulation setting, the feature variables are generated the same as before except that 7 baseline feature variables are from independent Unif[0,1] and ω_0 is from unif$[0.5, 1]$. The cumulative reward Y and risk R are generated using the following nonlinear models:

$$Y = 1 + 2X_2 + A_1(X_{8,1}^2 + 1) + A_2(X_{8,2}^2 + X_1^2) + \epsilon_Y,$$

$$R = 2 - X_2 + A_1(X_1 + 1) + A_2(A_1X_{8,2} + 1) + \epsilon_R,$$

where ϵ_Y and ϵ_R are generated the same way as in the first simulation setting. Note that for both simulation settings, the feature variables at each stage are $H_1 = (X_1, ..., X_7, X_{8,1})$ and $H_2 = (H_1, A_1, X_{8,2})$, respectively. We choose the risk constraint $\tau = 1$ for the first simulation setting and $\tau = 1.5$ for the second simulation setting.

For each simulation setting, we randomly generate the training data with sample sizes $n = 200$ and $n = 400$. For O-learning and MRL, both linear kernel and Gaussian radial basis kernel are implemented. When the Gaussian kernel is used, we follow Wu, Zhang and Liu (2010) to choose

$\sigma_{n,t}^{-1} = 1.25 * \text{median}_{A_{it} \neq A_{jt}} \|H_{it} - H_{jt}\|$. To choose the tuning parameters (λ_1, λ_2), we fix the tuning grid of $n\lambda_1$ and $n\lambda_2$ to be from $(2^{-8}, 2^{-6}, \cdots, 2^6, 2^8)$. The optimal tuning parameters are then determined via two-fold cross-validation, which yields the highest reward on the testing data.

Specifically to each algorithm in our proposed method, O-learning is conducted following both original OWL from Zhao et al. (2015) and AOWL from Liu et al. (2018). For MRL, we replace the original response variable with its residual as described in Section 3.2, where we estimate the conditional mean via Lasso regression. For each stage, the initial values of MRL are set to be estimated from regression Y_i on the kernel basis functions. The quadratic optimization problem in the DC algorithm can be solved using standard R functions such as *solve_osqp()* from package *osqp* (`https://cran.r-project.org/web/packages/osqp/index.html`). For both O-learning and MRL we determine $\widehat{\gamma}$ via (3.9) where we set shifting parameter $\eta = 10^{-4}$ and bisection termination condition $\epsilon = 10^{-3}$. We also include Q-learning for comparison, where at each stage of the backward learning, the Q-function is estimated using linear regression with the kernel basis functions and their interactions with treatments as predictors. Finally, to examine the impact of imposing risk control when learning DTRs, we also estimate the unconstrained optimal DTRs by setting $\tau = \infty$.

All simulation studies are repeated 500 times for each setting. To evaluate the performance of different methods, an independent testing dataset of sample size 5,000 is generated, and the estimated reward and risk on the independent testing data from each method are reported. To further quantify the benefit-risk tradeoff, we also report the efficacy ratio, one common measure to evaluate the benefit-risk tradeoff (Guo et al., 2010) defined as $r(\mathcal{D}) = (E^{\mathcal{D}}[Y] - E^{\mathcal{D}_0}[Y])/(E^{\mathcal{D}}[R] - E^{\mathcal{D}_0}[R])$, where $\mathcal{D}$ denotes the treatment rules being assessed and $\mathcal{D}_0$ represents the standard treatments. In our simulation study, the standard treatments are selected to be the safest treatment rules which induce the lowest cumulative risk among all four possible one-size-fits-all treatment rules. Since the standard comparison is set to be the treatment that yields the lowest expected risk, a higher efficacy ratio indicates that the treatment rule will gain more reward under the same risk increment as compared to a treatment rule with a lower efficacy ratio. Clearly, a treatment rule with a large efficacy ratio is preferable.

Table 3.1 presents the simulation results. For the first linear simulation setting, we note that when no risk constraint is imposed, the expected adverse risk under the unconstrained optimal rules is greater than 3.2, which is significantly higher than the prespecified risk constraint of $\tau = 1$. When the risk constraint is imposed, using either MRL, OWL, AOWL, or Q-learning yields the rules that give an expected risk below

Table 3.1: Complete simulation results for simulation studies.

			Setting I			**Setting II**		
Kernel	n	Method	Testing Reward	Testing Risk	Efficacy Ratio	Testing Reward	Testing Risk	Efficacy Ratio
Linear	200	MRL	1.657(0.166)	0.835(0.169)	1.023(0.150)	2.681(0.046)	1.257(0.139)	1.585(0.064)
		OWL	1.792(0.118)	0.894(0.146)	1.289(0.181)	2.983(0.143)	1.376(0.203)	1.591(0.093)
		AOWL	1.875(0.122)	0.962(0.160)	1.290(0.187)	3.071(0.153)	1.477(0.229)	1.553(0.095)
		Q-Learning	1.844(0.107)	1.062(0.172)	1.043(0.149)	2.959(0.142)	1.516(0.208)	1.476(0.077)
		Unconstrained	2.726(0.055)	3.184(0.090)	0.555(0.015)	4.584(0.006)	4.674(0.028)	0.908(0.003)
	400	MRL	1.707(0.114)	0.863(0.125)	1.098(0.117)	2.681(0.000)	1.257(0.000)	1.528(0.000)
		OWL	1.866(0.092)	0.931(0.126)	1.350(0.157)	3.039(0.101)	1.385(0.144)	1.605(0.065)
		AOWL	1.926(0.089)	0.988(0.127)	1.329(0.134)	3.093(0.109)	1.451(0.153)	1.576(0.064)
		Q-Learning	1.837(0.074)	1.028(0.110)	1.089(0.107)	2.965(0.110)	1.506(0.153)	1.482(0.057)
		Unconstrained	2.757(0.045)	3.223(0.057)	0.560(0.013)	4.587(0.005)	4.685(0.017)	0.907(0.002)
Gaussian	200	MRL	1.460(0.168)	0.726(0.188)	0.845(0.385)	2.805(0.128)	1.257(0.170)	1.586(0.086)
		OWL	1.738(0.105)	1.082(0.170)	0.883(0.213)	2.826(0.171)	1.355(0.231)	1.530(0.107)
		AOWL	1.812(0.140)	1.021(0.192)	1.025(0.183)	3.141(0.177)	1.793(0.277)	1.378(0.087)
		Q-Learning	1.912(0.129)	1.011(0.178)	1.267(0.176)	3.122(0.152)	1.524(0.213)	1.544(0.085)
		Unconstrained	2.825(0.010)	3.452(0.037)	0.535(0.006)	4.587(0.000)	4.702(0.000)	0.905(0.000)
	400	MRL	1.609(0.099)	0.805(0.099)	1.076(0.168)	2.895(0.125)	1.289(0.147)	1.611(0.080)
		OWL	1.784(0.076)	1.042(0.128)	1.052(0.235)	2.891(0.112)	1.333(0.158)	1.567(0.093)
		AOWL	1.849(0.101)	0.996(0.119)	1.200(0.150)	3.168(0.137)	1.740(0.201)	1.419(0.074)
		Q-Learning	1.928(0.087)	0.986(0.122)	1.337(0.141)	3.115(0.099)	1.483(0.147)	1.568(0.063)
		Unconstrained	2.825(0.021)	3.382(0.077)	0.554(0.016)	4.587(0.000)	4.702(0.000)	0.905(0.000)

Notes: Testing reward, testing risk, and efficacy ratio are reported in *median(dev)* format. *dev* denotes the median of the absolute value of the difference between the estimated value and the median estimated value.

or close to the risk constraint on the independent testing data. This suggests that the proposed estimation procedure is effective in finding the estimated rules that meet the risk restriction. In terms of the reward outcome, the theoretical maximum reward under the risk constraint $\tau = 1$ is approximately 2.17. As shown in Table 3.1, we note that the testing rewards of all four methods are all close to the optimal value. This demonstrates that our proposed estimation procedure does find the treatment rules that improve the beneficial reward while preserving safety. For O-learning, using either OWL or AOWL yields a similar result, with OWL having better performance in risk control. Comparing different methods, under the linear kernel, MRL and OWL tend to yield more stable and safer rules with median testing risk strictly lower than the risk constraint and with a smaller variability. In contrast, AOWL and Q-learning tend to underestimate the expected risk, leading to a testing risk close to or slightly higher than the risk constraint with larger variability. The performance of the four methods under the Gaussian kernel is similar to the performance under the linear kernel, except that OWL also tends to underestimate the expected risk and produces higher risk under the Gaussian kernel. In general, MRL tends to have the best risk control compared to OWL, AOWL, and Q-learning. No significant difference is observed between different kernels.

For the second simulation setting, the risk under the unconstrained optimal rules can be as high as 4.6 when no risk constraint is imposed, which is also significantly higher than the risk constraint $\tau = 1.5$. When risk restriction is imposed, and the linear kernel is used, the result shows that all methods can still well control the risk below or close to the risk constraint on the testing data. In terms of reward, the theoretical maximum reward under $\tau = 1.5$ is approximately 3.29. The expected reward on testing data using the linear kernel are all below but also close to the theoretical optimal reward indicating that our proposed estimation procedure can maintain its performance and balance reward and adverse risk under a different simulation setting. Compared with AOWL and Q-learning, MRL and OWL still shows better control of the adverse risk, with MRL having stricter risk control and more stability. When the Gaussian kernel is used, the performance of MRL, OWL, and Q-learning slightly improve, but AOWL starts to underestimate the risk with testing risk considerably exceeding $\tau = 1.5$. When the sample size is increased, all four methods using either linear or Gaussian kernel will improve, but AOWL under Gaussian kernel remains to underestimate the risk. The simulation results indicate that under this nonlinear setting, MRL, OWL, and Q-learning can still perform well, with MRL showing better control of the risk similar to the first simulation setting, while AOWL under the Gaussian kernel fails to strictly control the risk.

In terms of the efficacy ratio, in the first simulation setting, OWL has a higher efficacy ratio under the linear kernel, and Q-learning has a higher value under the Gaussian kernel. In the second simulation setting, MRL has roughly the same high efficacy ratio as OWL and a higher efficacy ratio than AOWL and Q-learning under the linear kernel. When the Gaussian kernel is used, MRL tends to achieve the highest efficacy ratio among the four methods. In summary, these simulation results show that MRL, OWL, AOWL, and Q-learning using our proposed estimation procedure can yield the rules that meet the risk restriction while maintaining a high beneficial reward for a CBR problem. MRL tends to have an overall better performance with stricter risk control. Additionally, we conduct a simulation setting with $T = 4$ and allow the treatment assignment probabilities to depend on the covariates at later stages. The results are reported in Section 3.10 and show a similar conclusion: MRL has better risk control and a higher efficacy ratio compared to the other competing methods.

3.5 Application to DURABLE Trial

In this section, we apply our proposed CBR framework to the DURABLE study and the description of DURABLE can be found in Section 2.5. For T2D, the A1c level is the main efficacy outcome measuring the patient's health condition, and we choose HbA1c reduction by the end of week 48 compared to baseline level (week 0) as the cumulative reward outcome. Weight gain is one of the common long-term side effects of insulin therapy, and in the analysis, we choose cumulative risk outcome to be BMI change by the end of week 48, with a lower BMI increment indicating a better risk control. Due to the DURABLE study design, not all patients were re-randomized during the second phase of the study. To implement CBR, we again make a practical assumption like Section 2.4 that for patients who had reached HbA1c of 7% at the end of the first phase and entered the maintenance study, their treatments received at the second phase of the study were optimal. Hence, for the patients who entered the maintenance study, only their first-stage treatment needs to be evaluated and optimized. With this assumption, for MRL, we solve the modified empirical problem

$$\begin{aligned}\max_{(f_1,f_2)\in\mathcal{G}_1\times\mathcal{G}_2}\ & \frac{1}{n}\sum_{i\in I_1}\widehat{O}_i^{+}\frac{\min(\psi(A_{i1}f_1(H_{i1})),\psi(A_{i2}f_2(H_{i2})))}{p(A_{i1}|H_{i1})p(A_{i2}|H_{i2})}\\ &+\frac{1}{n}\sum_{i\in I_1}\sum_{a_t\in\{-1,1\},a_t\neq A_{it}}\widehat{O}_i^{-}\frac{\min(\psi(a_1f_1(H_{i1})),\psi(a_2f_2(H_{i2})))}{p(A_{i1}|H_{i1})p(A_{i2}|H_{i2})}\\ &+\frac{1}{n}\sum_{i\in I_2}\widehat{O}_i^{+}\frac{\psi(A_{i1}f_1(H_{i1}))}{p(A_{i1}|H_{i1})}+\frac{1}{n}\sum_{i\in I_2}\sum_{a_1\in\{-1,1\},a_1\neq A_{i1}}\widehat{O}_i^{-}\frac{\psi(a_1f_1(H_{i1}))}{p(A_{i1}|H_{i1})}-\sum_{t=1}^{2}\lambda_{n,t}\|f_t\|_{\mathcal{G}_t}^2,\end{aligned}$$

where I_1 and I_2 denote the set of patients who entered the intensification study and maintenance study, respectively. For Q-learning and O-learning, in stage 2, only patients who entered the intensification study are used for estimation. In the first stage, we use all patients for estimation but update the outcomes by their estimated Q-functions for Q-learning, or inverse probability estimator $Y_i\mathbb{I}(A_{i2}\widehat{f}_2(H_{i2}) > 0)/p(A_{i2}|H_{i2})$ for the patients from the intensification study when applying O-learning.

To estimate the optimal decision rules, we extract 20 relevant feature variables as the baseline variables H_1. These variables include baseline HbA1c level, heart rate, systolic/diastolic blood pressure, body weight, body height, BMI, and 7 points self-monitored blood glucose measured at week 0, and demographic variables including patient's age, gender, along with the duration of T2D and 3 indicator variables of whether patients were receiving oral antihyperglycemic agent of metformin, thiazolidinedione, or sulfonylureas. The second stage feature variables H_2 include all variables in H_1, as well as the patient's stage 1 treatment assignment, heart rate, systolic/diastolic blood pressures, HbA1c, body weight, body height, BMI and the same 7 points self-monitored blood glucose measured at the beginning of phase 2 study (24 weeks). All covariates are normalized to have mean zero and variance one.

Our analysis includes 573 patients from the intensification study and 771 patients from the maintenance study. To reduce the impact due to sampling variability, we repeatedly sample 30% of patients as training data and use the remaining 70% of data as testing data to evaluate the performance of estimated rules. The population average BMI change is approximately 1.5, and we repeat the analysis with τ from 1.5 to 1.65, increased by 0.05. We still implement both OWL and AOWL for O-learning, and the tuning grid of $(n\lambda_1, n\lambda_2)$ for O-learning and MRL is chosen to be the same as used in the simulation study but exclude the pairs when $\max(\lambda_1, \lambda_2)/\min(\lambda_1, \lambda_2) > 4$, with $\eta = 10^{-4}$ and $\epsilon = 10^{-3}$. Preliminary exploratory analyses indicate that the optimal decision function is highly nonlinear, and hence we use the Gaussian kernel and select the bandwidth also similar to the simulation studies. For each risk constraint, we repeat the analysis 100 times using MRL, OWL, AOWL, and Q-learning. For comparison, we also conduct MRL and set the risk constraint to be infinite to estimate the globally optimal decision rules with no risk control.

The estimated reward and risk on testing data are reported in Table 3.2. From the results, we first note that when no constraint is imposed, the unconstrained estimated optimal treatment rules will yield an overall increment of BMI approximately equal to 1.75 with a gain of 1.70% HbA1c reduction over 48 weeks period, which is close to the expected BMI increment and HbA1c reduction induced by the most aggressive LMx2-MMx3 rules among all four one-size-fits-all rules shown in Table 3.3. In contrast, when the risk

τ	Method	BMI Increment	HbA1c Reduction	Efficacy Ratio	Percentage of LMx2 during Phase I	Percentage of LMx2/MMx3 during Phase II
1.50	MRL	1.508(0.190)	1.589(0.089)	0.516(0.118)	0.1	100.0
	OWL	1.493(0.082)	1.555(0.057)	0.482(0.177)	38.2	74.2
	AOWL	1.516(0.080)	1.585(0.064)	0.450(0.130)	53.5	85.5
	Q-learning	1.556(0.097)	1.583(0.057)	0.428(0.113)	64.3	64.2
1.55	MRL	1.541(0.187)	1.609(0.081)	0.547(0.150)	24.6	100.0
	OWL	1.491(0.083)	1.563(0.053)	0.482(0.164)	36.0	81.7
	AOWL	1.536(0.074)	1.593(0.058)	0.471(0.168)	55.7	88.1
	Q-learning	1.568(0.097)	1.584(0.046)	0.428(0.104)	69.3	86.1
1.60	MRL	1.615(0.139)	1.609(0.079)	0.527(0.115)	99.8	100.0
	OWL	1.506(0.082)	1.561(0.051)	0.482(0.168)	37.3	79.9
	AOWL	1.567(0.086)	1.596(0.049)	0.466(0.114)	65.8	93.9
	Q-learning	1.580(0.102)	1.585(0.055)	0.436(0.103)	72.9	68.8
1.65	MRL	1.622(0.140)	1.614(0.086)	0.502(0.125)	99.8	100.0
	OWL	1.512(0.084)	1.556(0.056)	0.455(0.150)	45.0	83.2
	AOWL	1.552(0.088)	1.589(0.058)	0.455(0.165)	61.6	86.4
	Q-learning	1.594(0.092)	1.592(0.056)	0.423(0.104)	75.4	70.9
∞	Unconstrained	1.746(0.056)	1.694(0.056)	0.501(0.065)	100.0	100.0

Table 3.2: Analysis results for the DURABLE study. Results are reported in median(dev) format as the simulation study. BMI, HbA1c, and efficacy ratio are estimated on repeatedly sampled testing data. Efficacy ratios are calculated using G-BBT as reference rules. The percentage of LMx2 during phase I is the proportion of patients recommended with LMx2 treatment as initial treatment. The percentage of LMx2/MMx3 during phase II is the proportion of patients recommended with LMx2/MMx3 as second phase intensification treatment when failed to reach HbA1c $\leq 7.0\%$. Treatment recommendation is estimated for all patients using maximum voting based on 100 repeated analyses.

constraint is imposed, the expected increment of BMI can decrease from 1.60 to roughly 1.50, which is significantly lower than the unconstrained expected BMI increment at the price of a smaller HbA1c reduction decreasing from 1.61% to roughly 1.56%. Comparing four different methods, both MRL, OWL, AOWL, and Q-learning can yield treatment rules with an expected BMI increment below or close to the prespecified constraint under different choices of τ. However, in terms of beneficial reward, MRL can always lead to an equal or higher HbA1c reduction than OWL, AOWL, and Q-learning under all choices of τ. The results indicate that all four proposed methods will still successfully yield treatment rules that meet the risk restriction in real data application, and MRL tends to have top performance with both ideal control over risk and higher gain in beneficial reward compared to OWL, AOWL, and Q-learning.

We also evaluate the capability of balancing the benefit-risk via efficacy ratio against the standard treatment rules. According to Table 3.3, assigning all patients with insulin glargine as the initial treatment and

Table 3.3: Mean HbA1c reduction/BMI increment at week 48 under 4 one-size-fits-all treatment rules. Efficacy ratios are calculated using G-BBT as reference rules.

Treatment Rules	Mean BMI Increment	Mean HbA1c Reduction	Efficacy Ratio
LMx2-MMx3	1.738	1.699	0.519
LMx2-BBT	1.683	1.640	0.456
G-LMx2	1.437	1.563	0.610
G-BBT	1.205	1.422	Ref

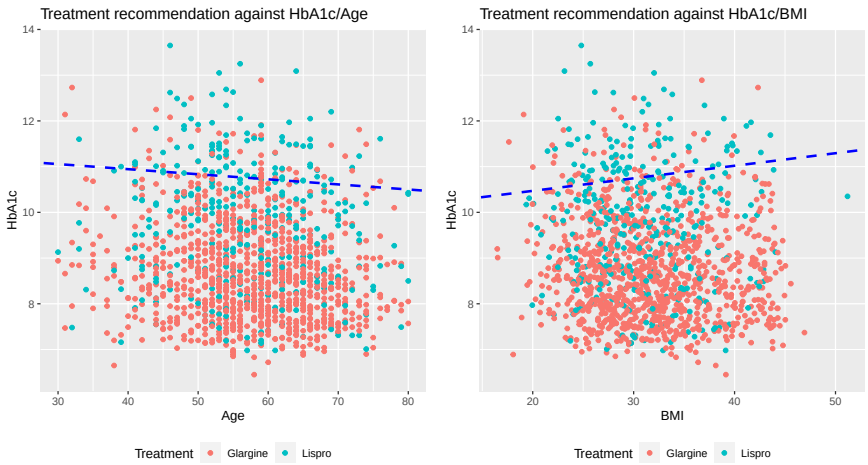

Figure 3.2: Scatter plot of baseline HbA1c against age or baseline BMI. The color indicates the treatment recommendation estimated from MRL given $\tau = 1.55$. The linear decision boundary is calculated using logistic regression.

reassigning patients who fail to reach HbA1c 7% with BBT (G-BBT) yields the lowest risk with an average increment of BMI equal to 1.20 and HbA1c reduction equal to 1.42%. When G-BBT is selected to be the standard treatment, the results in Table 3.2 show that MRL still achieves an overall higher efficacy ratio than OWL, AOWL, and Q-learning. Moreover, MRL can yield treatment rules with an efficacy ratio close to or higher than the unconstrained optimal rules for all τ. This demonstrates that considering risk impact can also lead to a better treatment regimen design with more efficient benefit-risk balancing in DURABLE, and MRL tends to have better performance compared with OWL, AOWL, and Q-learning.

The proportion of patients recommended to each treatment arm by MRL is displayed in Table 3.2. MRL recommends almost all patients to G-LMx2 rules when the risk constraint is $\tau = 1.50$ and almost all patients to LMx2-MMx3 when the risk constraint is above 1.60. When $\tau = 1.55$, the treatment rules estimated from MRL will recommend LMx2-MMx3 to 24.6% of patients and G-LMx2 to the remaining patients. We show the treatment recommendation distribution in Figure 3.2, which suggests that younger and more overweight patients with lower baseline HbA1c are more likely to be recommended with the less intensive

insulin glargine therapy as the initial treatment. Younger patients with lower HbA1c levels are in better health conditions, and thus less intensive insulin therapy is preferable following the general T2D management guidance. Moreover, clinical studies suggest that obesity is associated with insulin resistance (Saha and Schwarz, 2017), and additional exogenous insulin will cause increased weight gain among T2D patients with insulin resistance (McFarlane, 2009). Therefore, patients with higher BMI are more likely to be resistant to insulin and should be treated with less intensive insulin therapy to reduce the risk of weight gain unless the patient's HbA1c level is high. Figure 3.2 indicates that the treatment rules learned from MRL are consistent with clinical evidence and practices. These results suggest that our proposed method is capable of learning treatment rules that are clinically meaningful in practice while meeting the risk constraint in a real-world application.

3.6 Discussion

In this chapter, we proposed a general estimation procedure to solve the CBR problem where the goal is to find optimal treatment rules that maximize the cumulative reward, but the induced risk is no more than a pre-specified threshold. Our approach converts constrained optimization into solving a series of unconstrained optimization problems. Consequently, the proposed procedure can be easily implemented using many existing standard DTR methods or using the proposed simultaneous algorithm. Simulation studies and the real data example indicated that either using MRL, O-learning, or Q-learning along with the proposed procedure would yield well-performed DTRs with the risk being controlled under or close to the risk constraint.

The proposed MRL can be used to solve unconstrained DTRs problems. The key advantage of MRL is that it estimates the DTRs jointly without distinguishing early stages from later stages. This special property would allow one to impose a joint structure on stagewise decision rules to conduct simultaneous variable selection across all stages, which is not feasible using backward Q-learning or O-learning. From the computational perspective, the DC algorithm may be inefficient with a large sample size or a large number of stages. Possible improvement can be to use coordinate descent along with stochastic gradient descent to reduce the computation cost of each DC iteration or to consider a more smooth approximation to the objective functions so that quasi-Newton's methods are applicable.

In some applications, more than one adverse event needs to be controlled in the long run, and CBR can be generalized to handle multiple risk constraints. In addition, CBR can be extended to solve combined

short-term stagewise risk control along with cumulative risk control in DTRs. However, for the case with multiple constraints, further development is needed to address the computational challenges. The proposed method can also be extended to consider multicategory or even continuous treatments (Laber et al., 2018).

Although we focused on clinical trials, our method is applicable to analyze observational studies except that the treatment assignment probabilities, i.e., propensity scores, must be estimated from the data. Theoretically, when the propensity scores are estimated via parametric models, following the same arguments in Chen, Zeng and Wang (2021), the extra error due to this estimation, which is of order $O(n^{-\frac{1}{2}})$, should not affect the error bounds given in our results. Finally, when the positivity assumption is a concern, especially for observational studies, some existing techniques such as pessimistic learning (Fu et al., 2022; Zhou et al., 2023) can be incorporated into our framework to learn suboptimal DTRs by working on pessimistic Q-functions.

3.7 Details of the DC algorithm for solving MRL

The complete algorithm for solving the CBR problem via bisection is presented as Algorithm 2. To solve MRL, we consider the equivalent minimization problem

$$\min_{(f_1,...,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \sum_{t=1}^{T} C_{n,t}\|f_t\|_{\mathcal{G}_t}^2 - \sum_{i=1}^{n} O_i \frac{\min(\psi(A_{i1}f_1(H_{i1})),...,\psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} - \sum_{i\in I^-} O_i^- \frac{\min(\psi(|f_1(H_{i1})|),...,\psi(|f_T(H_{iT})|))}{\prod_{t=1}^{T} p(A_{it}|H_{it})}.$$

where we let $C_{n,t} = n\lambda_{n,t}$ in this section, and I^- denotes the indices of subjects whose response variable is negative, i.e., $\{i : O_i < 0\}$. Original expression (6) can be recovered by adding and subtracting the term

$$\frac{1}{n}\sum_{i=1}^{n} O_{i,\gamma}^- \frac{\min(\psi(A_{i1}f_1(H_{i1})),...,\psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^{T} p(A_{it}|H_{it})}.$$

The derivation is based on utilizing the decomposition property

$$\min(\psi(x_1),...,\psi(x_T)) = \min(x_1,...,x_T,1) - \min(x_1,...,x_T,0)$$

Algorithm 2 General Algorithm for Solving CBR Problem

Input: Training data $(Y, R, H_1, ..., H_T)$, risk constraint τ and termination condition ϵ.

Start with $\gamma_{\max} = 0$ and $\gamma_{min} = 1$, and solve the unconstrained problem

$$\max_D E\left[\{(1-\gamma)Y - \gamma R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]$$

for $\gamma = \gamma_{\min}$ and $\gamma = \gamma_{\max}$.

Let $\mathcal{D}^*_{\gamma_{\min\,/\,\max}}$ denote the solution associated with $\gamma_{\min}/\gamma_{\max}$ respectively, we define

$$\tau_{\max} = E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma_{\max}}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right],$$

$$\tau_{\min} = E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t = D^*_{t,\gamma_{\min}}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right].$$

if $\tau \geq \tau_{\max}$ **then** return $\mathcal{D}^*_{\gamma_{\max}}$

else

while $|\gamma_{\max} - \gamma_{min}| > \epsilon$ **do**

Set $\gamma = \frac{1}{2}(\gamma_{\min} + \gamma_{\max})$ and solve the unconstrained problem associated with new γ

Obtain the current risk

$$\tau_{\text{now}} = E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t = D^*_{t,\gamma}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right].$$

if $\tau_{\text{now}} \leq \tau$ **then** update $\gamma_{\min}$ by γ

else update $\gamma_{\max}$ by γ.

end if

end while

end if

Output: Return $\mathcal{D}^*_{\gamma_{\min}}$ as the solution under risk constraint τ.

and noticing that

$$\begin{aligned}
&-\min(\psi(|f_1(H_1)|), ..., \psi(|f_T(H_T)|)) \\
=&\max(-|f_1(H_1)|, ..., -|f_T(H_T)|, -1) - \max(-|f_1(H_1)|, ..., -|f_T(H_T)|, 0) \\
=&\max(-|f_1(H_1)|, ..., -|f_T(H_T)|, -1) \\
=&\max(\sum_{t\neq 1}|f_t(H_t)|, ..., \sum_{t\neq T}|f_t(H_t)|, \sum_{t=1}^T |f_t(H_t))| - 1) - \sum_{t=1}^T |f_t(H_t)|.
\end{aligned}$$

Two equations above imply that the objective function can be written as the difference of two convex functions $\widetilde{S}_1 - \widetilde{S}_2$ where

$$\begin{aligned}\widetilde{S}_1 =& \sum_{i=1}^{n} |O_i| \max(-A_{i1} f_1(H_{i1}), ..., -A_{iT} f_T(H_{iT}), -d_i) \\ &+ \sum_{i \in I^-} O_i^- \max(\sum_{t \neq 1} |f_t(H_{it})|, ..., \sum_{t \neq T} |f_t(H_{it})|, \sum_{t=1}^{T} |f_t(H_{it})| - 1) \\ &+ \sum_{t=1}^{T} C_{n,t} \|f_t\|_{\mathcal{G}_t}^2,\end{aligned}$$

$$\begin{aligned}\widetilde{S}_2 =& \sum_{i=1}^{n} |O_i| \max(-A_{i1} f_1(H_{i1}), ..., -A_{iT} f_T(H_{iT}), -(1 - d_i)) \\ &+ \sum_{i \in I^-} O_i^- \sum_{t=1}^{T} |f_t(H_{it})|.\end{aligned}$$

where $d_i = \mathbb{I}\{O_i > 0\}$. Assuming that $\{\mathcal{G}_t\}_{t=1}^T$ are RKHS, given initial coefficients $\{\beta_t^{(0)}\}_{t=1}^T$ and intercepts $\{\beta_{0t}^{(0)}\}_{t=1}^T$, one can use standard DC algorithm to solve the optimization problem where we update $(\beta_1^{(s+1)}, \beta_{01}^{(s+1)}, ..., \beta_T^{(s+1)}, \beta_{0T}^{(s+1)})$ sequentially via solving the optimization problem

$$\begin{aligned}\underset{\beta_1, \beta_{01}, ..., \beta_T, \beta_{0T}}{\arg\min} \quad & \sum_{t=1}^{T} C_{n,t} \beta_t^T K_t \beta_t + \sum_{t=1}^{T} C_{n,t} \beta_{0t}^2 + \widetilde{S}_1(\beta_1, \beta_{01}, ..., \beta_T, \beta_{0T}) \\ & - \sum_{t=1}^{T} \frac{\partial \widetilde{S}_2}{\partial \beta_t}(\beta_1^{(s)}, \beta_{01}^{(s)}, ..., \beta_T^{(s)}, \beta_{0T}^{(s)})(\beta_t - \beta_t^{(s)}) \\ & - \sum_{t=1}^{T} \frac{\partial \widetilde{S}_2}{\partial \beta_{0t}}(\beta_1^{(s)}, \beta_{01}^{(s)}, ..., \beta_T^{(s)}, \beta_{0T}^{(s)})(\beta_{0t} - \beta_{0t}^{(s)}).\end{aligned}$$

Here, we add additional term $\sum_{i=1}^n C_{n,t} \beta_{0t}^2$ to avoid determining optimal β_{0t} through exhausted search and speed up the estimation. K_t denotes the kernel matrix of stage t where the (i, j) entry is equal to $\langle H_{it}, H_{it} \rangle$ and $\frac{\partial \widetilde{S}_2}{\partial \beta}$ denotes the subgradients of $\widetilde{S}_2$ w.r.t. β.

To avoid involving subgradients and further improve the numerical performance, we adopt the smooth approximation technique from Nesterov (2005) and consider the smooth approximation function

$$l(x_1, ..., x_K) = \mu \log \left(\frac{1}{K} \sum_{k=1}^{K} e^{\frac{x_k}{\mu}} \right)$$

to replace non-smooth function $\max(x_1, ..., x_K)$. When μ is sufficient small, l is a good differentiable approximation of $\max(x_1, ..., x_K)$. Simulation shows that $\mu = 10^{-8}$ is sufficiently small and we recommend using $\mu = 10^{-8}$ in MRL. By introducing the smooth approximation function l, we can replace $\widetilde{S}_2$ by differentiable function

$$\begin{aligned}\bar{S}_2 = \sum_{i=1}^{n} |O_i| \mu \log\left(\frac{1}{T+1}\sum_{t=1}^{T} e^{-A_{it} f_t(H_{it})/\mu} + \frac{1}{T+1} e^{-(1-d_i)/\mu}\right) \\ + \sum_{i\in I^-} O_i^- \sum_{t=1}^{T} \mu \log\left(\frac{1}{2} e^{f_t(H_{it})/\mu} + \frac{1}{2} e^{-f_t(H_{it})/\mu}\right)\end{aligned}$$

By introducing slack variables ξ_i, η_i, ω_{it}, the optimization is equivalent to

$$\begin{aligned}\min_{\beta_t, \beta_{0t}, \xi_i, \eta_i, \omega_{it}} & \sum_{t=1}^{T} C_{n,t} \beta_t^T K_t \beta_t + \sum_{t=1}^{T} C_{n,t} \beta_{0t}^2 + \sum_{i=1}^{n} |O_i| \xi_i + \sum_{i\in I^-} O_i^- \eta_i \\ & - \sum_{t=1}^{T} \frac{\partial \bar{S}_2}{\partial \beta_t}(\beta_1^{(s)}, \beta_{01}^{(s)}, ..., \beta_T^{(s)}, \beta_{0T}^{(s)})(\beta_t - \beta_t^{(s)}) \\ & - \sum_{t=1}^{T} \frac{\partial \bar{S}_2}{\partial \beta_{0t}}(\beta_1^{(s)}, \beta_{01}^{(s)}, ..., \beta_T^{(s)}, \beta_{0T}^{(s)})(\beta_{0t} - \beta_{0t}^{(s)}), \qquad (3.12)\end{aligned}$$

$$\text{subject to } \xi_i \geq -A_{it}(K_{it}\beta_t + \beta_{0t}), \quad \xi_i \geq -d_i, \quad \eta_i \geq \sum_{s\neq t} \omega_{is}, \quad \eta_i \geq \sum_{t=1}^{T} \omega_{it} - 1, \quad \forall\{i,t\},$$

$$\omega_{it} \geq -(K_{it}\beta_t + \beta_{0t}), \quad \omega_{it} \geq K_{it}\beta_t + \beta_{0t}, \quad \forall i \in I^- \text{ and } t \in \{1, ..., T\}.$$

Here, K_{it} denotes the i-th row of kernel matrix K_t. The Lagrange function of (3.12) is given by

$$\begin{aligned}L(u_{it}, u_i, v_{it}, v_i, l_{it}^+, l_{it}^-) = & \sum_{t=1}^{T} C_{n,t} \beta_t^T K_t \beta^T + \sum_{t=1}^{T} C_{n,t} \beta_{0t}^2 + \sum_{i=1}^{n} |O_i| \xi_i + \sum_{i\in I^-} O_i^- \eta_i \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T} C_{it}^{(s)} |O_i| A_{it}(K_{it}\beta_t + \beta_{0t}) + \sum_{i\in I^-}\sum_{t=1}^{T} C_{it}^{-(s)} O_i^-(K_{it}\beta_t + \beta_{0t}) \\ & - \sum_{i=1}^{n}\sum_{t=1}^{T} u_{it}(\xi_i + A_{it}(K_{it}\beta_t + \beta_{0t})) - \sum_{i=1}^{n} u_i(\xi_i + d_i) \\ & - \sum_{i\in I^-}\sum_{t=1}^{T} v_{it}(\eta_i - \sum_{s\neq t} \omega_{is}) - \sum_{i\in I^-} v_i(\eta_i - \sum_{t=1}^{T} \omega_{it} + 1) \\ & - \sum_{i\in I^-}\sum_{t=1}^{T} l_{it}^+(\omega_{it} + K_{it}\beta_t + \beta_{0t}) - \sum_{i\in I^-}\sum_{t=1}^{T} l_{it}^-(\omega_{it} - K_{it}\beta_t - \beta_{0t}),\end{aligned}$$

where

$$C_{it}^{(s)} = \frac{e^{-A_{it}(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})/\mu}}{\sum_{t=1}^{T} e^{-A_{it}(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})/\mu} + e^{-(1-d_i)/\mu}}, \tag{3.13}$$

$$C_{it}^{-(s)} = \frac{e^{-(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})} - e^{(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})}}{e^{(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})} + e^{-(K_{it}\beta_t^{(s)}+\beta_{0t}^{(s)})}}. \tag{3.14}$$

Taking derivatives w.r.t. ξ_i, η_i and ω_{it} and setting the derivatives equal to 0, we can obtain

$$|O_i| - \sum_{t=1}^{T} u_{it} - u_i = 0,$$

$$O_i^- - \sum_{t=1}^{T} v_{it} - v_i = 0,$$

$$-l_{it}^+ - l_{it}^- + \sum_{s\neq t} v_{it} + v_i = 0.$$

The equations above yield

$$u_i = |O_i| - \sum_{t=1}^{T} u_{it},$$

$$v_{it} = O_i^- - l_{it}^+ - l_{it}^-,$$

$$v_i = \sum_{t=1}^{T} (l_{it}^+ + l_{it}^-) - (T-1)O_i^-.$$

Plugging in u_i, v_{it} and v_i back to L and taking derivatives w.r.t. β_t and β_{0t} yields

$$\beta_t = \frac{1}{C_{n,t}}\big[A_t(U_t - WC_t^{(s)}) + I_n^-(L_t^+ - L_t^- - W^-C_t^{-(s)})\big]$$

$$\beta_{0t} = \frac{1}{C_{n,t}}\big[\mathbb{1}_n A_t(U_t - WC_t^{(s)}) + \mathbb{1}_n^-(L_t^+ - L_t^- - W^-C_t^{-(s)})\big].$$

Here, $\mathbb{1}_n$ and $\mathbb{1}_n^{(-)}$ denote the one vector of length n and size of I^-, I_n^- denote the submatrix of identical matrix I_n where only columns within index set I^- are kept, $W = \text{diag}\{|O_i|\}$, $W^- = \text{diag}\{O_i^-\}$, $A_t = \text{diag}\{A_{it}\}$, $U_t = \{u_{1t}, ..., u_{nt}\}^T$, $L_t^+ = \{l_{it}^+\}_{i\in I^-}^T$, $L_t^- = \{l_{it}^-\}_{i\in I^-}^T$, $C_t^{(s)} = \{C_{1t}^{(s)}, ..., C_{nt}^{(s)}\}^T$, $C_t^{-(s)} = \{C_{it}^{-(s)}\}_{i\in I^-}^T$ defined in (3.13) and (3.14).

Plug in the expression of β_t and β_{0t} back to L and include the constraints $u_{it} \geq 0, \quad u_i \geq 0, v_{it} \geq 0$, $v_i \geq 0, l_{it}^+ \geq 0, l_{it}^- \geq 0$, we can obtain that the dual problem of (3.12) is given by

$$
\begin{aligned}
\min_{U_t, L_t^+, L_t^-} \quad & \sum_{t=1}^{T} C_{n,t}^{-1} \left[A_t W C_t^{(s)} + I_n^- W^- C_t^{-(s)} - A_t U_t - I_n^- L_t^+ + I_n^- L_t^- \right]^T \cdot K \\
& \cdot \left[A_t W C_t^{(s)} + I_n^- W^- C_t^{-(s)} - A_t U_t - I_n^- L_t^+ + I_n^- L_t^- \right] \\
& + \sum_{t=1}^{T} C_{n,t}^{-1} \left[\mathbb{1}_n A_t W C_t^{(s)} + \mathbb{1}_n^- W^- C_t^{-(s)} - \mathbb{1}_n A_t U_t - \mathbb{1}_n^- L_t^+ + \mathbb{1}_n^- L_t^- \right]^T \\
& \cdot \left[\mathbb{1}_n A_t W C_t^{(s)} + \mathbb{1}_n^- W^- C_t^{-(s)} - \mathbb{1}_n A_t U_t - \mathbb{1}_n^- L_t^+ + \mathbb{1}_n^- L_t^- \right] \\
& - \sum_{i=1}^{n} \sum_{t=1}^{T} u_{it} d_i + \sum_{i \in I^-} \sum_{t=1}^{T} (l_{it}^+ + l_{it}^-) \\
\text{subject to} \quad & \sum_{t=1}^{T} u_{it} \leq |O_i|, \;\; u_{it} \geq 0, \;\; \forall \{i, t\} \\
& l_{it}^+ + l_{it}^- \leq O_i^-, \;\; \sum_{t=1}^{T} (l_{it}^+ + l_{it}^-) \geq (T-1) O_i^-, \\
& l_{it}^+ \geq 0, \;\; l_{it}^- \geq 0, \;\; \forall \, i \in I^-, \; t \in \{1, ..., T\},
\end{aligned}
$$

or equivalently

$$
\begin{aligned}
\min_{U_t, L_t^+, L_t^-} \quad & \sum_{t=1}^{T} C_{n,t}^{-1} \left[A_t W C_t^{(s)} + I_n^- W^- C_t^{-(s)} - A_t U_t - I_n^- L_t^+ + I_n^- L_t^- \right]^T \cdot (K + \mathbb{1}_n^T \mathbb{1}_n) \\
& \cdot \left[A_t W C_t^{(s)} + I_n^- W^- C_t^{-(s)} - A_t U_t - I_n^- L_t^+ + I_n^- L_t^- \right] \\
& - \sum_{i=1}^{n} \sum_{t=1}^{T} u_{it} d_i + \sum_{i \in I^-} \sum_{t=1}^{T} (l_{it}^+ + l_{it}^-) \\
\text{subject to} \quad & \sum_{t=1}^{T} u_{it} \leq |O_i|, \;\; u_{it} \geq 0, \;\; \forall \{i, t\} \\
& l_{it}^+ + l_{it}^- \leq O_i^-, \;\; \sum_{t=1}^{T} (l_{it}^+ + l_{it}^-) \geq (T-1) O_i^-, \\
& l_{it}^+ \geq 0, \;\; l_{it}^- \geq 0, \;\; \forall \, i \in I^-, \; t \in \{1, ..., T\},
\end{aligned}
$$

The last optimization is a standard quadratic optimization and can be efficiently solved using the standard method such ADMM solver (Stellato et al., 2020).

3.8 Proof of Lemma 3.1 and Lemma 3.2

3.8.1 Proof of Lemma 3.1

We first verify that $\mathfrak{R}(\gamma)$ and $\mathfrak{Y}(\gamma)$ are non-increasing functions for $\gamma \in [0,1]$. By definition, for any γ_1 and γ_2 we have

$$E\left[\{(1-\gamma_1)Y-\gamma_1 R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right] \geq E\left[\{(1-\gamma_1)Y-\gamma_1 R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]$$

and

$$E\left[\{(1-\gamma_2)Y-\gamma_2 R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right] \geq E\left[\{(1-\gamma_2)Y-\gamma_2 R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right].$$

By reorganizing and adding the two inequalities above, we have

$$\begin{aligned}
&(1-\gamma_1)(\mathfrak{Y}(\gamma_1)-\mathfrak{Y}(\gamma_2)) \\
=&(1-\gamma_1)\left(E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]-E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) \\
\geq&\gamma_1\left(E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]-E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) \\
=&\gamma_1(\mathfrak{R}(\gamma_1)-\mathfrak{R}(\gamma_2))
\end{aligned} \tag{3.15}$$

and

$$\begin{aligned}
&(1-\gamma_2)(\mathfrak{Y}(\gamma_1)-\mathfrak{Y}(\gamma_2)) \\
=&(1-\gamma_2)\left(E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]-E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) \\
\leq&\gamma_2\left(E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_1}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]-E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t=\mathcal{D}^*_{t,\gamma_2}(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) \\
=&\gamma_2(\mathfrak{R}(\gamma_1)-\mathfrak{R}(\gamma_2)).
\end{aligned} \tag{3.16}$$

Hence, suppose that $\gamma_1 < \gamma_2$ and $\mathfrak{R}(\gamma_1) < \mathfrak{R}(\gamma_2)$, then by combining (3.15) and (3.16) we can obtain

$$\frac{\gamma_1}{1-\gamma_1} \geq \frac{\gamma_2}{1-\gamma_2}$$

and consequently $\gamma_1 \geq \gamma_2$, which is a contradiction. Therefore, we must have $\mathfrak{R}(\gamma_2) \leq \mathfrak{R}(\gamma_1)$ for any $\gamma_1 < \gamma_2$. By exactly the same argument, we can also show that $\mathfrak{Y}(\gamma_2) \leq \mathfrak{Y}(\gamma_1)$ for any $\gamma_1 < \gamma_2$. Hence, we have shown that both $\mathfrak{R}(\gamma)$ and $\mathfrak{Y}(\gamma)$ are non-increasing functions of γ.

We now complete the proof of Lemma 3.1. Suppose that $\mathfrak{R}(\gamma^*) = \tau$ and let $\mathcal{D} = (\mathcal{D}_1, ..., \mathcal{D}_T)$ be arbitrary decision rules such that

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \tau,$$

then it is sufficient to verify that one must have

$$E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \mathfrak{Y}(\gamma^*).$$

We prove the result by contradiction. Suppose that

$$E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] > \mathfrak{Y}(\gamma^*), \tag{3.17}$$

then we consider two possible cases:

Case 1: Suppose that

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] = \mathfrak{R}(\gamma^*) = \tau,$$

then Assumption 3.4 implies that $\gamma^* < 1$ since τ is assumed to be greater than $\mathfrak{R}(1)$ and we have

$$\begin{aligned}
0 <&(1-\gamma^*)\left(E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \mathfrak{Y}(\gamma^*)\right)\\
=&\left((1-\gamma^*)E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \gamma^* E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right)\\
&-\left((1-\gamma^*)E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma^*}(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \gamma^* E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma^*}(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right),
\end{aligned}$$

which implies that $\mathcal{D}$ should be the optimal solution of unconstrained problem (3) associated with multiplier γ^* and is contradictory to the definition of $\mathcal{D}^*_{\gamma^*}$.

Case 2: Suppose that

$$\mathfrak{R}(1) < E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] = \tau' < \tau = \mathfrak{R}(\gamma^*),$$

then by the continuity and non-increasing property of $\mathfrak{R}(\gamma)$ and recall that $\mathfrak{R}(1) < \tau$, we can always find γ' such that $\mathfrak{R}(\gamma') = \tau'$ for some $\gamma^* < \gamma' < 1$. Hence, by definition we have

$$\begin{aligned}
0 \leq & \left((1-\gamma^*)E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma'}(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \gamma^* E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma'}(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right) \\
& - \left((1-\gamma^*)E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \gamma^* E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right) \\
= & (1-\gamma^*)\left(E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma'}(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right) \\
= & (1-\gamma^*)\left(\mathfrak{Y}(\gamma') - E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\right),
\end{aligned}$$

which implies that

$$E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \mathfrak{Y}(\gamma').$$

On the other hand, the non-increasing property of $\mathfrak{Y}(\gamma)$ indicates that $\mathfrak{Y}(\gamma') \leq \mathfrak{Y}(\gamma^*)$ and, hence, we have

$$E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \mathfrak{Y}(\gamma^*),$$

which is contradictory to the assumption (3.17).

Hence, combine *Case 1* and *Case 2* and we obtain that

$$E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \mathfrak{Y}(\gamma^*)$$

holds for any decision rules $\mathcal{D}$ such that

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T}p(A_t|H_t)}\right] \leq \tau,$$

which indicates that $\mathcal{D}^*_{\gamma^*}$ is one of the optimal solutions of (3.1) under risk constraint τ.

3.8.2 Proof of Lemma 3.2

First note that for arbitrary weight O, by adding an additional term $E\big[O^-/\prod_{t=1}^T p(A_t|H_t)\big]$ which is independent of the choice of decision functions $\{f_t\}_{t=1}^T$, one can notice that the optimization problem

$$\underset{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T}{\arg\max} E\bigg[O\frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg]$$

is equivalent to maximizing

$$\begin{aligned}
&E\bigg[O\frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg] + E\bigg[\frac{O^-}{\prod_{t=1}^T p(A_t|H_t)}\bigg] \\
=&E\bigg[O^+\frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg] + E\bigg[\sum_{a_t\in\{-1,+1\},\ a_t\neq A_t} O^-\frac{\prod_{t=1}^T \mathbb{I}(a_t f_t(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg]
\end{aligned}$$

where now both O^+ and O^- are non-negative random variables. Hence, it is sufficient to prove Lemma 3.2 for non-negative weight O.

From now on, we assume that $O\geq 0$ and let

$$(f_1^*,...,f_T^*)\in\underset{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T}{\arg\max} E\bigg[O\frac{\min(\psi(A_1f_1(H_1)),...,\psi(A_Tf_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\bigg] \tag{3.18}$$

and

$$(g_1^*,...,g_T^*)\in\underset{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T}{\arg\max} E\bigg[O\frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg]$$

denote arbitrary optimal decision rules of MRL and the original problem for an arbitrary non-negative response variable O. Our goal is to show that $(\text{sign}(f_1^*),...,\text{sign}(T_T^*))$ is one of the choices for $(g_1^*,...,g_T^*)$.

To prove this, by conditioning on H_T the conditional objective function of (3.18) is equal to

$$\begin{aligned}
&E\bigg[\frac{O}{\prod_{t=1}^T p(A_t|H_t)}\min(\psi(A_1f_1(H_1)),...,\psi(A_Tf_T(H_T)))\bigg|H_T\bigg] \\
=&\frac{\min(c,\psi(|f_T(H_T)|))\mathbb{I}(f_T(H_T)>0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}E[O|H_T,A_T=1] \\
&+\frac{\min(c,\psi(|f_T(H_T)|))\mathbb{I}(f_T(H_T)<0)}{\prod_{t=1}^{T-1} p(A_t|H_t)}E[O|H_T,A_T=-1],
\end{aligned} \tag{3.19}$$

where

$$c = \min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-1} f_{T-1}(H_{T-1}))).$$

Note that the weight term $\min(c, \psi(|f_T(H_T)|))$ in (3.19) is non-negative and is maximized for any $f_T(\cdot)$ such that $|f_T(H_T)| \geq 1$ due to the definition of $\psi(\cdot)$. Thus, in the interval where $f_T(H_T) > 0$, the above term has the maximum value $\min(c, \psi(1))E[O|H_T, A_T = 1]/\prod_{t=1}^{T-1} p(A_t|H_t)$ which is attained for any $f_T(H_T) \geq 1$; in the interval where $f_T(H_T) < 0$, the maximum value is $\min(c, \psi(1))E[O|H_T, A_T = -1]/\prod_{t=1}^{T-1} p(A_t|H_t)$ and it is attained for any $f_T(H_T) \leq -1$. Comparing these two values, we conclude that $f_T^*(H_T)$ can be any function taking form $\omega(H_T)\text{sign}(E[O|H_T, A_T = 1] - E[O|H_T, A_T = -1])$ where $\omega(H_T) \geq 1$.On the other hand, according to Zhao et al. (2015) function $\{g_t^*\}_{t=1}^T$ should satisfy

$$\text{sign}(g_t^*(H_t)) = \text{sign}(E[U_{t+1}^*(H_{t+1})|H_t, A_t = 1] - E[U_{t+1}^*(H_{t+1})|H_t, A_t = -1])$$

almost surely, where

$$U_t^*(H_t) = E\left[O\frac{\prod_{s=t}^T \mathbb{I}(A_s g_s^*(H_s) > 0)}{\prod_{s=t}^T p(A_s|H_s)}\middle| H_t\right]$$

with $U_{T+1}^* = O$. Immediately, we have $\text{sign}(g_T^*(H_T)) = \text{sign}(f_T^*(H_T))$, which indicates that f_T^* is a Fisher consistent estimator of g_T^*.

By plugging f_T^* into (3.18), we can verify that $(f_1^*, ..., f_{T-1}^*)$ should maximize

$$\begin{aligned}
&E\left[O\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-1} f_{T-1}(H_{T-1})), \psi(A_T f_T^*(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
=&E\left[E\left[O\frac{\mathbb{I}(A_T f_T^*(H_T) > 0)}{p(A_T|H_T)}\middle| H_T\right]\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-1} f_{T-1}(H_{T-1})), 1)}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right] \\
=&E\left[U_T^*(H_T)\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-1} f_{T-1}(H_{T-1})))}{\prod_{t=1}^{T-1} p(A_t|H_t)}\right]
\end{aligned}$$

and the last inequality holds since $\psi(\cdot) \leq 1$. Repeating the same argument for $T-1$, we can also show that

$$\begin{aligned}
&\text{sign}(f_{T-1}^*(H_{T-1})) \\
=&\text{sign}(E[U_T^*(H_T)|H_{T-1}, A_{T-1} = 1] - E[U_T^*(H_T)|H_{T-1}, A_{T-1} = -1]) \\
=&\text{sign}(g_{T-1}^*(H_{T-1}))
\end{aligned}$$

and $(f_1^*, ..., f_{T-2}^*)$ should maximize

$$
\begin{aligned}
&E\left[O\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-1} f_{T-1}^*(H_{T-1})), \psi(A_T f_T^*(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
=&E\left[E\left[O\frac{\mathbb{I}(A_T f_T^*(H_T) > 0)\mathbb{I}(A_{T-1} f_{T-1}^*(H_{T-1}) > 0)}{p(A_T|H_T)p(A_{T-1}|H_{T-1})}\middle| H_{T-1}\right]\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-2} f_{T-2}(H_{T-2})))}{\prod_{t=1}^{T-2} p(A_t|H_t)}\right] \\
=&E\left[U_{T-1}^*(H_{T-1})\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-2} f_{T-2}(H_{T-2})))}{\prod_{t=1}^{T-2} p(A_t|H_t)}\right].
\end{aligned}
\tag{3.20}
$$

The first equality in (3.20) holds since previous arguments indicate that $|f_T^*(H_T)| \geq 1$ and $|f_{T-1}^*(H_{T-1})| \geq 1$ almost surely, which implies that

$$
\min(\psi(A_{T-1} f_{T-1}^*(H_{T-1})), \psi(A_T f_T^*(H_T))) = \begin{cases} 1, & \text{if } A_{T-1} f_{T-1}^*(H_{T-1}) > 0 \ \& \ A_T f_T^*(H_T) > 0 \\ 0, & \text{if } A_{T-1} f_{T-1}^*(H_{T-1}) < 0 \text{ or } A_T f_T^*(H_T) < 0 \end{cases},
$$

and consequently,

$$
\begin{aligned}
&\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-2} f_{T-2}(H_{T-2})), \psi(A_{T-1} f_{T-1}^*(H_{T-1})), \psi(A_T f_T^*(H_T))) \\
=&\mathbb{I}(A_{T-1} f_{T-1}^*(H_{T-1}) > 0)\mathbb{I}(A_T f_T^*(H_T) > 0)\min(\psi(A_1 f_1(H_1)), ..., \psi(A_{T-2} f_{T-2}(H_{T-2}))),
\end{aligned}
$$

The proof is completed by repeating the argument from $T-2$ to 1.

3.9 Proof of Theorem 3.1 and Theorem 3.2

We aim to prove Theorem 3.1 and Theorem 3.2 in this section. The structure is organized as follows: in section Section 3.9.1, we prove a general preliminary lemma which will be used to derive the non-asymptotic rates in Theorem 3.1 and Theorem 3.2. We then complete the proof of Theorem 3.1 and state the propositions and lemmas used in the proof in Section 3.9.2. The proof of Theorem 3.2 is given in Section 3.9.3 along with additional lemmas used to complete the proof. The proof relies on the covering number property Proposition 2.1 stated in Section 2.9.

3.9.1 Proof of a general lemma

We start with proving a general lemma for establishing the non-asymptotic convergence in Theorem 3.1 and Theorem 3.2. Recall that we use $\{(g^*_{1,\gamma}, ..., g^*_{T,\gamma})\}_{\gamma\in[0,1]}$ to denote the true optimal decision functions of unconstrained DTRs problem with response variable $O_\gamma = (1-\gamma)Y - \gamma R$. Lemma 3.1 indicates that the true optimal decisions functions of a CBR problem under risk constraint τ are given by

$$(g^*_1, ..., g^*_T) = (g^*_{1,\gamma^*}, ..., g^*_{T,\gamma^*}).$$

Let $\{(\widehat{f}_{1,\gamma}, ..., \widehat{f}_{T,\gamma})\}_{\gamma\in[0,1]}$ be arbitrary estimators of $\{(g^*_{1,\gamma}, ..., g^*_{T,\gamma})\}_{\gamma\in[0,1]}$ and let $\widehat{\gamma}$ denote an arbitrary estimator of true multiplier γ^*. Given constants $0 < a_n$, $0 < b_n$ and $\zeta \in (0,1)$, for arbitrary $\delta_1 > 0$ and $\delta_2 > 0$ we consider following three conditions:

(C1) $\sup_{\gamma\in[0,1]} \left| E\left[\{(1-\gamma)Y - \gamma R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\gamma}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - E\left[\{(1-\gamma)Y - \gamma R\}\frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\gamma}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]\right| \le a_n + \delta_1$

(C2) $\left| E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - \tau\right| \le b_n + \delta_2$

(C3) $\widehat{\gamma} < 1 - \zeta$.

The main result of this section is stated as Lemma 3.3 below:

Lemma 3.3 *Suppose that there exist $0 < a_n$, $0 < b_n$ and $\zeta \in (0,1)$ such that conditions (C1)-(C3) hold with probability at least $1 - ce^{-c' \min(\delta_1^2, \delta_2^2)n}$ for any sufficient small $\delta_1 > 0$ and $\delta_2 > 0$. Then, we also have*

$$E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] \ge E\left[Y\frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - \frac{1}{\zeta}(a_n + b_n + \delta_1 + \delta_2) \qquad (3.21)$$

holds with probability at least $1 - ce^{-c' \min(\delta_1^2, \delta_2^2)n}$. Here, c and c' are two arbitrary positive constants that do not depend on sample size n.

Remark: Note that condition (C2) characterizes the expected adverse risk under the estimated decision functions against risk constraint τ, which provides an upper bound of the expected risk. Lemma 3.3 indicates that if in addition the estimation of each unconstrained DTRs problem w.r.t. response variable O_γ can be uniformly good for any $\gamma \in [0,1]$ and the estimated multiplier $\widehat{\gamma}$ is guaranteed to be bounded away from 1, i.e. conditions (C1) and (C3) also hold, then the characterization of the expected adverse risk will also ensure

that the beneficial reward under this estimated rules will not be significantly lower than the optimal expected reward. In the proof of Theorem 3.1 and Theorem 3.2, we will verify that conditions (C2) and (C3) can be obtained from condition (C1) under Assumption 3.6 and Assumption 3.7. Hence, the establishment of the non-asymptotic convergence rates consists of first establishing the uniform concentration inequality (C1) and then verifying conditions (C2) and (C3).

Proof: Condition (C1) implies that

$$
\begin{aligned}
&E\bigg[\{(1-\widehat{\gamma})Y-\widehat{\gamma}R\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]\\
&\geq E\bigg[\{(1-\widehat{\gamma})Y-\widehat{\gamma}R\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-a_n-\delta_1
\end{aligned}
$$

with probability at least $1-ce^{-c'\min(\delta_1^2,\delta_2^2)n}$. Moreover, by the definition of $(g^*_{1,\gamma},...,g^*_{T,\gamma})$ we also have

$$
E\bigg[\{(1-\widehat{\gamma})Y-\widehat{\gamma}R\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]\geq E\bigg[\{(1-\widehat{\gamma})Y-\widehat{\gamma}R\}\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg].
$$

Hence, the two inequalities above imply that

$$
\begin{aligned}
&(1-\widehat{\gamma})E\bigg[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-(1-\widehat{\gamma})E\bigg[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]\\
\geq&\widehat{\gamma}E\bigg[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-\widehat{\gamma}E\bigg[\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-a_n-\delta_1\\
=&\widehat{\gamma}\bigg(E\bigg[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-\tau\bigg)-a_n-\delta_1.
\end{aligned}
\tag{3.22}
$$

Note that condition (C2) indicates that (3.22) is further lower bounded by $-(a_n+b_n+\delta_1+\delta_2)$ with probability at least $1-ce^{-c'\min(\delta_1^2,\delta_2^2)n}$. Since $\widehat{\gamma}<1-\zeta$, by dividing $1-\widehat{\gamma}$ from both side of (3.22) we have

$$
E\bigg[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]\geq E\bigg[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t g^*_{t}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\bigg]-\frac{1}{\zeta}(a_n+b_n+\delta_1+\delta_2),
$$

hold with probability at least $1-ce^{-c'\min(\delta_1^2,\delta_2^2)n}$, which completes the proof of lemma. □

3.9.2 Proof of Theorem 3.1

We complete the proof of Theorem 3.1 in this section. Throughout Section 3.9.2, we let

$$\epsilon_n = \sum_{t=1}^{T} c_1^{-(1-t)} C_{1,t} \left(\frac{1}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} + \lambda_{n,t} \sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \right),$$

$$\xi_n = T c_1^{-2T} \sum_{t=1}^{T} C_{2,t} \frac{1}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2} \lambda_{n,t}^{-\nu_t'/4} \epsilon_n^{-\nu_t'/2}.$$

The proof is completed by verifying (C1) to (C3).

Step 1 - verify condition (C1): Let

$$\mathcal{V}_t(f_t, ..., f_T; \gamma) = E\left[\{(1-\gamma)Y - \gamma R\} \frac{\prod_{s=t}^{T} \mathbb{I}(A_s f_s(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \right]$$

and

$$\mathcal{V}_{\phi,t}(f_t, ..., f_T; \gamma) = -E\left[\{(1-\gamma)Y - \gamma R\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s f_s(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \phi(A_t f_t(H_t)) \right].$$

Then following the same proof of Theorem 3.4 in Zhao et al. (2012) and Theorem 3.4 in Zhao et al. (2015), under Assumption 3.5 it can be shown that both

$$\begin{aligned} 0 \leq & \mathcal{V}_t(g_{t,\gamma}^*, ..., g_{T,\gamma}^*; \gamma) - \mathcal{V}_t(\widehat{f}_{t,\gamma}, ..., \widehat{f}_{T,\gamma}; \gamma) \\ \leq & \sum_{s=t}^{T} c_1^{-(t-s)} C_{1,t} \left(\frac{1}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} + \lambda_{n,t} \sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \right) + \delta \\ \leq & \epsilon_n + \delta \end{aligned} \tag{3.23}$$

$$\begin{aligned} 0 \leq & \mathcal{V}_{\phi,t}(g_{t,\gamma}^*, ..., g_{T,\gamma}^*; \gamma) - \mathcal{V}_{\phi,t}(\widehat{f}_{t,\gamma}, ..., \widehat{f}_{T,\gamma}; \gamma) \\ \leq & \sum_{s=t}^{T} c_1^{-(t-s)} C_{1,t} \left(\frac{1}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} + \lambda_{n,t} \sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \right) + \delta \\ \leq & \epsilon_n + \delta \end{aligned} \tag{3.24}$$

holds with probability for at least $1 - \sum_{s=t}^{T} e^{-c_t' \delta^2 n}$ for any $\delta > 0$, $\gamma \in [0, 1]$ and $t \in \{1, ..., T\}$, where c_t' is a positive constant which depend on $(\nu_t, \theta_t, d_t, M, c_1)$ and are independent of n for t=1,...,T.

Remark: Liu et al. (2018) shows that using AOWL to estimate an unconstrained optimal DTRs problem

will lead to the same non-asymptotic errors as OWL obtained in (3.23) and (3.24). Hence, the proof of OWL and AOWL can follow exactly the same proof scheme which leads to the same error bounds as stated in Theorem 3.1.

In particular, when $t = 1$ we can obtain that for any $\delta > 0$, $\gamma \in [0, 1]$ and $t = 1, ..., T$,

$$\begin{aligned}
0 \leq & \mathcal{V}_1(g^*_{1,\gamma}, ..., g^*_{T,\gamma}; \gamma) - \mathcal{V}_1(\widehat{f}_{1,\gamma}, ..., \widehat{f}_{T,\gamma}; \gamma) \\
= & \left| E\left[\{(1-\gamma)Y - \gamma R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\gamma}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - E\left[\{(1-\gamma)Y - \gamma R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\gamma}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] \right| \\
\leq & \epsilon_n + \delta
\end{aligned}$$

holds with probability for at least $1 - \sum_{t=1}^T e^{-c_t' \delta^2 n}$, which indicates that condition (C1) holds by choosing $a_n = \epsilon_n$ and $\delta_1 = \delta$.

Step 2 - verify condition (C2): To verify (C2), following the proof of Theorem 3.4 in Zhao et al. (2012) one can also show that $\widehat{f}_{t,\widehat{\gamma}} \in \mathcal{B}_{\mathcal{G}_t}(M c_1^{-1} \lambda_{n,t}^{-1/2})$ and Lemma 3.4 given at the end of this section implies that

$$\left| E\left[R \frac{\min(\psi(A_1 \widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta), ..., \psi(A_T \widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right] - \tau \right| \leq \delta + \xi_n \tag{3.25}$$

holds with probability at least $1 - e^{-\frac{1}{2} c_1^{2T} M^{-2} \delta^2 n}$ where ξ_n is the term defined in (3.41) which depends on the shifting parameter η.

One the other hand, let

$$\widehat{g}_{t,\widehat{\gamma}}(H_t) = \begin{cases} \widehat{f}_{t,\widehat{\gamma}}(H_t), & \text{if } |\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq 1 \\ \text{sign}(\widehat{f}_{t,\widehat{\gamma}}(H_t)), & \text{if } |\widehat{f}_{t,\widehat{\gamma}}(H_t)| > 1, \end{cases}$$

which is equal to $\widehat{f}_{t,\widehat{\gamma}}$ truncated at ± 1 for $t = 1, ..., T$. Using the definition of the hinge loss function and, without loss of generality, assuming that $(1 - \widehat{\gamma})Y - \widehat{\gamma} R$ are non-negative, which can always be achieved via simply adding constant M to the response variable or using the augmentation, we can obtain that the surrogate regret under $\gamma = \widehat{\gamma}$ up to stage t satisfies

$$
\begin{aligned}
&\mathcal{V}_{\phi,t}(g^*_{t,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}};\widehat{\gamma}) - \mathcal{V}_{\phi,t}(\widehat{f}_{t,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}};\widehat{\gamma}) \\
\geq &\mathcal{V}_{\phi,t}(g^*_{t,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}};\widehat{\gamma}) - \mathcal{V}_{\phi,t}(\widehat{g}_{t,\widehat{\gamma}},\widehat{f}_{t+1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}};\widehat{\gamma}) \\
= &- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t g^*_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
&- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s \widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right].
\end{aligned} \tag{3.26}
$$

Since $(g^*_{1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})$ are the optimal decision functions associated with response variable $(1-\widehat{\gamma})Y - \widehat{\gamma}R$, the first two terms in the last expression of (3.26) is lower bounded by

$$
\begin{aligned}
&- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t g^*_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
\geq &- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\mathbb{I}(|\widehat{g}_{t,\widehat{\gamma}}(H_t)|\leq\epsilon_n)\phi(A_t g^*_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\mathbb{I}(|\widehat{g}_{t,\widehat{\gamma}}(H_t)|\leq\epsilon_n)\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
= &- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)|\leq\epsilon_n)\phi(A_t g^*_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)|\leq\epsilon_n)\phi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t))\right].
\end{aligned} \tag{3.27}
$$

Moreover, using conclusion (3.23) with $t+1$ the third and fourth terms of (3.26) is lower bounded by

$$
\begin{aligned}
&- E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
&+ E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\}\frac{\prod_{s=t+1}^{T}\mathbb{I}(A_s \widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=t}^{T}p(A_s|H_s)}\phi(A_t \widehat{g}_{t,\widehat{\gamma}}(H_t))\right] \\
\geq &- (\epsilon_n + \delta)
\end{aligned} \tag{3.28}
$$

with probability at least $1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n}$. Hence, by combining (3.27) and (3.28) we can obtain that

$$\begin{aligned}
&\mathcal{V}_{\phi,t}(g^*_{t,\widehat{\gamma}}, ..., g^*_{T,\widehat{\gamma}}; \widehat{\gamma}) - \mathcal{V}_{\phi,t}(\widehat{f}_{t,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}}; \widehat{\gamma}) \\
\geq & - E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n) \phi(A_t g^*_{t,\widehat{\gamma}}(H_t))\right] \\
& + E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n) \phi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t))\right] \\
& - (\epsilon_n + \delta)
\end{aligned} \tag{3.29}$$

holds with probability at least $1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n}$. On the other hand, note that

$$\begin{aligned}
& E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s g^*_{s,\widehat{\gamma}}(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \phi(A_t g^*_{t,\widehat{\gamma}}(H_t)) \middle| H_t = h_t\right] \\
= & -2\min(Q_{t,\widehat{\gamma}}(h_t, 1), Q_{t,\widehat{\gamma}}(h_t, -1))
\end{aligned}$$

and

$$\begin{aligned}
& E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma}R\} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_s g^*_{t,\widehat{\gamma}}(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)} \mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n) \phi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t)) \middle| H_t = h_t\right] \\
\leq & \max\{(1+\epsilon_n) Q_{t,\widehat{\gamma}}(h_t, 1) + (1-\epsilon_n) Q_{t,\widehat{\gamma}}(h_t, -1), (1-\epsilon_n) Q_{t,\widehat{\gamma}}(h_t, 1) + (1+\epsilon_n) Q_{t,\widehat{\gamma}}(h_t, -1)\},
\end{aligned}$$

we can show that (3.29) is lower bounded by

$$\begin{aligned}
& \mathcal{V}_{\phi,t}(g^*_{t,\widehat{\gamma}}, ..., g^*_{T,\widehat{\gamma}}; \widehat{\gamma}) - \mathcal{V}_{\phi,t}(\widehat{f}_{t,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}}; \widehat{\gamma}) \\
& \quad \geq (1-\epsilon_n) E[|Q_{t,\widehat{\gamma}}(H_t, 1) - Q_{t,\widehat{\gamma}}(H_t, -1)| \mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n)] - (\epsilon_n + \delta).
\end{aligned} \tag{3.30}$$

Assumption 3.6 implies that we either have

$$P_H(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n) \leq 3K_1(\epsilon_n + \delta)$$

or

$$E[|Q_{t,\widehat{\gamma}}(H_t, 1) - Q_{t,\widehat{\gamma}}(H_t, -1)| \mathbb{I}(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n)] \geq 3(\epsilon_n + \delta). \tag{3.31}$$

When the second case happens and assume that ϵ_n is sufficient small, (3.30) will then imply that

$$\mathcal{V}_{\phi,t}(g^*_{t,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}};\widehat{\gamma}) - \mathcal{V}_{\phi,t}(\widehat{f}_{t,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}};\widehat{\gamma}) > \epsilon_n + \delta.$$

Since the inequality above will only hold with probability no more than $\sum_{t=1}^{T} e^{-c'_t\delta^2 n}$ according to (3.24), therefore we must have

$$P_H(|\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n) \leq 3K_1(\epsilon_n + \delta) \tag{3.32}$$

holds with probability no more than $\sum_{t=1}^{T} e^{-c'_t\delta^2 n}$ for $t = 1, ..., T$.

Consequently, let

$$\mathcal{D} = \{(H_1, ..., H_T) : \exists t \text{ such that } |\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n\},$$

then by taking union of (3.32) for $t = 1$ to T, with probability at least $1 - \sum_{t=1}^{T} e^{-c'_t\delta^2 n}$ we have

$$P(\mathcal{D}) \leq 3K_1T(\epsilon_n + \delta).$$

As a result, when $\eta \leq \epsilon_n$ we have

$$\begin{aligned}
&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \\
=&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\mathbb{I}((H_1,...,H_T) \in \mathcal{D})\right] \\
&+ E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\mathbb{I}((H_1,...,H_T) \in \mathcal{D}^c)\right] \\
\geq&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T) \in \mathcal{D}^c)\right] \\
=&E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\mathbb{I}((H_1,...,H_T) \in \mathcal{D}^c)\right] \\
\geq&E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \\
&- E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\mathbb{I}((H_1,...,H_T) \in \mathcal{D})\right] \\
\geq&E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)}\right] - 3K_1TMc_1^{-T}(\epsilon_n + \delta).
\end{aligned}$$

In addition, by definition we also have

$$E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right] \leq E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]. \tag{3.33}$$

Hence, combine the two inequalities above we can obtain that

$$\left|E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right]\right| \leq C_3 T c_1^{-T}(\epsilon_n+\delta)$$

holds with probability at least $1-\sum_{t=1}^T e^{-c_t'\delta^2 n}$. Along with (3.25), we have

$$\left|E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - \tau\right| \leq C_3 T c_1^{-T}(\epsilon_n+\delta)+\xi_n$$

holds with probability at least $1-\sum_{t=1}^T e^{-c_t'\delta^2 n} - e^{-\frac{1}{2}c_1^T M^{-2}\delta^2 n}$. Therefore, condition (C2) is satisfied by choosing $\eta=\epsilon_n$ and $b_n = C_3 T c_1^{-T}\epsilon_n + \xi_n$ which will hold with probability at least $1-\sum_{t=1}^T e^{-c_t'\delta^2 n} - e^{-\frac{1}{2}c_1^T M^{-2}\delta^2 n}$ for $\delta_2 = C_3 T c_1^{-T}\delta$.

Step 3 - verify condition (C3): We will verify that $\widehat{\gamma}$ is bounded away from 1 with high probability in this step. Recall that

$$\mathfrak{R}(\gamma) = E\left[R\frac{\prod_{t=1}^T \mathbb{I}(A_t g_{t,\gamma}^*(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right],$$

and in the proof of Lemma 3.1, we have shown that $\mathfrak{R}(\gamma)$ is a non-increasing function of γ.

Since we have also assumed $\mathfrak{R}(\gamma)$ is continuous w.r.t. γ and $\tau > \mathfrak{R}(1)$ in Assumption 3.4, one can always find $\tau > \tau_0 > \mathfrak{R}(1)$ and $\zeta > 0$ such that

$$\mathfrak{R}(1-\zeta) \leq \tau_0 < \tau$$

and

$$\zeta \leq \min\left\{\frac{\tau-\tau_0}{6M}, \frac{1}{3}\right\}.$$

Without loss of generality, we assume that ζ is the largest positive constant that satisfies the two requirements above.

We complete the proof by verifying that

$$P(\widehat{\gamma} > 1-\zeta) \leq 1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n} - e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n}$$

holds for any $0 < \delta$ and n such that $x_n \leq \frac{\tau - \tau_0}{6}$ where we define $x_n = C_3 T c_1^{-T}(\epsilon_n + \delta) + \xi_n$. We now show that $\widehat{\gamma} > 1-\zeta$ implies

$$\left| E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] - E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] \right| \geq \epsilon_n + \delta. \quad (3.34)$$

We verify this by contradiction. If not, then it follows that

$$\begin{aligned} 0 \geq & E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] \\ & - E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] \geq -\epsilon_n - \delta. \end{aligned} \quad (3.35)$$

By rearranging (3.35), we can obtain that

$$\begin{aligned} & - \widehat{\gamma} \left(E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] - E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] \right) \\ \geq & -\epsilon_n - \delta - 2M(1-\widehat{\gamma}) \\ \geq & -\epsilon_n - \delta - 2M\zeta \\ \geq & -\epsilon_n - \delta - \frac{\tau - \tau_0}{3} \\ > & -\frac{1}{2}(\tau - \tau_0) \end{aligned} \quad (3.36)$$

where the last inequality since $\epsilon_n + \delta \leq x_n$ assuming that $K_1 \geq 1$ without loss of generality and $x_n \leq \frac{\tau - \tau_0}{6}$ by requirement. On the other hand, since we have also assumed that $\widehat{\gamma} > 1 - \zeta$, the non-increasing property of $\mathfrak{R}(\gamma)$ obtained in Lemma 3.1 implies that

$$E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)} \right] \leq \tau_0.$$

Moreover, in *step 2* we have shown that

$$\left| E\left[\frac{\prod_{t=1}^{T} \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] - \tau \right| \leq x_n.$$

holds with probability at least $1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n} - e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n}$. Therefore, the left-hand side of (3.36) satisfies

$$\begin{aligned} &- \widehat{\gamma} \bigg(E\left[R \frac{\prod_{t=1}^{T} \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] - E\left[R \frac{\prod_{t=1}^{T} \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] \bigg) \\ \leq & \widehat{\gamma} x_n - \widehat{\gamma}(\tau - \tau_0) \\ \leq & x_n - \frac{2}{3}(\tau - \tau_0) \end{aligned} \tag{3.37}$$

with probability at least $1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n} - e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n}$. Again, since $x_n \leq \frac{\tau - \tau_0}{6}$, (3.36) and (3.37) imply that

$$-\frac{1}{2}(\tau - \tau_0) < -\frac{2}{3}(\tau - \tau_0) + \frac{1}{6}(\tau - \tau_0) \leq -\frac{1}{2}(\tau - \tau_0)$$

which is a contradiction. This indicates that $\widehat{\gamma} > 1 - \zeta$ implies (3.35) holds with probability at least $1 - \sum_{t=1}^{T} e^{-c_t' \delta^2 n} - e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n}$. Consequently, according to condition (C1) verified in *step 1* we have

$$\begin{aligned} & P(\widehat{\gamma} > 1 - \zeta) \\ \leq & P\bigg(\bigg| E\bigg[\{(1 - \widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \bigg] \\ & - E\bigg[\{(1 - \widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \bigg] \bigg| \geq \epsilon_n + \delta \bigg) + O\Big(\sum_{t=1}^{T} e^{-c_t' \delta^2 n} + e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n} \Big) \\ \leq & O\Big(\sum_{t=1}^{T} e^{-c_t' \delta^2 n} + e^{-\frac{1}{2} c_1^T M^{-2} \delta^2 n} \Big). \end{aligned} \tag{3.38}$$

Indeed, the proof can be completed by conditioning on two events

$$\left| E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] - E\left[\{(1-\widehat{\gamma})Y - \widehat{\gamma} R\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t g^*_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] \right| \leq \epsilon_n + \delta$$

and

$$\left| E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-\tau\right| \leq C_3Tc_1^{-T}(\epsilon_n+\delta)+\xi_n,$$

which will improve the inequality (3.38) to

$$P(\widehat{\gamma}>1-\zeta)\leq 1-\sum_{t=1}^{T}e^{-c_t'\delta^2 n}-e^{-\frac{1}{2}c_1^TM^{-2}\delta^2 n}.$$

Complete the proof of Theorem 3.1: step1 - 3 indicate that both condition (C1), (C2) and (C3) hold with probability at least $1-\sum_{t=1}^{T}e^{-c_t'\delta^2 n}-e^{-\frac{1}{2}c_1^TM^{-2}\delta^2 n}$ with $a_n=\epsilon_n$, $b_n=C_3Tc_1^{-T}\epsilon_n+\xi_n$, $\delta_1=\delta$ and $\delta_2=C_3Tc_1^{-T}\delta$ for sufficient large n and sufficient small δ. Thus, Theorem 3.1 will be proved by directly applying the conclusion of Lemma 3.3 up to a constant $C_3Tc_1^{-T}$ for δ. '

We now state and prove Lemma 3.4 used in the proof of Theorem 3.1. The goal of Lemma 3.4 is to establish a concentration inequality for the smooth estimator on the left-hand side of (3.39). Recall that by the choice of $\widehat{\gamma}$ we have

$$\mathbb{P}_n\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]=\tau. \tag{3.39}$$

The additional Proposition 3.1 used to complete the proof of Lemma 3.4 is stated by the end of the proof. To state Lemma 3.4, let $\mathcal{G}$ denote a Gaussian RKHS defined on $\mathbb{R}^d$, then we use $\mathcal{B}_{\mathcal{G}}(r)$ to denote the ball of radius r centered at 0 for $\mathcal{G}$ w.r.t. the norm induced by the inner product equipped by $\mathcal{G}$.

Lemma 3.4 *Suppose that the estimated decision functions* $(\widehat{f}_{1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}})$ *satisfy*

$$\widehat{f}_{t,\widehat{\gamma}}\in\mathcal{B}_{\mathcal{G}_t}(\lambda_{n,t}^{-1/2}),\quad t=1,...,T$$

with probability 1, then for any $0<\nu_t'\leq 2$, $0<\theta_t'$ *for* $t=1,...,T$, *we have*

$$P\left(\left|E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-\tau\right|\geq\delta+\xi_n\right)\leq e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}, \tag{3.40}$$

where

$$\xi_n=Tc_1^{-2T}\big(\sum_{t=1}^{T}C_{2,t}\sigma_{n,t}^{(1-\nu_t/2')(1+\theta_t')d_t/2}\lambda_{n,t}^{-\nu_t'/4}\eta^{-\nu_t'/2}/\sqrt{n}\big). \tag{3.41}$$

Here, $C_{2,t}$ is a positive constants which depend on parameters (ν_t', θ_t', d_t) and M but does not depend on sample size n for $t = 1, ..., T$.

Proof: Since $(\widehat{f}_{1,\widehat{\gamma}}, ..., \widehat{f}_{T,\widehat{\gamma}})$ satisfy (3.39), it suffices to derive a uniform bound for

$$\sup_{f \in \mathcal{W}} |\mathbb{P}_n[f] - E[f]|,$$

where

$$\mathcal{W} = \left\{ R\frac{\min(\psi(A_1 f_1(H_1)/\eta), ..., \psi(A_T f_T(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)} : f_1 \in \mathcal{B}_{\mathcal{G}_1}(\lambda_{n,1}^{-1/2}), ..., f_T \in \mathcal{B}_{\mathcal{G}_T}(\lambda_{n,T}^{-1/2}) \right\}.$$

To establish the uniform bound, let f_1 and f_2 be two functions from set $\mathcal{W}$ associated with stagewise decision functions $(f_{11}, ..., f_{1T})$ and $(f_{21}, ..., f_{2T})$, then the square of the empirical L_2 norm, which we give the definition by the end of this section and denote it as $L_2(\mathbb{P}_n)$, between f_1 and f_2 satisfies

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_1(x_i) - f_2(x_i))^2 \\
\overset{(i)}{\leq} &\frac{1}{n}\sum_{i=1}^{n} M^2 c_1^{-2T}(\sum_{t=1}^{T} |\psi(A_{it} f_{1t}(H_{it})/\eta) - \psi(A_{it} f_{2t}(H_{it})/\eta)|)^2 \\
\overset{(ii)}{\leq} &\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} T M^2 c_1^{-2T}[\psi(A_{it} f_{1t}(H_{it})/\eta) - \psi(A_{it} f_{2t}(H_{it})/\eta)]^2 \\
\overset{(iii)}{\leq} &\sum_{t=1}^{T} \frac{T M^2 c_1^{-2T}\eta^{-2}}{n}\sum_{i=1}^{n}(f_{1t}(H_{it}) - f_{2t}(H_{it}))^2
\end{aligned} \tag{3.42}$$

Here, inequality (i) is guaranteed by Proposition 3.1 stated by the end of the proof and note that $|R/\prod_{t=1}^{T} p(A_t|H_t)| \leq M c_1^{-T}$, (ii) is followed by the Cauchy-Schwarz inequality and the last inequality (iii) holds by noting that $\psi(\cdot/\eta)$ is a η^{-1}-Lipschitz function. Inequality (3.42) implies that

$$\mathcal{N}(\epsilon; \mathcal{W}, L_2(\mathbb{P}_n)) \leq \prod_{t=1}^{T} \mathcal{N}(M^{-1}T^{-1}c_1^T\eta\epsilon; \mathcal{B}_{\mathcal{G}_t}(\lambda_{n,t}^{-1/2}), L_2(\mathbb{P}_n)). \tag{3.43}$$

By Theorem 4.10 from Wainwright (2019), for any $\delta > 0$ we have

$$P(\sup_{f \in \mathcal{W}} |E[f] - \mathbb{P}_n[f]| \geq \delta + 2\text{Rad}_n(\mathcal{W})) \leq e^{-\frac{c_1^{2T}}{2M^2}n\delta^2}, \tag{3.44}$$

where $\text{Rad}_n(\mathcal{F})$ denotes the Rademacher complexity of $\mathcal{F}$ defined as

$$\text{Rad}_n(\mathcal{F}) = E_X E_\epsilon \left| \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|, \quad \epsilon_i \sim i.i.d.\ P(\epsilon_i = \pm 1) = 0.5.$$

Following Example 5.24 from Wainwright (2019), the covering number inequalities (3.43) and the covering number bound from Proposition 2.1 stated in Section 2.9, we can obtain

$$\begin{aligned}
\text{Rad}_n(\mathcal{W}) &\leq \frac{24}{\sqrt{n}} E\left[\int_0^{2Mc_1^{-T}} \sqrt{\log \mathcal{N}(M^{-1}T^{-1}c_1^T \eta\epsilon; \mathcal{W}, L_2(\mathbb{P}_n))} d\epsilon \right] \\
&\leq \frac{24}{\sqrt{n}} E\left[\int_0^{2Mc_1^{-T}} \sqrt{\left(\sum_{t=1}^{T} C_{2,t} \sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t} \lambda_{n,t}^{-\nu_t'/2} (M^{-1}T^{-1}c_1^T \eta\epsilon)^{-\nu_t'} \right)} d\epsilon \right] \\
&\leq \sum_{t=1}^{T} C_{2,t} \frac{c_1^{-2T} T}{\sqrt{n}} \sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2} \lambda_{n,t}^{-\nu_t'/4} \eta^{-\nu_t'/2},
\end{aligned}$$

which completes the proof. □

Proposition 3.1 shows that the difference of the objective function of MRL is bounded by the sum of the difference of each entry.

Proposition 3.1 *For any $z_t \in \mathbb{R}$, $z_t' \in \mathbb{R}$ and $t \in \{1, ..., T\}$, we have*

$$|\min(\psi(z_1), ..., \psi(z_T)) - \min(\psi(z_1'), ..., \psi(z_T'))| \leq \sum_{t=1}^{T} |z_t - z_t'|.$$

Proof: Without loss of generality, we assume that

$$\min(\psi(z_1'), ..., \psi(z_T')) \leq \min(\psi(z_1), ..., \psi(z_T))$$

and suppose that $\min(\psi(z_1'), ..., \psi(z_T')) = z_{t_0}'$, then it is easy to verify that

$$\begin{aligned}
\min(\psi(z_1), ..., \psi(z_T)) - \min(\psi(z_1'), ..., \psi(z_T')) &= \min(\psi(z_1), ..., \psi(z_T)) - \psi(z_{t_0}') \\
&\leq |\psi(z_{t_0}) - \psi(z_{t_0}')| \\
&\leq \sum_{t=1}^{T} |\psi(z_t) - \psi(z_t')| \\
&\leq \sum_{t=1}^{T} |z_t - z_t'|
\end{aligned}$$

where in the last inequality we have used the fact that ψ is a 1-Lipschitz function, which completes the proof. □

3.9.3 Proof of Theorem 3.2

We complete the proof of Theorem 3.2 in this section. For convenience, we define

$$\mathcal{V}(f_1, ..., f_T) = E\bigg[O\frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\bigg]$$

and define the surrogate reward under MRL to be

$$\begin{aligned}\mathcal{V}_\psi(f_1, ..., f_T) =& E\bigg[O\frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\bigg] \\ &+ E\bigg[O^-\frac{\min(\psi(|f_1(H_1)|), ..., \psi(|f_T(H_T)|))}{\prod_{t=1}^T p(A_t|H_t)}\bigg].\end{aligned}$$

Moreover, let

$$L_1(f_1, ..., f_T) = -\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_T))) + 1,$$

$$L_2(f_1, ..., f_T) = -\min(\psi(|f_1(H_1)|), ..., \psi(|f_T(H_T)|)) + 1,$$

and let $O_\gamma = (1-\gamma)Y - \gamma R$. Then by definition, we have

$$(g^*_{1,\gamma}, ..., g^*_{T,\gamma}) = \underset{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T}{\arg\min} E[\mathfrak{L}(f_1, ..., f_T; \gamma)],$$

where

$$\mathfrak{L}(f_1, ..., f_T; \gamma) = O_\gamma\frac{L_1(f_1, ..., f_T; \gamma)}{\prod_{t=1}^T p(A_t|H_t)} + O_\gamma^-\frac{L_2(f_1, ..., f_T; \gamma)}{\prod_{t=1}^T p(A_t|H_t)}.$$

From now on, given finite sample we use $\widehat{f}_\gamma = (\widehat{f}_{1,\gamma}, ..., \widehat{f}_{T,\gamma})$ to denote the solution of

$$\min_{(f_1,...,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \mathbb{P}_n[\mathfrak{L}(f_1, ..., f_T; \gamma)] + \sum_{t=1}^T \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t}.$$

where $\mathcal{G}_t = \mathcal{G}(\sigma_{n,t})$. Throughout Section 3.9.3, we let

$$\epsilon_n = c_1^{-T}\sum_{t=1}^T C_{1,t}\bigg(\frac{1}{\sqrt{n}}\bigg(\sqrt{T} + T^{\nu_t/2}c_1^{-3T\nu_t/4}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4}\bigg) + \lambda_{n,t}\sigma_{n,t}^{d_t} + c_1^{-T}\sigma_{n,t}^{-\alpha_t d_t}\bigg),$$

$$\xi_n = Tc_1^{-2T}\sum_{t=1}^{T} C_{2,t}\left(\frac{1}{\sqrt{n}}c_1^{-T\nu_t'/4}\sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2}\lambda_{n,t}^{-\nu_t'/4}\right),$$

and use c_x to denote a constant that depends on parameter x. The proof of Theorem 3.2 is also completed by checking conditions (C1) to (C3).

Step 1 - verify condition (C1): According to Lemma 3.5 given at the end of the section, we have

$$\mathcal{V}(g_1^*, ..., g_T^*) - \mathcal{V}(f_1, ..., f_T) \leq \mathcal{V}_\psi(g_1^*, ..., g_T^*) - \mathcal{V}_\psi(f_1, ..., f_T)$$

hold for any $(f_1, ..., f_T)$, it is sufficient to prove

$$P\left(\sup_{\gamma\in[0,1]} |E[\mathfrak{L}(\widehat{f}_{1,\gamma}, ..., \widehat{f}_{T,\gamma}; \gamma)] - E[\mathfrak{L}(g_{1,\gamma}^*, ..., g_{T,\gamma}^*; \gamma)]| \geq \delta + \epsilon_n\right) \leq 2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}. \tag{3.45}$$

The key is to construct an adequately good approximation of $g_{t,\gamma}^*$ from $\mathcal{G}_t$ for any $t = 1, .., T$ and $\gamma \in [0, 1]$, which will be denoted as $\widetilde{f}_{t,\gamma}$ from now on.

The construction of such $\widetilde{f}_{t,\gamma}$ basically follows the idea of the proof of Theorem 2.7 in Steinwart and Scovel (2007). Specifically, our goal is to find function $\widetilde{f}_{t,\gamma} \in \mathcal{G}_t$ such that

$$\|\widetilde{f}_{t,\gamma}\|_{\mathcal{G}_t}^2 \leq c_{d_t}\sigma_{n,t}^{d_t} \tag{3.46}$$

and

$$|\widetilde{f}_{t,\gamma}(h_t) - g_{t,\gamma}^*(h_t)| \leq 8e^{-\sigma_{n,t}^2\Delta_{t,\gamma}^2(h_t)/2d_t} \tag{3.47}$$

holds for any $h_t \in \mathcal{H}_t$.

The proof follows the same argument used for proving Theorem 2.7 in Steinwart and Scovel (2007). Using the same argument, we can construct an extension of $g_{t,\gamma}^*$, denoted as $\bar{g}_{t,\gamma}$, such that $\bar{g}_{t,\gamma}$ is well defined on $3\bar{\mathcal{H}}_t$ and $g_{t,\gamma}^*(h_t) = \bar{g}_{t,\gamma}(h_t)$ holds for any $h_t \in \mathcal{H}_t$. The construction of $\widetilde{f}_{t,\gamma}$ will use the fact that the linear operator $V_\sigma : L_2(\mathbb{R}^d) \to \mathcal{G}_\sigma$ defined by

$$V_\sigma g(x) = \frac{(2\sigma)^{d/2}}{\pi^{d/4}}\int_{\mathbb{R}^d} e^{-2\sigma^2\|x-y\|^2} g(y)dy$$

is an isometric isomorphism. Let $\widetilde{g}_{t,\gamma} = (\sigma_{n,t}^2/\pi)^{d_t/4}\bar{g}_{t,\gamma}$, then one can verify that

$$\|\widetilde{g}_{t,\gamma}\|_{L_2}^2 \leq \text{Vol}(d_t)^2\sigma_{n,t}^{d_t} = c_{d_t}\sigma_{n,t}^{d_t},$$

where $\text{Vol}(d_t)$ is the volume of the unit ball in $\mathbb{R}^{d_t}$.

We now show that $\widetilde{f}_{t,\gamma} = V_{\sigma_{n,t}}\widetilde{g}_{t,\gamma}$ satisfies the requirements (3.46) and (3.47). Since $V_{\sigma_{n,t}}$ is an isometric isomorphism, we have

$$\|\widetilde{f}_{t,\gamma}\|_{\mathcal{G}_t}^2 = \|V_{\sigma_{n,t}}\widetilde{g}_{t,\gamma}\|_{\mathcal{G}_t}^2 = \|\widetilde{g}_{t,\gamma}\|_{L_2}^2 \leq c_{d_t}\sigma_{n,t}^{d_t}.$$

On the other hand, following the argument of Lemma 4.1 from Steinwart and Scovel (2007) it can be shown that for any h_t such that $g_{t,\gamma}^*(h_t) = 1$, by construction we will have $\bar{g}_{t,\gamma}(h) = 1$ for any $h \in B(h_t, \Delta_{t,\gamma}(h_t))$. Consequently, we have

$$\begin{aligned}
1 \geq V_{\sigma_{n,t}}g(h_t) &= \left(\frac{2\sigma_{n,t}}{\pi}\right)^{d_t/2}\int_{\mathbb{R}^{d_t}} e^{-2\sigma_{n,t}^2\|h_t-y\|}\bar{g}_{t,\gamma}(y)dy \\
&\geq \left(\frac{2\sigma_{n,t}}{\pi}\right)^{d_t/2}\int_{B(h_t,\Delta_{t,\gamma}(h_t))} e^{-2\sigma_{n,t}^2\|h_t-y\|}(\bar{g}_{t,\gamma}(y)+1)dy - 1 \\
&= 2\left(\frac{2\sigma_{n,t}}{\pi}\right)^{d_t/2}\int_{B(h_t,\Delta_{t,\gamma}(h_t))} e^{-2\sigma_{n,t}^2\|h_t-y\|}dy - 1 \\
&\geq 1 - 2P(|U| \geq \Delta_{t,\gamma}(h_t)),
\end{aligned}$$

where U is a random variable following spherical Gaussian distribution in $\mathbb{R}^{d_t}$. Hence, we can obtain

$$|\widetilde{f}_{t,\gamma}(h_t) - 1| = |\widetilde{f}_{t,\gamma}(h_t) - g_{t,\gamma}^*(h_t)| \leq 1 - 2P(|U| \geq \Delta_{t,\gamma}(h_t))$$

for any h_t such that $g_{t,\gamma}^*(h_t) = 1$. Using inequality (3.5) from Ledoux and Talagrand (1991), the right-hand side of the inequality above is bounded by

$$1 - 2P(|U| \geq \Delta_{t,\gamma}(h_t)) \leq 1 - 8e^{-\sigma_{n,t}^2\Delta_{t,\gamma}^2(h_t)/2d_t} \tag{3.48}$$

By repeating the same argument, we can verify that (3.48) will also hold for any $h_t \in \{h \in \mathcal{H}_t : g_{t,\gamma}^*(h) < 0\}$.

We now show that using (3.48) and Assumption 3.5, we can obtain that $(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})$ satisfies

$$E[\mathfrak{L}(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma};\gamma)]+\sum_{t=1}^{T}\lambda_{n,t}\|\widetilde{f}_{t,\gamma}\|^2_{\mathcal{G}_t}-E[\mathfrak{L}(g^*_{1,\gamma},...,g^*_{T,\gamma};\gamma)]\leq\sum_{t=1}^{T}c_{d_t,K,M}(\lambda_{n,t}\sigma^{d_t}_{n,t}+c_1^{-T}\sigma^{-\alpha_t d_t}_{n,t}). \tag{3.49}$$

To prove (3.49), first note that (3.46) indicates that

$$\sum_{t=1}^{T}\lambda_{n,t}\|\widetilde{f}_{t,\gamma}\|^2_{\mathcal{G}_t}\leq\sum_{t=1}^{T}c_{d_t}\lambda_{n,t}\sigma^{d_t}_{n,t}.$$

On the other hand, Proposition 3.1 implies that

$$\begin{aligned}
&E\left[\frac{O_\gamma L_1(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-E\left[\frac{O_\gamma L_1(g^*_{1,\gamma},...,g^*_{T,\gamma})}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\\
\leq& Mc_1^{-T}E[|L_1(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})-L_1(g^*_{1,\gamma},...,g^*_{T,\gamma})|]\\
\overset{(i)}{\leq}& Mc_1^{-T}\sum_{t=1}^{T}E[|\widetilde{f}_{t,\gamma}(H_t)-g^*_{t,\gamma}(H_t)|]\\
\overset{(ii)}{\leq}& Mc_1^{-T}\sum_{t=1}^{T}E[e^{-\sigma^2_{n,t}\Delta^2_{t,\gamma}(H_t)/2d_t}]\\
\overset{(iii)}{\leq}& C_{M,K}c_1^{-T}\sum_{t=1}^{T}\sigma^{-\alpha_t d_t}_{n,t},
\end{aligned}$$

where to obtain (i) we implement the inequality in Proposition 3.1, (ii) is ensured by (3.47) and (3.48), (iii) follows from applying Assumption 3.5. Similarly, using the fact that $||x_1|-|x_2||\leq|x_1-x_2|$ we can also show that

$$E\left[\frac{O^-_\gamma L_2(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-E\left[\frac{O^-_\gamma L_2(g^*_{1,\gamma},...,g^*_{T,\gamma})}{\prod_{t=1}^{T}p(A_t|H_t)}\right]\leq c_{M,K}c_1^{-T}\sum_{t=1}^{T}\sigma^{-\alpha_t d_t}_{n,t}$$

which completes the verification of (3.49).

We now start verifying condition (C1). Note that for any γ, we have following decomposition

$$\begin{aligned}
0\leq& E[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma};\gamma)]-E[\mathfrak{L}(g^*_{1,\gamma},...,g^*_{T,\gamma};\gamma)]\\
\leq& \underbrace{E[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma};\gamma)]-\mathbb{P}_n[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma};\gamma)]}_{(I)}
\end{aligned}$$

$$
\begin{aligned}
&\underbrace{+\mathbb{P}_n[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma};\gamma)]+\sum_{t=1}^{T}\lambda_{n,t}\|\widehat{f}_{t,\gamma}\|_{\mathcal{G}_t}^2-\mathbb{P}_n[\mathfrak{L}(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma};\gamma)]-\sum_{t=1}^{T}\lambda_{n,t}\|\widetilde{f}_{t,\gamma}\|_{\mathcal{G}_t}^2}_{(II)}\\
&\underbrace{+\mathbb{P}_n[\mathfrak{L}(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})]-E[\mathfrak{L}(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})]}_{(III)}\\
&\underbrace{+E[\mathfrak{L}(\widetilde{f}_{1,\gamma},...,\widetilde{f}_{T,\gamma})]+\sum_{t=1}^{T}\lambda_{n,t}\|\widetilde{f}_{t,\gamma}\|_{\mathcal{G}_t}^2-E[\mathfrak{L}(g^*_{1,\gamma},...,g^*_{T,\gamma};\gamma)]}_{(IV)},
\end{aligned}
$$

By the definition of $\{\widehat{f}_{t,\gamma}\}_{t=1}^T$, we know that $(II)\leq 0$. Moreover, (3.49) indicates that

$$
(IV)\leq\sum_{t=1}^{T}c_{d_t,K,M}(\lambda_{n,t}\sigma_{n,t}^{d_t}+c_1^{-T}\sigma_{n,t}^{-\alpha_t d_t}).
$$

Hence, it remains to establish a non-asymptotic upper bound for (I) and (III) w.r.t. γ. Note that by definition

$$
\begin{aligned}
&\mathbb{P}_n[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma};\gamma)]+\sum_{t=1}^{T}\lambda_{n,t}\|\widehat{f}_{t,\gamma}\|_{\mathcal{G}_t}^2\\
\leq&\mathbb{P}_n[\mathfrak{L}(0,...,0;\gamma)]+0\\
\leq&Mc_1^{-T},
\end{aligned}
$$

and consequently, we have

$$
\sum_{t=1}^{T}\lambda_{n,t}\|\widehat{f}_{t,\gamma}\|_{\mathcal{G}_t}^2\leq-\mathbb{P}_n[\mathfrak{L}(\widehat{f}_{1,\gamma},...,\widehat{f}_{T,\gamma})]+\mathbb{P}_n[\mathfrak{L}(0,...,0;\gamma)]\leq 2Mc_1^{-T}.
$$

The previous inequality implies that

$$
\widehat{f}_{t,\gamma}\in\mathcal{B}_{\mathcal{G}_t}((2M)^{-1/2}c_1^{-T/2}\lambda_{n,t}^{-1/2}),\tag{3.50}
$$

and (3.46) indicates that

$$
\widetilde{f}_{t,\gamma}\in\mathcal{B}_{\mathcal{G}_t}(c_{d_t}\sigma_{n,t}^{d_t/2}).
$$

Since we assumed $\lambda_{n,t}\sigma_{n,t}^{d_t}\to 0$, without loss of generality we can further assume that both $\widehat{f}_{t,\gamma}$ and $\widetilde{f}_{t,\gamma}$ belong to $\mathcal{B}_{\mathcal{G}_t}(c_{d_t,M}c_1^{-T/2}\lambda_{n,t}^{-1/2})$ for some positive constant $c_{d_t,M}$. Therefore, it is sufficient to obtain a

uniform concentration inequality for $|\mathbb{P}_n[\mathfrak{L}(f_1,...,f_T)] - E[\mathfrak{L}(f_1,...,f_T)]|$ for any $\gamma \in [0,1]$ and $f_t \in \mathcal{B}_{\mathcal{G}_t}(c_{d_t,M}c_1^{-T/2}\lambda_{n,t}^{-1/2})$ with $t = 1,...,T$.

To achieve this, recall that

$$\mathfrak{L}(f_1,...,f_T;\gamma) = O_\gamma \frac{L_1(f_1,...,f_T;\gamma)}{\prod_{t=1}^T p(A_t|H_t)} + O_\gamma^- \frac{L_2(f_1,...,f_T;\gamma)}{\prod_{t=1}^T p(A_t|H_t)}.$$

Thus, it suffices to establish the uniform value bounds for both

$$\sup_{f\in\mathcal{W}_1} |E[f] - \mathbb{P}_n[f]|, \quad \sup_{f\in\mathcal{W}_2} |E[f] - \mathbb{P}_n[f]|,$$

where

$$\mathcal{W}_1 = \left\{ \{(1-\gamma)Y - \gamma R\} L_1(f_1,...,f_T) : \gamma \in [0,1], f_t \in \mathcal{B}_{\mathcal{G}_t}(c_{d_t,M}c_1^{-T/2}\lambda_{n,t}^{-1/2}), t = 1,...,T \right\},$$

$$\mathcal{W}_2 = \left\{ \{(1-\gamma)Y - \gamma R\}^- L_2(|f_1|,...,|f_T|) : \gamma \in [0,1], f_t \in \mathcal{B}_{\mathcal{G}_t}(c_{d_t,M}c_1^{-T/2}\lambda_{n,t}^{-1/2}), t = 1,...,T \right\}.$$

We first derive a concentration inequality for $\sup_{f\in\mathcal{W}_1} |E[f] - \mathbb{P}_n[f]|$. Again, by Theorem 4.10 from Wainwright (2019), for any $\delta > 0$ we have

$$P(\sup_{f\in\mathcal{W}_1} |E[f] - \mathbb{P}_n[f]| \geq \delta + 2\text{Rad}_n(\mathcal{W}_1)) \leq e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}, \tag{3.51}$$

and following Example 5.24 from Wainwright (2019) and the upper bound of covering number obtained in Lemma 3.6 given by the end of the section we have

$$\begin{aligned}
&\text{Rad}_n(\mathcal{W}_1) \\
\leq &\frac{24}{\sqrt{n}} E\left[\int_0^{2Mc_1^{-T}} \sqrt{\log \mathcal{N}(\epsilon; \mathcal{W}_1, L_2(\mathbb{P}_n))} d\epsilon \right] \\
\leq &\frac{24}{\sqrt{n}} E\left[\int_0^{2Mc_1^{-T}} \sqrt{c_{M,c_1}\left(-\log\epsilon + T + \sum_{t=1}^T \log \mathcal{N}(c_1^T\epsilon/(2MT); \mathcal{B}_{\mathcal{G}_t}(c_{d_t,M}c_1^{-T/2}\lambda_{n,t}^{-1/2}), L_2(\mathbb{P}_n)) \right)} d\epsilon \right] \\
\overset{(i)}{\leq} &\frac{24}{\sqrt{n}} E\left[\int_0^{2Mc_1^{-T}} \sqrt{c_{M,c_1}\left(-\log\epsilon + T + \sum_{t=1}^T c_1^{-3T\nu_t/2} T^{\nu_t} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t} \lambda_{n,t}^{-\nu_t/2} \epsilon^{-\nu_t} \right)} d\epsilon \right] \\
\leq &\frac{C_{M,c_1}}{\sqrt{n}} c_1^{-T} \left(\sqrt{T} + \sum_{t=1}^T C_{1,t} T^{\nu_t/2} c_1^{-3T\nu_t/4} \sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2} \lambda_{n,t}^{-\nu_t/4} \right),
\end{aligned} \tag{3.52}$$

where to obtain (i) we again use the covering number property Proposition 2.1. The last inequality in (3.52) implies

$$P\bigg(\sup_{f\in\mathcal{W}_1}|E[f]-\mathbb{P}_n[f]|\geq\delta+\frac{C_{M,c_1}}{\sqrt{n}}c_1^{-T}\bigg(\sqrt{T}+\sum_{t=1}^{T}C_{1,t}T^{\nu_t/2}c_1^{-3T\nu_t/4}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4}\bigg)\bigg)$$
$$\leq e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}.$$

Repeating analogous argument for $\mathcal{W}_2$, we can also show that

$$P\bigg(\sup_{f\in\mathcal{W}_2}|E[f]-\mathbb{P}_n[f]|\geq\delta+\frac{C_{M,c_1}}{\sqrt{n}}c_1^{-T}\bigg(\sqrt{T}+\sum_{t=1}^{T}C_{1,t}T^{\nu_t/2}c_1^{-3T\nu_t/4}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4}\bigg)\bigg)$$
$$\leq e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}.$$

Combine two concentration inequalities above, we can obtain that

$$P\bigg(\sup|\mathbb{P}_n[\mathfrak{L}(f_1,...,f_T;\gamma)]-E[\mathfrak{L}(f_1,...,f_T;\gamma)]|$$
$$\geq\delta+\frac{C_{M,c_1}}{\sqrt{n}}c_1^{-T}\bigg(\sqrt{T}+\sum_{t=1}^{T}C_{1,t}T^{\nu_t/2}c_1^{-3T\nu_t/4}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4}\bigg)\bigg)\leq 2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n} \tag{3.53}$$

where the supreme in (3.53) is taken w.r.t.

$$\gamma\in[0,1],\quad (f_1,...,f_T)\in\mathcal{B}_{\mathcal{G}_1}(c_{d_1,M}c_1^{-T/2}\lambda_{n,1}^{-1/2})\times\cdots\times\mathcal{B}_{\mathcal{G}_T}(c_{d_T,M}c_1^{-T/2}\lambda_{n,T}^{-1/2}).$$

Apply (3.53) to (I) and (III), we can show that both two terms are bounded by the left-hand side of (3.53) with probability at least $1-2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. The verification is completed by combing $(I)-(IV)$.

Step 2 - verify condition (C2): We complete verifying (C2) by first proving a preliminary inequality. Let

$$\Gamma(\epsilon)=\sum_{t=1}^{T}P_H\bigg(\bigg\{H_t\in\mathcal{H}_t:\prod_{s=1}^{t}\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)=1,|\widehat{f}_{t,\widehat{\gamma}}(H_t)|\leq\epsilon\bigg\}\bigg).$$

Here, we again use index H to emphasize that the expectation is taken w.r.t. to H_T with $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^{T}$ being treated as fixed function, so $\Gamma(\epsilon)$ is a random variable w.r.t. sample. Our goal is to show that for any

$\omega \geq 2(\epsilon_n + \delta)$, we will obtain

$$P(\Gamma(\epsilon_n/M) > K_2 T\omega) \leq 3e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}. \tag{3.54}$$

To verify (3.54), we first note that according to the augmentation decomposition used in the proof of Lemma 3.5 presented at the end of this section, the U-function defined in Assumption 3.7 can be written as the summation of expectations w.r.t. non-negative response variables. Therefore, without loss of generality, we can assume that O_γ is non-negative during the proof. In this case, U_t in Assumption 3.7 can be simplified as

$$U_t(H_t; f_t, ..., f_T; \gamma) = E\left[O_\gamma \frac{\prod_{s=t}^T \mathbb{I}(A_s f_s(H_s) > 0)}{\prod_{s=t}^T p(A_s|H_s)} \middle| H_t\right].$$

For convenience, we use $\widehat{g}_t$ to denote $\text{sign}(\widehat{f}_{t,\widehat{\gamma}})$ and let

$$\widehat{S}_t = \{H_t \in \mathcal{H}_t : |\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \epsilon_n/M\},$$

and

$$\widehat{D}_t = \left\{(H_1, A_1, ..., H_t, A_t) : \prod_{s=1}^t \mathbb{I}(A_s \widehat{f}_{s,\widehat{\gamma}}(H_s)) = 1, H_t \in \widehat{S}_t\right\}.$$

We consider two situations:

Case 1: Suppose

$$E\left[\frac{\mathbb{I}(H_t \in \widehat{D}_t)}{\prod_{s=1}^t p(A_s|H_s)} U_{t+1}(H_{t+1}; g^*_{t+1,\widehat{\gamma}}, ..., g^*_{T,\widehat{\gamma}})\right] \leq \omega,$$

then by Assumption 3.7 we have $P_H(\widehat{D}_t) \leq C\omega$.

Case 2: Otherwise, suppose

$$E\left[\frac{\mathbb{I}(H_t \in \widehat{D}_t)}{\prod_{s=1}^t p(A_s|H_s)} U_{t+1}(H_{t+1}; g^*_{t+1,\widehat{\gamma}}, ..., g^*_{T,\widehat{\gamma}})\right] > \omega, \tag{3.55}$$

in this case, we aim at showing that

$$E\left[O_{\widehat{\gamma}} \frac{\min(\psi(A_1 g^*_{1,\widehat{\gamma}}(H_1)), ..., \psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right] - E\left[O_{\widehat{\gamma}} \frac{\min(\psi(A_1 \widehat{f}_{1,\widehat{\gamma}}(H_1)), ..., \psi(A_T \widehat{f}_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right] \geq \epsilon_n + \delta.$$

First, using the fact that

$$
\begin{aligned}
&E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 g^*_{1,\widehat{\gamma}}(H_1)),...,\psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
\geq &E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_t \widehat{g}_t(H_t)),\psi(A_{t+1} g^*_{t+1,\widehat{\gamma}}(H_{t+1})),...,\psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{f}_{1,\widehat{\gamma}}(H_1)),...,\psi(A_T \widehat{f}_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
\leq &E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_{t-1} \widehat{g}_{t-1}(H_{t-1})),\psi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t)),\psi(A_{t+1} \widehat{g}_{t+1}(H_{t+1})),...,\psi(A_T \widehat{g}_T(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]
\end{aligned}
$$

where the first inequality is guaranteed by the optimality of $(g^*_{1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})$ and the second inequality is guaranteed by noting that $\psi(\cdot)$ is a non-decreasing function, we can obtain that

$$
\begin{aligned}
&E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 g^*_{1,\widehat{\gamma}}(H_1)),...,\psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right] - E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{f}_{1,\widehat{\gamma}}(H_1)),...,\psi(A_T \widehat{f}_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
\geq &E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_t \widehat{g}_t(H_t)),\psi(A_{t+1} g^*_{t+1,\widehat{\gamma}}(H_{t+1})),...,\psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
&-E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_{t-1} \widehat{g}_{t-1}(H_{t-1})),\psi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t)),\psi(A_{t+1} \widehat{g}_{t+1}(H_{t+1})),...,\psi(A_T \widehat{g}_T(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right].
\end{aligned}
\tag{3.56}
$$

Sequentially taking conditional expectation w.r.t. H_T to H_{t+1} in backward order, the last term of (3.56) is equal to

$$
\begin{aligned}
&E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_t \widehat{g}_t(H_t)),\psi(A_{t+1} g^*_{t+1,\widehat{\gamma}}(H_{t+1})),...,\psi(A_T g^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
&- E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_{t-1} \widehat{g}_{t-1}(H_{t-1})),\psi(A_t \widehat{f}_{t,\widehat{\gamma}}(H_t)),\psi(A_{t+1} \widehat{g}_{t+1}(H_{t+1})),...,\psi(A_T \widehat{g}_T(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\\
=&E\left[\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_t \widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}U_{t+1}(H_{t+1};g^*_{t+1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})\right]\\
&- E\left[\frac{\min(\psi(A_1 \widehat{g}_1(H_1)),...,\psi(A_t \widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}|\psi(\widehat{f}_{t,\widehat{\gamma}}(H_t))|U_{t+1}(H_{t+1};\widehat{g}_{t+1},...,\widehat{g}_T)\right]
\end{aligned}
\tag{3.57}
$$

Since we have

$$
|\psi(\widehat{f}_{t,\widehat{\gamma}}(H_t))|U_t(H_t;\widehat{f}_{t,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}}) \leq U_t(H_t;\widehat{f}_{t,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}}) \leq U_t(H_t;g^*_{t,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}}),
$$

(3.57) is further lower bounded by

$$
\begin{aligned}
&E\left[\frac{\min(\psi(A_1\widehat{g}_1(H_1)),...,\psi(A_t\widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}U_{t+1}(H_{t+1};g^*_{t+1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})\right]\\
&-E\left[\frac{\min(\psi(A_1\widehat{g}_1(H_1)),...,\psi(A_t\widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}|\psi(\widehat{f}_{t,\widehat{\gamma}}(H_t))|U_{t+1}(H_{t+1};\widehat{g}_{t+1},...,\widehat{g}_T)\right]\\
\geq&E\left[\frac{\min(\psi(A_1\widehat{g}_1(H_1)),...,\psi(A_t\widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)U_{t+1}(H_{t+1};g^*_{t+1},...,g^*_T)\right]\\
&-E\left[\frac{\min(\psi(A_1\widehat{g}_1(H_1)),...,\psi(A_t\widehat{g}_t(H_t)))}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)|\psi(\widehat{f}_{t,\widehat{\gamma}}(H_t))|U_{t+1}(H_{t+1};\widehat{g}_{t+1},...,\widehat{g}_T)\right]\\
=&E\left[\frac{\prod_{s=1}^t\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)U_{t+1}(H_{t+1};g^*_{t+1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})\right]\\
&-E\left[\frac{\prod_{s=1}^t\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)|\widehat{f}_{t,\widehat{\gamma}}(H_t)|U_{t+1}(H_{t+1};\widehat{f}_{t+1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}})\right]\\
=&E\left[\frac{\mathbb{I}(H_t\in\widehat{D}_t)}{\prod_{s=1}^t p(A_s|H_s)}U_{t+1}(H_{t+1};g^*_{t+1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})\right]\\
&-E\left[\frac{\prod_{s=1}^t\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)|\widehat{f}_{t,\widehat{\gamma}}(H_t)|U_{t+1}(H_{t+1};\widehat{f}_{t+1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}})\right]
\end{aligned}
\tag{3.58}
$$

Hence, when (3.55) holds, (3.58) will be lower bounded by

$$
\begin{aligned}
&E\left[\frac{\mathbb{I}(H_t\in\widehat{D}_t)}{\prod_{s=1}^t p(A_s|H_s)}U_{t+1}(H_{t+1};g^*_{t+1,\widehat{\gamma}},...,g^*_{T,\widehat{\gamma}})\right]\\
&-E\left[\frac{\prod_{s=1}^t\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=1}^t p(A_s|H_s)}\mathbb{I}(H_t\in\widehat{S}_t)|\widehat{f}_{t,\widehat{\gamma}}(H_t)|U_{t+1}(H_{t+1};\widehat{f}_{t+1,\widehat{\gamma}},...,\widehat{f}_{T,\widehat{\gamma}})\right]\\
\geq&\omega-E\left[\frac{\prod_{s=1}^t\mathbb{I}(A_s\widehat{f}_{s,\widehat{\gamma}}(H_s)>0)}{\prod_{s=1}^t p(A_s|H_s)}\frac{\epsilon_n}{M}M\right]\\
>&\epsilon_n+\delta.
\end{aligned}
\tag{3.59}
$$

Combine (3.56), (3.57), (3.58) and (3.59), we have shown that (3.55) implies

$$
E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1g^*_{1,\widehat{\gamma}}(H_1)),...,\psi(A_Tg^*_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]-E\left[O_{\widehat{\gamma}}\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)))}{\prod_{s=1}^T p(A_s|H_s)}\right]\geq\epsilon_n+\delta,
$$

which holds with probability no more than $2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$ as verified in *step 1*.

The discussion of *case 1* and *case 2* indicates that

$$
P_H(\widehat{D}_t)>C\omega
$$

can only hold with probability no more than $2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. Note that the argument of *case 1* and *case 2* holds for any $t \in \{1, ..., T\}$, taking summation w.r.t. t we can obtain that

$$\begin{aligned}
&P(\Gamma(\epsilon_n/M) \geq K_2 T\omega) \\
=&P(\sum_{t=1}^{T} P_H(\widehat{D}_t) \geq K_2 T\omega) \\
\leq&P(\sup_t P_H(\widehat{D}_t) > C\omega) \\
\leq&2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n},
\end{aligned}$$

which completes the verification of (3.54).

We now begin to verify (C2) by utilizing (3.54). Recall that our goal is to verify

$$\left| E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] - \tau \right| \leq c(\epsilon_n + \xi_n + \delta).$$

Note that the estimated decision functions satisfy (3.50), by applying Lemma 3.4 we can obtain that the inequality (3.40) also holds with high probability. Hence, the key to complete the proof is to show that

$$\left| E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta), ..., \psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)} \right] - E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T} p(A_t|H_t)} \right] \right| \leq O(\epsilon_n).$$

To achieve this, we let

$$\begin{aligned}
\mathcal{D} :=&\left\{ (H_1, A_1, ..., H_T, A_T) : \prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0) = 1, \exists t \in \{1, ..., T\} \text{ such that } |\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \frac{\epsilon_n}{M} \right\} \\
\subseteq&\bigcup_{t=1}^{T}\left\{ (H_1, A_1, ..., H_t, A_t) : \prod_{s=1}^{t}\mathbb{I}(A_s\widehat{f}_{t,\widehat{\gamma}}(H_s) > 0) = 1, |\widehat{f}_{t,\widehat{\gamma}}(H_t)| \leq \frac{\epsilon_n}{M} \right\}
\end{aligned}$$

and choose $\omega = 2(\epsilon_n + \delta)$, then (3.54) implies that

$$P_H(\mathcal{D}) \leq 2K_2 T(\epsilon_n + \delta) \tag{3.60}$$

holds with probability of at least $1 - 2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. Therefore, we have

$$
\begin{aligned}
&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right]\\
=&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T,A_T)\in\mathcal{D})\right]\\
&+E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T,A_T)\in\mathcal{D}^c)\right]\\
\geq&E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T,A_T)\in\mathcal{D}^c)\right]\\
\overset{(i)}{=}&E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T,A_T)\in\mathcal{D}^c)\right]\\
\geq&E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]\\
&-E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\mathbb{I}((H_1,A_1,...,H_T,A_T)\in\mathcal{D})\right]\\
\overset{(ii)}{\geq}&E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]-C_3Tc_1^{-T}T(\epsilon_n+\delta).
\end{aligned}
\tag{3.61}
$$

holds with probability at least $1-2e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. Here, equality (i) is followed by noting that for any $\eta\leq\frac{\epsilon_n}{M}$, $|\widehat{f}_{t,\widehat{\gamma}}(H_t)|\geq\frac{\epsilon_n}{M}$ implies

$$
\psi(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)/\eta)=\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0),
$$

and (ii) is guaranteed by (3.60). Combine (3.61) and (3.33), we obtain that

$$
\left|E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right]-E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]\right|\leq C_3(\epsilon+\delta).
$$

On the other hand, recall that the estimated decision functions satisfy (3.50), by applying Lemma 3.4 we can obtain that

$$
\left|E\left[R\frac{\min(\psi(A_1\widehat{f}_{1,\widehat{\gamma}}(H_1)/\eta),...,\psi(A_T\widehat{f}_{T,\widehat{\gamma}}(H_T)/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right]-\tau\right|\leq\xi_n+\delta
$$

holds with probability at least $1-e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. Combining the two inequalities above, we have that

$$
\left|E\left[R\frac{\prod_{t=1}^T\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^T p(A_t|H_t)}\right]-\tau\right|\leq O(\epsilon_n+\xi_n+\delta)
$$

holds with probability at least $1 - 3e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$. Hence, (C2) holds by choosing $b_n = C_3 T c_1^{-T}\epsilon_n + \xi_n$, $\delta_2 = C_3 T c_1^{-T}\delta$ with η choosing to be $\eta = \epsilon_n / N$ in ξ_n.

Step 3 - verify condition (C3): Using exactly the same argument of *step 3* in the proof of Theorem 3.1, we can show that

$$P(\widehat{\gamma} > 1 - \zeta) \leq 1 - 3e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$$

holds for any $0 < \delta$ and n such that $C_3 T c_1^{-T}(\epsilon_n + \delta) + \xi_n \leq \frac{\tau - \tau_0}{6}$, which completes the verification of (C3).

Complete the proof of Theorem 3.2: step1 - 3 indicate that both condition (C1), (C2) and (C3) hold with probability at least $1 - 3e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$ with $a_n = \epsilon_n$, $b_n = C_3 T c_1^{-T}\epsilon_n + \xi_n$, $\delta_1 = \delta$ and $\delta_2 = C_3 T c_1^{-T}\delta$ for sufficient large n and sufficient small δ. Therefore, Theorem 3.2 will again be proved by directly applying the conclusion of Lemma 3.3.

We complete the proof of two additional lemmas used in the proof of Theorem 3.2. Lemma 3.5 shows that the excessive risk of arbitrary decision function $(f_1, ..., f_T)$ is upper bounded by the excessive risk under the surrogate objective function of MRL.

Lemma 3.5 *For any random variable O and any decision functions $(f_1, ..., f_T)$,*

$$\mathcal{V}(g_1^*, ..., g_T^*) - \mathcal{V}(f_1, ..., f_T) \leq \mathcal{V}_\psi(g_1^*, ..., g_T^*) - \mathcal{V}_\psi(f_1, ..., f_T).$$

Proof: Lemma 3.2 ensures that $\mathcal{V}_\psi(g_1^*, ..., g_T^*) = \mathcal{V}(g_1^*, ..., g_T^*)$. Moreover, using the augmentation decomposition we have

$$\begin{aligned}
&\mathcal{V}(g_1^*, ..., g_T^*) - \mathcal{V}(f_1, ..., f_T) \\
=&\left(\mathcal{V}(g_1^*, ..., g_T^*) + E\left[\frac{O^-}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) - \left(\mathcal{V}(f_1, ..., f_T) + E\left[\frac{O^-}{\prod_{t=1}^T p(A_t|H_t)}\right]\right) \\
=&E\left[O^+ \frac{\prod_{t=1}^T \mathbb{I}(A_t g_t^*(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] + E\left[\sum_{a_t \in \{-1,+1\},\, a_t \neq A_t} O^- \frac{\prod_{t=1}^T \mathbb{I}(a_t g_t^*(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
&- E\left[O^+ \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] - E\left[\sum_{a_t \in \{-1,+1\},\, a_t \neq A_t} O^- \frac{\prod_{t=1}^T \mathbb{I}(a_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right],
\end{aligned}$$

and by definition

$$
\begin{aligned}
&\mathcal{V}_\psi(g_1^*, ..., g_T^*) - \mathcal{V}_\psi(f_1, ..., f_T) \\
=&E\left[O^+ \frac{\min(\psi(A_1 g_1^*(H_1)), ..., \psi(A_T g_T^*(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
&+ E\left[\sum_{a_t \in \{-1,+1\},\ a_t \neq A_t} O^- \frac{\min(\psi(a_1 g_1^*(H_1)), ..., \psi(a_T g_T^*(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
&- E\left[O^+ \frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\
&- E\left[\sum_{a_t \in \{-1,+1\},\ a_t \neq A_t} O^- \frac{\min(\psi(a_1 f_1(H_1)), ..., \psi(a_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right],
\end{aligned}
$$

where now both O^+ and O^- are non-negative response variables. Hence, without loss of generality, it is sufficient to prove the result for non-negative weight O.

When O is non-negative, we have

$$
\mathcal{V}_\psi(f_1, ..., f_T) = E\left[O \frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right].
$$

Note that

$$
\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_T))) \leq \prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)
$$

holds for any $(H_1, A_1, ..., H_T, A_t)$. Consequently, $\mathcal{V}_\psi(f_1, ..., f_T) \leq \mathcal{V}(f_1, ..., f_T)$ holds for any decision functions which implies that

$$
\mathcal{V}(g_1^*, ..., g_T^*) - \mathcal{V}(f_1, ..., f_T) \leq \mathcal{V}_\psi(g_1^*, ..., g_T^*) - \mathcal{V}_\psi(f_1, ..., f_T).
$$

This completes the proof of Lemma 3.5. □

Lemma 3.6 provides an upper bound of the covering number of MRL. In the proof of Lemma 3.6, we abuse the notation and use $\mathcal{F}_t$ to denote an arbitrary set of measurable functions defined on $\mathcal{H}_t$.

Lemma 3.6 *Let*

$$
\mathcal{W}_1 = \left\{\{(1-\gamma)Y - \gamma R\} L_1(f_1, ..., f_T) : \gamma \in [0,1], f_1 \in \mathcal{F}_1, ..., f_T \in \mathcal{F}_T\right\},
$$

$$\mathcal{W}_2 = \left\{ \{(1-\gamma)Y - \gamma R\}^- L_2(|f_1|, ..., |f_T|) : \gamma \in [0,1], f_1 \in \mathcal{F}_1, ..., f_T \in \mathcal{F}_T \right\},$$

then for any positive $\epsilon \to 0$ *we have*

$$\log \mathcal{N}(\epsilon; \mathcal{W}_1, L_2(\mathbb{P}_n)) \leq C(-\log \epsilon + T + \sum_{t=1}^{T} \log \mathcal{N}(c_1^T \epsilon/(2MT); \mathcal{F}_t, L_2(\mathbb{P}_n))) \tag{3.62}$$

and

$$\log \mathcal{N}(\epsilon; \mathcal{W}_2, L_2(\mathbb{P}_n)) \leq C(-\log \epsilon + T + \sum_{t=1}^{T} \log \mathcal{N}(c_1^T \epsilon/(2MT); \mathcal{F}_t, L_2(\mathbb{P}_n))), \tag{3.63}$$

where C is a positive constant that only depends on (M, c_1).

Proof: We first prove the conclusion (3.62). Note that any function in $\mathcal{W}_1$ is the product of two functions from functional space

$$\mathcal{W}_{11} = \{(1-\gamma)Y - \gamma R | \gamma \in [0,1]\}$$

and

$$\mathcal{W}_{12} = \left\{ \frac{\min(\psi(A_1 f_1(H_1)), ..., \psi(A_T f_T(H_t)))}{\prod_{t=1}^{T} p(A_t|H_t)} \middle| f_1 \in \mathcal{F}_1, ..., f_T \in \mathcal{F}_T \right\}.$$

Hence, for any function l_1 and l_2 in $\mathcal{W}_1$, using inequality

$$(a+b)^2 \leq 2a^2 + 2b^2 \quad \forall a, b$$

the $L_2(\mathbb{P}_n)$ distance between l_1 and l_2 satisfies

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(l_1(x_i) - l_2(x_i))^2 =& \frac{1}{n}\sum_{i=1}^{n}(g_1(x_i)f_1(x_i) - g_2(x_i)f_2(x_i))^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}(g_1(x_i)f_1(x_i) - g_1(x_i)f_2(x_i) + g_1(x_i)f_2(x_i) - g_2(x_i)f_2(x_i))^2 \\
\leq& \frac{2}{n}\sum_{i=1}^{n} g_1^2(x_i)(f_1(x_i) - f_2(x_i))^2 + \frac{2}{n}\sum_{i=1}^{n} f_2^2(x_i)(g_1(x_i) - g_2(x_i))^2 \\
\leq& \frac{2M^2}{n}\sum_{i=1}^{n}(f_1(x_i) - f_2(x_i))^2 + \frac{2c_1^{-2T}}{n}\sum_{i=1}^{n}(g_1(x_i) - g_2(x_i))^2
\end{aligned}$$

where $g_1, g_2 \in \mathcal{W}_{11}$ and $f_1, f_2 \in \mathcal{W}_{12}$. The inequality above implies that

$$\mathcal{N}(\epsilon; \mathcal{W}_1, L_2(\mathbb{P}_n)) \leq \mathcal{N}(c_1^T \epsilon/2; \mathcal{W}_{11}, L_2(\mathbb{P}_n)) \mathcal{N}(\epsilon/(2M); \mathcal{W}_{12}, L_2(\mathbb{P}_n)).$$

For $\mathcal{W}_{11}$, it is easy to verify

$$\mathcal{N}(c_1^T \epsilon/2; \mathcal{W}_{11}, L_2(\mathbb{P}_n)) \leq 2M/c_1^T \epsilon.$$

Furthermore, analogous to the proof of Lemma 3.4 we can show that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_1(x_i) - f_2(x_i))^2 \\
\overset{(i)}{\leq} &\frac{1}{n}\sum_{i=1}^{n} c_1^{-2T}(\sum_{t=1}^{T}|\psi(A_{it}f_{1t}(H_{it})) - \psi(A_{it}f_{2t}(H_{it}))|)^2 \\
\overset{(ii)}{\leq} &\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} T c_1^{-2T}[\psi(A_{it}f_{1t}(H_{it})) - \psi(A_{it}f_{2t}(H_{it}))]^2 \\
\overset{(iii)}{\leq} &\sum_{t=1}^{T}\frac{T c_1^{-2T}}{n}\sum_{i=1}^{n}(f_{1t}(H_{it}) - f_{2t}(H_{it}))^2
\end{aligned} \tag{3.64}$$

where again we use f_{it} to denote the t-th stage decision function of f_i, i.e., $f_i = L_1(f_{i1}, ..., f_{iT})$. Inequality (i) is guaranteed by Proposition 3.1 and note that any function from $\mathcal{W}_{12}$ is bounded by c_1^{-T}, (ii) is followed by the Cauchy-Schwarz inequality and the last inequality (iii) holds by noting that $\psi(\cdot)$ is a 1-Lipschitz function. Inequality (3.64) implies that

$$\mathcal{N}(\epsilon; \mathcal{W}_{12}, L_2(\mathbb{P}_n)) \leq \prod_{t=1}^{T}\mathcal{N}(T^{-1}c_1^T\epsilon; \mathcal{F}_t, L_2(\mathbb{P}_n)). \tag{3.65}$$

Taking logarithm transformation, we have

$$\begin{aligned}
&\log \mathcal{N}(\epsilon; \mathcal{W}_1, L_2(\mathbb{P}_n)) \\
\leq &\log \mathcal{N}(c_1^T\epsilon/2; \mathcal{W}_{11}, L_2(\mathbb{P}_n)) + \log \mathcal{N}(\epsilon/(2M); \mathcal{W}_{12}, L_2(\mathbb{P}_n)) \\
\leq &\log(2Mc_1^{-T}) - \log\epsilon + \sum_{t=1}^{T}\log \mathcal{N}(c_1^T\epsilon/(2MT); \mathcal{F}_t, L_2(\mathbb{P}_n)) \\
= &C(-\log\epsilon + T + \sum_{t=1}^{T}\log \mathcal{N}(c_1^T\epsilon/(2MT); \mathcal{F}_t, L_2(\mathbb{P}_n)))
\end{aligned} \tag{3.66}$$

which completes the proof of (3.62).

Inequality (3.63) can be established using the same argument and the fact that $l(x) = x^-$ and $l(x) = |x|$ are both 1-Lipschitz functions. Specifically, let

$$\mathcal{W}_{21} = \{[(1-\gamma)Y - \gamma R]^- | \gamma \in [0,1]\},$$

$$\mathcal{W}_{22} = \left\{ \frac{\min(\psi(f_1), ..., \psi(f_T))}{\prod_{t=1}^T p(A_t|H_t)} \middle| f_1 \in |\mathcal{F}_1|, ..., f_T \in |\mathcal{F}_T| \right\}$$

where $|\mathcal{F}_t| = \{|f| : f \in \mathcal{F}_t\}$, then by repeating the same argument for $\mathcal{W}_1$ we can obtain

$$\mathcal{N}(\epsilon; \mathcal{W}_{21}, L_2(\mathbb{P}_n)) \leq \mathcal{N}(\epsilon; \mathcal{W}_{11}, L_2(\mathbb{P}_n)),$$

and

$$\begin{aligned}\mathcal{N}(\epsilon; \mathcal{W}_{22}, L_2(\mathbb{P}_n)) \leq & \prod_{t=1}^T \mathcal{N}(c_1^T \epsilon/(2MT); |\mathcal{F}_t|, L_2(\mathbb{P}_n)) \\ \leq & \prod_{t=1}^T \mathcal{N}(c_1^T \epsilon/(2MT); \mathcal{F}_t, L_2(\mathbb{P}_n)),\end{aligned}$$

and (3.63) can be obtained via exactly the same argument as (3.66). □

3.10 Additional Simulation Results

In the first part, we present the additional simulation result with $T = 4$ under a more complicated treatment design. In this simulation, $(Z_1, ..., Z_8)$ are 8 independent baseline feature variables generated from the uniform distribution $U[0,1]$. We use $(Y_0, ..., Y_4)$ and $(R_0, ..., R_4)$ to denote the reward and risk outcome at baseline and time point $t = 1, ..., 4$. Y_0 and R_0 are generated according to

$$Y_0 = Z_1 + Z_2/2 + \epsilon,$$

$$R_0 = Z_1/2 + Z_2 + \epsilon.$$

where ϵ denotes the noisy term generated from independent normal distribution N[0,1] truncated at ± 0.5. For $t = 1, .., 4$, (A_t, Y_t, R_t) are generated according to

- $t = 1$:

$$\text{logit}P(A_1 = 1|H_1) = 0.5,$$

$$Y_1 = Z_1 + A_1(Y_0 - 1) + \epsilon,$$

$$R_1 = Z_2 + A_1(R_0 - 1) + \epsilon.$$

- $t = 2$:

$$\text{logit}P(A_2 = 1|H_2) = 0.5,$$

$$Y_2 = Z_1 + A_2(Y_1 - 1) + \epsilon,$$

$$R_2 = Z_2 + A_2(R_1 - 1) + \epsilon,$$

- $t = 3$:

$$\text{logit}P(A_3 = 1|H_3) \sim Y_2/4 - R_2/4,$$

$$Y_3 = Z_1 + 2A_3(Y_2/2 + R_2/2 - 0.5) + \epsilon,$$

$$R_3 = Z_2 + 2A_3(R_2 - 0.5) + \epsilon,$$

- $t = 4$:

$$\text{logit}P(A_4 = 1|H_4) \sim Y_3/4 - R_3/4,$$

$$Y_4 = Z_1 + 2A_4(Y_3/2 + R_3/2 + 0.5) + \epsilon,$$

$$R_4 = Z_2 + 2A_4(+R_3 - 0.5) + \epsilon.$$

The feature variable at each stage is set to be

$$H_1 = (Z_1, ..., Z_8, Y_0, R_0),$$

$$H_2 = (Z_1, ..., Z_8, Y_0, R_0, A_1, Y_1, R_1),$$

$$H_3 = (Z_1, ..., Z_8, Y_0, R_0, ..., A_2, Y_2, R_2),$$

$$H_4 = (Z_1, ..., Z_8, Y_0, R_0, ..., A_3, Y_3, R_3),$$

N	Kernel	Method	Testing Reward	Testing Risk	Efficacy Ratio
200	Linear	MRL	2.320(0.431)	1.312(0.500)	1.916(0.376)
		OWL	2.520(0.347)	1.657(0.418)	1.670(0.206)
		AOWL	2.333(0.439)	1.686(0.425)	1.490(0.136)
		Q-learning	3.131(0.638)	2.248(0.857)	1.486(0.250)
		Unconstrained	3.980(0.138)	3.802(0.188)	1.102(0.025)
	Gaussian	MRL	2.568(0.455)	1.720(0.515)	1.564(0.221)
		OWL	2.532(0.474)	2.165(0.471)	1.271(0.078)
		AOWL	2.627(0.447)	2.120(0.452)	1.325(0.087)
		Q-learning	2.962(0.636)	2.200(0.770)	1.460(0.221)
		Unconstrained	3.974(0.145)	3.854(0.161)	1.095(0.017)
400	Linear	MRL	2.491(0.378)	1.465(0.438)	1.859(0.299)
		OWL	2.647(0.289)	1.677(0.337)	1.710(0.174)
		AOWL	2.335(0.408)	1.606(0.414)	1.511(0.142)
		Q-learning	3.020(0.451)	2.103(0.603)	1.540(0.209)
		Unconstrained	4.058(0.150)	3.892(0.171)	1.106(0.026)
	Gaussian	MRL	2.605(0.396)	1.725(0.474)	1.619(0.213)
		OWL	2.806(0.307)	2.278(0.300)	1.326(0.074)
		AOWL	2.603(0.297)	2.053(0.314)	1.357(0.065)
		Q-learning	2.835(0.599)	1.935(0.756)	1.580(0.253)
		Unconstrained	4.001(0.128)	3.866(0.148)	1.096(0.018)

Table 3.4: Summary table for additional simulation result. The results are reported in *median(dev)* format the same as Section 3.4.

and the cumulative reward and risk outcomes are set to be $Y = \sum_{t=1}^{4} Y_t$ and $R = \sum_{t=1}^{4} R_t$.

In this simulation, we assume that the treatment assignment probability at each stage is unknown and estimate the treatment assignment probability using training data repeatedly via Lasso logistic regression. We analyze the training data of size $n = 200$ and 400 and repeat the analyses 500 times. The risk constraint is set to be $\tau = 2$ and we repeated the analysis for both MRL, OWL, AOWL, and Q-learning following the same setting in Section 3.4, except that for MRL, OWL, and AOWL, we reduce the tuning grid by forcing $\lambda_1 = \cdots = \lambda_4$ and choose the optimal tuning parameters from $\lambda_t \in 2^{-8,-6,\cdots,+6,+8}$. The performance of the estimated rules is estimated on independent testing data of size $n = 5,000$. The efficacy ratio is also reported with the reference treatment rules set to be the safest rules that induce the minimum risk among 16 possible one-size-fits-all rules. The analysis result is summarized in Table 3.4. From the table, we observe that MRL still achieves the best risk control with the highest efficacy ratio compared with OWL, AOWL, and Q-learning under both $n = 200$ and $n = 400$ for both two kernels. This result is consistent with the findings in Section 3.4.

η	Method	Testing Reward	Testing Risk	Efficacy Ratio
10^{-2}	MRL	1.667(0.150)	0.845(0.156)	1.041(0.151)
	OWL	1.797(0.114)	0.893(0.143)	1.292(0.179)
	AOWL	1.878(0.123)	0.975(0.171)	1.287(0.198)
	Q-learning	1.868(0.106)	1.096(0.174)	1.028(0.143)
10^{-3}	MRL	1.649(0.162)	0.840(0.175)	1.020(0.170)
	OWL	1.795(0.118)	0.896(0.152)	1.284(0.183)
	AOWL	1.875(0.122)	0.959(0.165)	1.291(0.188)
	Q-learning	1.846(0.107)	1.066(0.170)	1.042(0.146)
10^{-4}	MRL	1.658(0.165)	0.837(0.167)	1.022(0.158)
	OWL	1.792(0.118)	0.894(0.146)	1.289(0.181)
	AOWL	1.875(0.122)	0.962(0.160)	1.290(0.187)
	Q-learning	1.844(0.107)	1.062(0.172)	1.043(0.149)
10^{-5}	MRL	1.635(0.175)	0.829(0.173)	1.017(0.173)
	OWL	1.792(0.118)	0.894(0.146)	1.289(0.181)
	AOWL	1.875(0.122)	0.962(0.161)	1.288(0.186)
	Q-learning	1.844(0.108)	1.062(0.173)	1.044(0.149)

Table 3.5: Sensitive analysis of CBR under the different choice of η.

In the second part, we repeat the simulation study under Setting I with $n = 200$ in Section 3.4 under different η to evaluate the impact of parameter η on the performance of each method. The simulation is conducted following the same description as Section 3.4 except that η is varied in $\{10^{-2}, ..., 10^{-5}\}$. The results are displayed in Table 3.5. From the table, we observe that when η decreases from 10^{-2} to 10^{-5}, the risk control of all four methods will slightly improve but the performances are roughly the same under different choices of η. This indicates that Algorithm 2 is not sensitive against η when η is small.

CHAPTER 4: SIMULTANEOUS VARIABLE SELECTION AND LEARNING FOR DYNAMIC TREATMENT REGIMENS

4.1 Introduction

The demand for identifying key biomarkers that helps treatment design in precision medicine has raised the concern of developing new DTRs method with the capability of both maximizing patients' beneficial outcome and eliminating unimportant variables that have little contribution to improving patients' health condition. Driven by this, many new DTRs methods have been designed to address both goals. Among them, most of the proposed methods refine and extend existing regression-based approaches and achieve variable selection by incorporating additional penalty terms to impose sparsity over the estimated decision rules. These methods include A-learning based methods (Gunter, Zhu and Murphy, 2011; Shi et al., 2018), Q-learning based methods (Qian and Murphy, 2011; Song, Wang, Zeng and Kosorok, 2015; Ghosh et al., 2022), G-estimation based method (Bian et al., 2021) or semi-parametric modeling method (Guo, Zhou and Ma, 2021). For machine learning-based approaches, the variable selection is usually achieved by adding proper penalty term with variable selection capability to the objective function and the optimal DTRs are then learned via optimizing the penalized objective function, existing methods including L_∞-penalty extension of O-learning (Lu, Zhang and Zeng, 2013; Li et al., 2018; He et al., 2021), SCAD-penalty extension of O-learning (Song, Kosorok, Zeng, Zhao, Laber and Yuan, 2015) and ramp loss L_∞-penalty method (Huang, 2015).

However, serval limitations exist for each type of method. For regression-based methods, method performance will be significantly affected when regression models are misspecified like the standard case. For machine learning-based approaches, the methods mentioned early are all designed to handle single-stage optimal treatment regimen problems. More importantly, though existing single-stage machine learning-based approaches mentioned early can be extended to multistage DTRs setting by adopting the backward induction technique introduced in Zhao et al. (2015), to our best knowledge, all existing regression-based and machine learning-based approaches are only capable of learning optimal DTRs stage by stage separately in backward order. This causes a main disadvantage for existing methods that the decision rule estimation and variable

selection of later stages cannot fully utilize the information from the early stage, while in practice an important biomarker for early stages' treatment decision-making is also a strong implication that the biomarker will be important in later stages' decision making.

To overcome the disadvantages of existing methods, in this chapter, we propose a new machine learning-based approach to learn optimal DTRs with simultaneous variable selection capability across all stages. Specifically, we extend the multistage ramp loss (MRL) function proposed in Chapter 3 by incorporating an additional Lasso-type penalty term to impose sparsity over the estimated coefficients. We name this new method as L1-MRL. Our proposed method learns the optimal DTRs by maximizing a single objective function in terms of unknown decision functions of all stages, which is guaranteed to use all information to learn DTRs and conduct variable selection across all stages. Numerically, the optimization problem can be efficiently solved using DC algorithm (Tao and An, 1997) where the optimization problem can be reduced to a simple optimization problem of a piecewise linear function in each DC iteration.

The remaining sections are organized as follows. In Section 4.2, we briefly introduce the MRL and present the details of the L1-MRL framework in the same section. Theoretical justification of L1-MRL is provided in Section 4.3. In Section 4.4, the performance of our proposed method is demonstrated via extensive simulation studies and comparisons with some of the existing methods. In Section 4.5, we apply our method to real observational electronic health record (EHR) data of type II diabetes (T2D) patients. The discussion and future extension of this chapter are presented in Section 4.6. The algorithm for solving L1-MRL and proofs are presented in Section 4.7 and Section 4.8.

4.2 Method

4.2.1 Learning optimal DTRs via multistage ramp loss (MRL)

Consider a T-stage decision-making problem with Y denoting the cumulative reward observed by the end of stage T and $\{A_t\}_{t=1}^T$ denoting the treatment assignment at each stage. Throughout this chapter, we assume that a higher value of Y indicates the better decision strategy and we assume two treatments, denoted as $\{-1, +1\}$, are available at each stage. For each t, we use H_t to denote patients' feature variables prior to stage t and let $\mathcal{H}_t \subset \mathbb{R}^{d_t}$ denote the feature space at each t. A dynamic treatment regimen is any function

from functional space

$$\mathcal{D} : \mathcal{D}_1 \times \cdots \times \mathcal{D}_T \to \{-1, +1\}^T, \quad \mathcal{D}_t : \mathcal{H}_t \to \{-1, +1\},$$

and the optimal DTRs $\mathcal{D}^*$ is defined to be the decision rules that maximize

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} \ E^{\mathcal{D}}[Y],$$

where $E^{\mathcal{D}}[\cdot]$ denotes the expectation with A_t forcing to be $\mathcal{D}(H_t)$.

To estimate $\mathcal{D}^*$, additional causal assumptions are required to ensure that $E^{\mathcal{D}}[Y]$ is estimable given observed data. Throughout this chapter, we assume that three standard causal assumptions made in Chapter 1 hold and repeat the assumptions below:

Assumption 4.1 *Stable Unit Treatment Value (SUTV): A subject's cumulative potential outcome is not influenced by other subjects' treatment allocation, i.e.,* $Y = Y(\bar{a}_T)$ *if* $\bar{A}_T = \bar{a}_T$.

Assumption 4.2 *No Unmeasured Confounders (NUC): For any* $t \in \{1, ..., T\}$ *and* $\bar{a}_T \in \{-1, +1\}^T$, $A_t \perp\!\!\!\perp (H_{t+1}(\bar{a}_t), ..., H_T(\bar{a}_{T-1}), Y(\bar{a}_T))\big|H_t$.

Assumption 4.3 *Positivity: For any* $t \in \{1, ..., T\}$, *there exists universal constants* $0 < c_1 \leq c_2 < 1$ *such that the treatment assignment probability at stage* t *satisfies* $c_1 \leq p(A_t = 1|H_t) \leq c_2$ *for* H_t *almost surely.*

Again under Assumption 4.1 to 4.3, Qian and Murphy (2011) shows that

$$E^{\mathcal{D}}[Y] = E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right].$$

Furthermore, suppose that the optimal DTRs are given by the signs of a series of optimal decision functions, i.e., there exists $f_t^* \in \mathcal{F}_t$, where $\mathcal{F}_t$ denotes the set of all measurable functions from $\mathcal{H}_t$ to $\mathbb{R}$, such that $\mathcal{D}^*(H_t) = \text{sign}(f_t^*(H_t))$ almost surely for $t = 1, ..., T$, then estimating the optimal DTRs is equivalent to the optimization problem

$$(f_1^*, ..., f_T^*) = \underset{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T}{\arg\max} E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right]. \tag{4.1}$$

Note that solving (4.1) directly is NP-hard due to the existence of the indicator functions, several methods have been proposed to address the computation challenge and estimate the optimal decision functions efficiently. In this chapter, we consider the multistage ramp loss (MRL) framework proposed in Chapter 3. Specifically, in Chapter 3 we consider the surrogate optimization problem

$$\begin{aligned}\max_{(f_1,...,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T} \; & E\left[Y^+\frac{\min(\psi(A_1f_1(H_1)),...,\psi(A_Tf_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] \\ & + E\left[\sum_{a_t\neq A_t} Y^-\frac{\min(\psi(a_1f_1(H_1)),...,\psi(a_Tf_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right],\end{aligned} \tag{4.2}$$

where Y^+ and Y^- denote the positive and negative part of Y defined as $Y^+ = \max(Y, 0)$ and $Y^- = \max(-Y, 0)$, and $\psi(x) = \max(\min(x, 1), 0)$. The objective function above can be viewed as a multistage extension of the shifted ramp loss function proposed in Huang, Shi and Suykens (2014). As one of the key properties of MRL, Lemma 3.2 in Chapter 3 shows that the optimization problem (4.2) is guaranteed to yield Fisher consistent estimators of $(f_1^*, ..., f_T^*)$. Moreover, from the expression (4.2) one can notice that f_{t_1} is not necessary to be a decision function before or after f_{t_2} for any t_1 and t_2. This special property indicates that MRL does determine the optimal decision function simultaneously without identifying early stages from later stages, which, unlike other methods such as Q-learning or OWL that must determine each stage's decision function separately, ensures that all information is used while determining the decision functions of all stages. As one of the main advantages due to the simultaneous property, for MRL the estimation of the early stage can provide feedback to the estimation of later stages; while for the backward induction-based method, the estimation error of later stages will be fixed and cumulated into the estimation of early stages once the later stages' decision rules have been estimated.

4.2.2 Variable selection via penalized MRL with adaptive coefficients

In many real applications, identifying important variables that contribute to treatment optimization is as important as finding optimal DTRs that yield the highest possible reward. In this chapter, we consider the specific decision-making problem where $H_t = (O_t, W_t)$ with O_t having fixed length P for $t = 1, ..., T$. From now on, we focus on linear decision rules and assume that the optimal decision function is linear in terms of feature variables, i.e., there exists $\boldsymbol{\theta}_t^* = (\boldsymbol{\alpha}_t^*, \boldsymbol{\beta}_t^*, \gamma_t^*) \in \mathbb{R}^{d_t+1}$ for $t = 1, ..., T$ such that

$$f_t^*(H_t) = W_t^T\boldsymbol{\alpha}_t^* + O_t^T\boldsymbol{\beta}_t^* + \gamma_t^*.$$

We assume that $\{\boldsymbol{\beta}_t^*\}_{t=1}^T$ are sparse and our goal is to learn optimal DTRs while recovering the sparsity within $\{\boldsymbol{\beta}_t^*\}_{t=1}^T$. In real applications, W_t can be viewed as already-unknown important variables for decision-making, such as treatment assignment from previous stages, so variable selection is not necessary for these features. $\{(O_{1p}, ..., O_{Tp})\}_{p=1}^P$ can be selected as P candidate variables for tailoring treatment to patients, which can take the same value to represent the impact of certain time-independent baseline covariates, or different values for time-varying variables such as patient's health test results obtained prior to each decision stage. To substantiate variable selection while learning the optimal rules, we consider adding a penalty term to MRL (4.2) to enforce sparsity over $\{\boldsymbol{\beta}_t\}_{t=1}^T$. Specifically, when treatment rules are assumed to be linear, we introduce L_1-penalty to MRL and consider following the L1-MRL optimization problem

$$\begin{aligned}\max_{\boldsymbol{\theta}} E\Bigg[Y^+ \frac{\min(\psi(A_1(H_1^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(A_T(H_T^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^T p(A_t|H_t)}\Bigg] \\ + E\Bigg[\sum_{a_t \neq A_t} Y^- \frac{\min(\psi(a_1(H_1^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(a_T(H_T^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^T p(A_t|H_t)}\Bigg] \\ - \lambda_n \sum_{p=1}^P \sum_{t=1}^T \frac{|\beta_{t,p}|}{\sqrt{\sum_{s=1}^T |\widetilde{\beta}_{s,p}|^2}}.\end{aligned} \tag{4.3}$$

Here, we assume that H_t includes an intercept term to simplify the notation, and we also abuse the notation and use $\{\widetilde{\beta}_{t,p}\}$ to denote the optimal solution of (4.1) normalized so that $\|(\widetilde{\beta}_{t,1}, ..., \widetilde{\beta}_{t,P})\|_2 = 1$ for all t. The L_1-penalty is a typical choice of penalty term to impose sparsity over the estimated coefficients and the additional adaptive coefficient $1/\sqrt{\sum_{t=1}^T |\widetilde{\beta}_{t,p}|^2}$ aims at imposing a stronger penalty to coefficients that are not significant across all stages. Like standard Lasso, λ_n is the tuning parameter to control the sparsity, and we also introduce shifting parameter $\eta_n \in (0, 1]$ as an additional tuning parameter to adjust for possible model misspecification under linearity assumption. We note that when $T = 1$, the L1-MRL framework will reduce to the ramp loss framework studied in Huang (2015).

Given finite samples, we estimate $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_T^*)$ by solving the empirical version of (4.3)

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ \frac{1}{n}\sum_{i=1}^n Y_i^+ \frac{\min(\psi(A_{i1}(H_{i1}^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(A_{iT}(H_{iT}^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^T p(A_{it}|H_{it})} \\ + \frac{1}{n}\sum_{i=1}^n \sum_{a_t \neq A_{it}} Y_i^- \frac{\min(\psi(a_1(H_{i1}^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(a_T(H_{iT}^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^T p(A_{it}|H_{it})} \\ - \lambda_n \sum_{p=1}^P \sum_{t=1}^T \frac{|\beta_{t,p}|}{\sqrt{\sum_{s=1}^T |\widehat{\widetilde{\beta}}_{s,p}|^2}},\end{aligned} \tag{4.4}$$

where $\widehat{\widetilde{\beta}}_{t,p}$ can be arbitrary estimators of $\widetilde{\beta}_{t,p}$ without penalty over coefficients. In application, we recommend using O-learning (Zhao et al., 2015; Liu et al., 2018) to estimate the unpenalized optimal coefficients. As another advantage of MRL, though being a non-convex optimization problem, the objective function of the minimization equivalence of (4.4) can be written as the difference between two convex functions. Hence, the empirical problem (4.4) can be solved efficiently via the DC algorithm. When the optimal decision functions are linear as assumed in this chapter, in each DC iteration the problem can be reduced to an optimization problem with the objective function being piecewise linear, which can be further solved efficiently by calculating the derivatives explicitly and grid search. In particular, note that (4.1) is invariant if we subtract Y by any function of H_1, one can further refine (4.4) as

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \; & \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i^{+} \frac{\min(\psi(A_{i1}(H_{i1}^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(A_{iT}(H_{iT}^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} \\ & + \frac{1}{n}\sum_{i=1}^{n} \sum_{a_t \neq A_{it}} \widehat{Y}_i^{-} \frac{\min(\psi(a_1(H_{i1}^T\boldsymbol{\theta}_1)/\eta_n), ..., \psi(a_T(H_{iT}^T\boldsymbol{\theta}_T)/\eta_n))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} \\ & - \lambda_n \sum_{p=1}^{P}\sum_{t=1}^{T} \frac{|\beta_{t,p}|}{\sqrt{\sum_{s=1}^{T} |\widehat{\widetilde{\beta}}_{s,p}|^2}},\end{aligned} \tag{4.5}$$

where we replace Y_i by corresponding estimated residual $\widehat{Y}_i = Y_i - \widehat{E}[Y_i|H_{i1}]$ following the idea of AOWL, which turns out to have better performance under the MRL framework with the proposed L_1-penalty term. In practice, we proposed to estimate $E[Y|H_1]$ via standard linear or Lasso regression. A coordinate descent DC algorithm-based procedure for solving (4.4) and (4.5) is provided in Section 4.7.

L1-MRL requires that treatment assignment probability $\{p(A_t|H_t)\}_{t=1}^T$ are known which can be satisfied when data is collected from a simple randomized clinical trial or a SMART. When $\{p(A_t|H_t)\}_{t=1}^T$ are unknown such as in an observational study, to implement L1-MRL one can assume that $\{p(A_t|H_t)\}_{t=1}^T$ can be consistently estimated from observed data and replace $p(A_t|H_t)$ by $\widehat{p}(A_t|H_t)$ for $t = 1, ..., T$ in (4.4) or (4.5) respectively.

4.2.3 Choose optimal tuning parameters

When implementing L1-MRL, one must prespecify the choice of tuning parameters (λ_n, η_n). In practice, we propose to choose the optimal tuning parameters via cross-validation and consider following AIC-type

(Bozdogan, 1987) criterion

$$n * \log\left(\frac{\widehat{\mathcal{R}}(\widehat{\boldsymbol{\theta}}(\lambda_n,\eta_n),\eta_n)}{\widehat{\mathcal{R}}(\widehat{\boldsymbol{\theta}}(0,\eta_n),\eta_n)}\right) - k(\lambda_n,\eta_n). \tag{4.6}$$

Here,

$$\begin{aligned}\widehat{\mathcal{R}}(\boldsymbol{\theta},\eta) =& \sum_{i=1}^{n} \widehat{Y}_i^{+} \frac{\min(\psi(A_{i1}(H_{i1}^T\boldsymbol{\theta}_1)/\eta),...,\psi(A_{iT}(H_{iT}^T\boldsymbol{\theta}_T)/\eta))}{\prod_{t=1}^{T} p(A_{it}|H_{it})} \\ &+ \sum_{i=1}^{n} \sum_{a_t \neq A_{it}} \widehat{Y}_i^{-} \frac{\min(\psi(a_1(H_{i1}^T\boldsymbol{\theta}_1)/\eta),...,\psi(a_T(H_{iT}^T\boldsymbol{\theta}_T)/\eta))}{\prod_{t=1}^{T} p(A_{it}|H_{it})}\end{aligned}$$

denotes the empirical estimator of the MRL objective function with coefficients $\boldsymbol{\theta}$ and shifting parameter η. $\widehat{\boldsymbol{\theta}}(\lambda,\eta)$ denotes the estimated coefficients of (4.4) or (4.5) under tuning pair (λ,η) and $k(\lambda,\eta)$ denotes the number of nonzero coefficients of $\widehat{\boldsymbol{\theta}}(\lambda,\eta)$. Similar BIC-type criteria were early adopted in Shi et al. (2018) under an A-learning framework and He et al. (2021) under a doubly robust outcome weighted learning framework. When implementing L1-MRL, we choose the optimal tuning parameter that maximizes (4.6) on testing data under cross-validation.

4.3 Theoretical Results

We present the theoretical result and show that the estimated coefficients obtained via L1-MRL enjoy the oracle property under mild conditions in this section. To state the necessary conditions, we let

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) =& - E\left[Y \frac{\min(\psi(A_1 H_1^T\boldsymbol{\theta}_1/\eta),...,\psi(A_T H_T^T\boldsymbol{\theta}_T/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\right] \\ &- E\left[Y^{-} \frac{\min(\psi(|H_1^T\boldsymbol{\theta}_1|/\eta),...,\psi(|H_T^T\boldsymbol{\theta}_T|/\eta))}{\prod_{t=1}^{T} p(A_t|H_t)}\right],\end{aligned}$$

so $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*,...,\boldsymbol{\theta}_T^*)$ will be the minimizer of $\mathcal{L}(\boldsymbol{\theta})$ by definition. In addition to the standard causal assumptions, we assume that the following two assumptions hold.

Assumption 4.4 *$\mathcal{H}_t \subset \mathbb{R}^{d_t}$ is compact for $t = 1,...,T$. Moreover, there exists a constant $B > 0$ such that $\|\boldsymbol{\theta}_t^*\|_\infty \leq B$ holds for $t = 1,...,T$.*

Assumption 4.5 *$\mathcal{L}(\boldsymbol{\theta})$ is three times differentiable,*

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^*) = 0, \quad \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*) \succ 0,$$

and the third order derivatives of $\mathcal{L}(\boldsymbol{\theta})$ is bounded.

Assumption 4.4 and 4.5 are regularity assumptions analogous to the regularity assumptions adopted in Fan and Li (2001) for establishing the oracle property for non-convex penalized regression. Basically, Assumption 4.4 and 4.5 ensure that the optimal solution for the unpenalized problem exists and the objective function is locally strictly convex near the optimal. Under the additional assumptions, we have

Theorem 4.1 *Assume that Assumption 4.4 and Assumption 4.5 also hold and for any fixed $\xi \in (0, \frac{1}{2})$, we have $a_n n^{\frac{1}{4}-\frac{\xi}{2}} = O(1)$ where*

$$a_n = \max\{\lambda_{np} : |\beta_{tp}^*| \neq 0\}.$$

Let shifting parameter η be a fixed constant, then there exists a local minimizer $\widehat{\boldsymbol{\theta}}_n$ of the empirical problem (4.4) such that

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n\|_2 = O_p(n^{-\frac{1}{4}+\frac{\xi}{2}} + a_n).$$

Consequently, we have $P(\max\{|\widehat{\beta}_{n,tp}| : \beta_{tp}^* = 0\} = 0) = 1$.

Theorem 4.1 implies that as n goes to infinite, there always exists a local minimizer of the empirical problem (4.4) within the ball of radius $O_p(n^{-\frac{1}{4}+\frac{\xi}{2}})$ centered true optimal solution $\boldsymbol{\theta}^*$. Consequently, $\widehat{\boldsymbol{\theta}}_n$ can recover the true sparsity of $\boldsymbol{\theta}^*$ with probability 1 as the sample size n goes to infinity. The proof of Theorem 4.1 is presented in Section 4.8.

4.4 Simulation Studies

We conduct simulation studies to assess the performance of L1-MRL in this section. We consider a two-stage SMART design and first generate 12 time-dependent baseline covariates $(Z_1, ..., Z_{12})$ from a multivariate normal distribution with mean 0, variance 1 and

$$Cov(Z_i, Z_j) = \begin{cases} 0.2, & \forall 1 \leq i < j \leq 6, \\ 0, & \text{for all other } i \neq j. \end{cases}$$

In addition, we generate 3 time-dependent covariates, denoted as (X_{11}, X_{12}), (X_{21}, X_{22}) and (X_{31}, X_{32}), independently according to

$$X_{i1} = Z_1\omega_i + \alpha_i + \epsilon_{i1}$$
$$X_{i2} = Z_1\omega_i + \alpha_i(1 + A_1/2) + \epsilon_{i2},$$

where $\{\omega_i\}_{i=1,2,3}$ are independently sampled from uniform distribution Unif$[0, 1]$, $\{\alpha_i\}_{i=1,2,3}$ are independently sampled from standard normal distribution N(0,1), and $\{(\epsilon_{i1}, \epsilon_{i2})\}_{i=1,2,3}$ are independent noisy terms sampled from standard normal distribution N(0,1).

In the simulation study, we consider the following two settings:

- **Setting I:**

$$\begin{aligned} Y = 1 + Z_1 + Z_3 \\ + A_1(3Z_1 + 3Z_2 - 2Z_7 - 2Z_8 - 2X_{11}) \\ + A_2(3Z_1 + 3Z_2 - 2Z_7 - 2Z_8 - 2X_{12}) + \epsilon_Y. \end{aligned}$$

- **Setting II:**

$$\begin{aligned} Y = 1 + Z_1 + Z_2 + \frac{1}{2}(Z_3^2 + Z_4^2) \\ + A_1\bigg((Z_1 + 3)^2 + (Z_2 + 3)^2 + (Z_7 - 3)^2 + (Z_8 - 3)^2 + \frac{2}{3}(X_{11} - 5)^2 - 60\bigg) \\ + A_2\bigg((Z_1 + 3)^2 + (Z_2 + 3)^2 + (Z_7 - 3)^2 + (Z_8 - 3)^2 + \frac{2}{3}(X_{12} - 5)^2 - 60\bigg) + \epsilon_Y. \end{aligned}$$

For both settings, ϵ_Y is generated from a standard normal distribution N(0,1) and treatments are randomly assigned following the regression model

$$\text{logit } P(A_1 = 1|Z, X_{11}, X_{21}, X_{31}) = \frac{1}{3}X_{11},$$
$$\text{logit } P(A_2 = 1|Z, X_{12}, X_{22}, X_{32}, A_1) = \frac{1}{3}X_{12} + \frac{1}{2}A_1.$$

In this simulation, we assume that the treatment assignment models are unknown. The feature variables for each stage are set to be

$$H_1 = (Z_1, ..., Z_{12}, X_{11}, X_{21}, X_{31}), \;\; H_2 = (Z_1, ..., Z_{12}, X_{12}, X_{22}, X_{32}, A_1),$$

and we repeatedly conduct the analyses for two settings 100 times with a training sample size of n=200 and 400. The performance under each setting is evaluated on an independent testing dataset of size 5,000 where the expected reward under the estimated rules is approximated via the Monte-Carlo method.

For L1-MRL, we solve the refinement problem (4.5) and select the optimal tuning parameters from tuning grid $(\lambda_n, \eta_n) \in 10^{-3:-5} \times 2^{0:10}$ according to AIC criterion (4.6) via two-folds CV. In this study, we use standard Lasso regression to estimate the conditional mean model $E[Y|H_1]$. When estimating the coefficients given tuning parameter (λ_n, η_n), we choose the initial iteration point of the DC algorithm to be the estimated coefficients of (4.5) under $(0, \eta_n)$ and using the solution of AOWL proposed by Liu et al. (2018) as the initial iteration point when estimating the coefficients under $(0, \eta_n)$. For the second stage, A_1 is excluded from variable selection, leaving $P = 15$ and only 12 time-independent and 3 time-dependent being penalized.

To demonstrate the performance of L1-MRL, we also compare our method with the following 4 competing methods:

- *Q-learning with L_1-penalty (Qian and Murphy, 2011):* assume that the Q-function, which is defined as

$$Q_t(h_t, a_t) = E[\max_{a_{t+1} \in \{-1,+1\}} Q_{t+1}(H_{t+1}, a_{t+1}) | H_t = h_t, A_t = a_t],$$

  follows linear regression model

$$Q_t(H_t, A_t) = H_t^T \boldsymbol{\omega}_t + A_t H_t^T \boldsymbol{\theta}_t,$$

  where we again assume that H_t includes an intercept term. Then, it can be shown that

$$\mathcal{D}^*(H_t) = \text{sign}(Q_t(H_t, 1) - Q_t(H_t, -1)).$$

almost surely. The Q-learning with L_1-penalty (Qian and Murphy, 2011) estimates the optimal decision functions and substantiates variable selection through estimating $Q_t(H_t, A_t)$ via standard Lasso regression and backward induction from $t = T$ to $t = 1$.

- *A-learning (Shi et al., 2018):* the A-learning approach assumes that the Q-function satisfies

$$Q_t(H_t, A_t) = H_t^T \boldsymbol{\omega}_t + \frac{1}{2}(1 + A_t) H_t^T \boldsymbol{\theta}_t.$$

  In Shi et al. (2018), $Q_t(H_t, A_t)$ is estimated by solving a Danzig selector optimization problem Candes and Tao (2007), which will impose an L_1-penalty over the estimated coefficients and ensure sparsity over the final estimated rules.

- *Weight least square (Bian et al., 2021):* Bian et al. (2021) proposes to use G-estimation (Robins, 2004) to estimate the optimal DTRs and conduct variable selection by imposing a hierarchical Lasso type penalty to enforce sparsity.

- *L_1-penalty O-learning:* Zhao et al. (2015) proposes to estimate the optimal DTRs by solving a series of weighted support vector machine (SVM) (Cortes and Vapnik, 1995) problem

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_{is} H_{is}^T \widehat{\boldsymbol{\theta}}_s^* > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} \phi(A_{it} H_{it}^T \boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_2^2$$

  via backward induction, where $\phi(\cdot)$ denotes the hinge loss function defined as $\phi(x) = (1 - x)_+$. To incorporate variable selection, we consider the modified L_1-penalty O-learning and estimate the coefficients by solving

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_{is} H_{is}^T \widehat{\boldsymbol{\theta}}_s^* > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} \phi(A_{it} H_{it}^T \boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1$$

  also in backward order. In the final simulation study, we implement the AOWL where all standard weighted SVM optimizations involved are replaced by corresponding L_1-penalized weighted SVM.

For L1-MRL and all competing methods requiring known treatment assignment probability, we use Lasso logistic regression to estimate the treatment assignment probability using all available feature variables as

predictors, which will yield consistent estimators of the true treatment assignment probability because of the choice of true treatment assignment models.

The simulation results are displayed in Table 4.1. For setting I, we first notice that in terms of the reward, L1-MRL and A-learning can attain an average reward of around 9 on testing data under n=200 and 400. In contrast, Q-learning, O-learning and pdwols can only attain a significantly lower average reward of less than 7. This indicates that Q-learning, O-learning and pdwols fail to achieve reward optimization after incorporating variable selection. Compared with A-learning, when the sample size is n=200, L1-MRL tends to yield a lower expected reward on testing data with slightly more selected variables than A-learning and higher variability. When the sample size is increased to n=400, the performance of L1-MRL is improved with higher reward, fewer selected variables around 4.5 for both stages and lower variability, which becomes closer to the performance of A-learning under n=400, but the reward of A-learning is still slightly higher than L1-MRL. As the conclusion, for the first setting and when treatment effects are approximately linear in terms of feature variables, L1-MRL and A-learning tend to have superior performance than other competing methods, with L1-MRL tending to have worse small sample performance and analogous but still slightly worse performance than A-learning. Note that A-learning is a regression-based method and under the first linear setting the decision rule of the second stage is correctly specified. This indicates that L1-MRL has a performance close to a regression-based approach with nearly correctly specified regression models when the sample size is large.

For the second simulation setting, in terms of the reward we still observe that L1-MRL and A-learning can attain an average reward of around 20 which is significantly higher than the reward of Q-learning, O-learning and pdwols, which only have an average reward lower than 13. This is consistent with the first setting that competing methods Q-learning, O-learning and pdwols fail to preserve reward optimization while implementing the variable selection. Compared with A-learning, L1-MRL now tends to attain a higher reward under both n=200 and 400, with L1-MRL selecting around 6 important variables and A-learning selecting around 3 important variables for each stage. Note that in the second setting, reward Y is generated according to a nonlinear function w.r.t. feature variables, so the linear decision function is misspecified. The lower expected reward of A-learning implies that the performance of A-learning will be affected and worsened by the model misspecification, leading to relatively worse performance in terms of maximizing the beneficial reward. In contrast, L1-MRL is designed to maximize (a surrogate function of) the expected reward directly, which can be more robust against model misspecification and maintain a high reward when decision function

Table 4.1: Summary of simulation studies. Results are summarized in mean(sd) format. Reward denotes the estimated expected reward under the estimated rules on the independent testing data. N1 and N2 denote the number of coefficients with absolute values greater than 10^{-6} among 15 candidate variables.

		n=200			n=400		
Setting	Method	Reward	N1	N2	Reward	N1	N2
Setting I	L1-MRL	8.560(0.825)	5.280(1.995)	5.260(2.048)	8.867(0.691)	4.540(0.771)	4.620(1.003)
	A-learning	9.051(0.433)	4.390(0.634)	4.010(0.718)	9.517(0.338)	4.730(0.468)	4.460(0.688)
	O-learning	6.699(0.408)	5.600(1.912)	5.900(1.691)	6.977(0.157)	5.380(2.173)	5.560(2.194)
	Q-learning	5.370(1.271)	2.240(2.767)	8.550(1.982)	5.418(1.391)	2.670(2.785)	8.770(1.932)
	pdwols	5.842(1.385)	13.040(3.752)	11.320(4.746)	6.195(1.388)	14.760(1.590)	14.140(2.458)
Setting II	L1-MRL	22.731(1.555)	7.860(2.503)	7.060(2.215)	24.539(0.850)	6.000(1.544)	5.790(1.465)
	A-learning	19.523(2.007)	2.929(0.746)	2.505(0.825)	20.941(2.082)	3.340(0.807)	2.910(0.922)
	O-learning	11.713(1.218)	5.394(2.208)	5.152(2.447)	12.610(0.612)	5.520(2.052)	5.230(2.457)
	Q-learning	12.920(2.414)	1.667(2.259)	8.616(1.888)	12.760(2.743)	2.250(2.622)	8.590(1.776)
	pdwols	10.931(3.778)	9.515(4.637)	7.475(4.803)	10.831(3.127)	9.430(5.823)	7.330(5.800)

models are incorrect. In terms of variable selection, we notice that A-learning tends to select much fewer variables than L1-MRL. By simulation design, Z_1 is an important variable that correlates with X_1 and Z_2 and will consequently both directly and indirectly influence the cumulative reward. By checking the selected variables when n=400, A-learning is less in favor of selecting Z_1 as an important variable for both two stages, which only has a 27% selecting rate during the first stage and a 16% selecting rate during the second stage, while L1-MRL will select Z_1 as an important variable with more than 90% selection rate for both two stages among 100 repeated analyses. The failure of selecting Z_1 as an important variable and a significantly lower number of selected variables implies that A-learning tends to be overconservative and omits important feature variables under the second simulation setting compared with L1-MRL when the model is misspecified. To sum up, under the second simulation setting L1-MRL tends to have better performance than A-learning and significantly superior performance than the other 3 competing methods. As an overall conclusion, simulation studies suggest that L1-MRL and A-learning have overall better performance than other competing methods, with L1-MRL tending to have slightly worse but comparable performance than A-learning when the simulation setting is close to being linear, while tending to be more robust and have significantly better performance than A-learning when models are misspecified.

4.5 Application to T2D EHR Data

We apply L1-MRL to an observational electronic health record (EHR) data of T2D patients. The raw data consists of EHR data of 55,246 T2D patients collected from the Ohio State University Hospital system between 2008 to 2018. The final cumulative reward Y is set to be

$$Y = -\left(\frac{Y_3 - Y_1}{T_3 - T_1}\right) * 365,$$

which is the cumulative HbA1c reduction at 180 days since the initial of the second stage treatment rescaled to 1 year, so higher Y indicates better treatment performance. The feature variables of the first stage H_1 consist of all 3 time-independent variables and 6 time-dependent variables at T_1, and the second stage feature variables H_2 consist of all 3 time-independent variables and 6 time-dependent variables at T_2 plus the treatment assignment of the first stage and the duration of the first stage treatment.

In this section, we implement L1-MRL and also compare the performance with 4 competing methods in the simulation studies. For each method, we conducted repeated analysis 100 times by sampling 50% of

patients as training data and evaluating the expected reward under the estimated rules using the remaining 50% data as testing data. Since EHR data is observational data, we estimate the treatment assignment probability model via Lasso logistic regression using sampled training data and use the estimated model to calculate the treatment assignment probability for testing data repeatedly. To eliminate the impact of extreme weights, the treatment assignment probability is truncated at 25% and 75% quantile of the estimated treatment assignment probability of training data. For L1-MRL, we impose variable selection for 3 time-independent and 6 time-dependent variables. The implementation of L1-MRL follows the same description as the simulation studies, except that we fix the adaptive coefficients to be the coefficients calculated from the estimated coefficients obtain from AOWL using all available data as training data, without recalculating the adaptive coefficients for each sampled training data. Ohio State University Hospital between 2008 to 2018. Patients' treatment stage and medication received are inferred using patients' medication prescription records where two medication prescriptions that happened less than 90 days are combined and counted into the same treatment stage (which extends the treatment ending time correspondingly). In the screening stage, we require that qualified candidate patients must satisfy the following criteria:

1. Patients must have three identifiable treatment stages since 2008.

2. The duration of the first and second treatment stage must be both longer than 180 days.

3. Patients' inferred treatment of the first stage must be either fast-acting insulin monotherapy or fast-acting insulin plus long-acting insulin combined therapy.

4. Let T_1, T_2 denote the starting point of the first and second stage treatment, then patients must have at least one HbA1c lab test result before T_1, one test result between $(T_1, T_2]$ and one lab test results after T_2.

After the screening stage, 624 patients satisfy the screening requirements. To apply L1-MRL, we generate a two-stage dataset with two arms available at each stage. The treatment arm is defined as follows:

- Stage I:
  - Fast-acting insulin (F) arm: if patient adopted fast-acting insulin monotherapy.
  - Fast-acting + long-acting insulin (FL) arm: if patient adopted fast-acting insulin plus long-acting insulin combined therapy.

- Stage II:
  - Intensified arm (I): for the fast-acting insulin arm, patients are classified into intensified arm if patients kept using fast-acting insulin plus long-acting insulin or at least one type of other types T2D medication; for fast-acting + long-acting insulin arm, patients are classified into intensified arm if patients kept using fast-acting and long-acting plus at least one type of other types T2D medication.
  - Maintained/Reduced arm (MR): for fast-acting insulin arm, patients are classified into maintained/reduced arm if patients maintained fast-acting insulin monotherapy or stopped using fast-acting insulin; for fast-acting plus long-acting insulin arm, patients are classified into maintained/reduced arm if patients maintained fast-acting plus long-acting insulin combined treatment, or stopped using either fast-acting insulin or long-acting insulin.

For each patient, we extract 3 time-independent feature variables including patients' age, gender, and smoking status at the beginning of the first stage. We also extract 6 time-dependent biomarkers including patients' BMI, systolic blood pressure, level of low-density lipoprotein, high-density lipoprotein, triglyceride and HbA1c level measured at the beginning of each treatment stage. For time-dependent variables apart from HbA1c, the value at each time point is approximated via linear interpolation using the most recent test results before and after the time point, and missing values for patients who have insufficient lab tests are imputed by the population mean. Extreme observed values greater than 95% or smaller than 5% quantile are truncated at 95%/5% quantiles before imputation to eliminate the impact of extreme values for 5 time-dependent variables. For HbA1c level, we let $T_3 = T_2 + 180$ and use (Y_1, Y_2, Y_3) to denote the HbA1c level at (T_1, T_2, T_3) respectively. In this analysis, we impute Y_1 using the most recent HbA1c test result before T_1 and similarly impute Y_2 using the most recent HbA1c test result between $(T_1, T_2]$. For Y_3, the HbA1c level is either imputed using the most recent lab test result within $[T_3 - 90, T_3 + 90]$ if any lab test result exists in the time interval or impute the value by fitting a simple linear regression using all lab test results since T_2 (assume that the imputed value at T_2 is the true HbA1c level at T_2) till the end of the second stage treatment if no lab test result exists within $[T_3 - 90, T_3 + 90]$. For patients with Y_3 imputed using linear regression, extreme imputed HbA1c values greater than 14% and below 4% are truncated at 14% and 4 % correspondingly.

The final cumulative reward Y is set to be

$$Y = -\left(\frac{Y_3 - Y_1}{T_3 - T_1}\right) * 365,$$

which is the cumulative HbA1c reduction at 180 days since the initial of the second stage treatment rescaled to 1 year, so higher Y indicates better treatment performance. The feature variables of the first stage H_1 consist of all 3 time-independent variables and 6 time-dependent variables at T_1, and the second stage feature variables H_2 consist of all 3 time-independent variables and 6 time-dependent variables at T_2 plus the treatment assignment of the first stage and the duration of the first stage treatment.

In this section, we implement L1-MRL and also compare the performance with 4 competing methods in the simulation studies. For each method, we conducted repeated analysis 100 times by sampling 50% of patients as training data and evaluating the expected reward under the estimated rules using the remaining 50% data as testing data. Since EHR data is observational data, we estimate the treatment assignment probability model via Lasso logistic regression using sampled training data and use the estimated model to calculate the treatment assignment probability for testing data repeatedly. To eliminate the impact of extreme weights, the treatment assignment probability is truncated at 25% and 75% quantile of the estimated treatment assignment probability of training data. For L1-MRL, we impose variable selection for 3 time-independent and 6 time-dependent variables. The implementation of L1-MRL follows the same description as the simulation studies, except that we fix the adaptive coefficients to be the coefficients calculated from the estimated coefficients obtain from AOWL using all available data as training data, without recalculating the adaptive coefficients for each sampled training data.

The real data analysis results are displayed in Table 4.2. From the table, we first note that compared with 4 competing methods, MRL produces the highest reward on testing data, which indicates that the treatment rules learned by L1-MRL have the best clinical performance. In particular, the expected testing reward under the estimated rules of L1-MRL is significantly higher than all 4 possible one-size-fits-all rules. This suggests that the L1-MRL method can preserve treatment optimization capability and improve the treatment decision via personalizing treatments to patients while conducting variable selection at the same time on real observational data, where competing methods including A-learning fail to achieve comparable reward gain.

In terms of variable selection, L1-MRL tends to select 4.5 important variables for the first stage and roughly the same number of important variables for the second stage, which is higher than A-learning and

Method	Testing Reward (%)	N1	N2
L1-MRL	0.101(0.049)	4.550(2.728)	4.630(2.493)
A-learning	0.091(0.045)	2.410(0.767)	2.480(1.150)
Q-learning	0.079(0.045)	2.190(2.465)	1.790(2.056)
O-learning-L1	0.078(0.046)	1.250(1.666)	4.710(2.630)
pdwols	0.068(0.031)	3.150(2.066)	0.050(0.500)

	FL-I	FL-MR	F-I	F-MR
Reward	0.072(0.037)	0.060(0.041)	0.037(0.037)	-0.013(0.021)

Table 4.2: Summary of the expected testing reward, the number of selected variables under the estimated rules and the expected testing reward under all 4 possible one-size-fits-all rules. Estimation results are reported in the same format as simulation studies. Variables with estimated coefficients with an absolute value greater than 10^{-6} are identified as important variables. Expected rewards are calculated using the stabilized inverse probability estimator defined as the inverse probability estimator divided by the mean of the inverse propensity weight.

Variable	L1-MRL		A-learning		Q-learning		O-learning-L1		pdwols	
	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II
Age	48	52	9	37	28	26	18	74	30	1
Gender	67	74	74	51	27	41	45	1	55	1
Smoking	51	46	21	19	25	11	14	62	23	0
BMI	88	69	59	4	22	4	0	59	12	1
SBP	44	71	3	37	20	28	3	67	15	0
LDL	25	14	9	4	16	4	12	40	26	0
HDL	26	21	7	9	26	6	10	46	23	0
Triglyceride	37	21	10	17	24	13	9	44	31	1
HbA1c	69	95	49	70	31	46	14	78	100	1
Jaccard Index	0.389	0.474	0.370	0.306	0.300	0.312	0.317	0.441	0.422	0.981

Table 4.3: Selected time as important variables for each candidate feature variable and average Jaccard index between selected variables across 100 repeated analyses.

other competing methods. However, by checking the number of times selected as important variables for each candidate variable in Table 4.3, we note that Q-learning, O-learning-L1 and pdwols fail to produce reasonable variable selection results. For Q-learning and O-learning-L1, the decision rule of the first stage learned by Q-learning shows no preference for any variable, while the decision rule learned by O-learning-L1 tends to only prefer gender as an important variable for stage I but discontinue to select the same variable during the second stage, both of which is less meaningful from the clinical perspective. For pdwols, the method only identifies important variables for the first stage and strongly prefers HbA1c, but selects no variable and does not conduct any personalization during the second stage, which is also less meaningful in practice. The unideal variable selection performance of 3 methods is consistent with the result in Table 4.2 where

Q-learning, O-learning-L1 and pdwols can only attain a lower expected reward compared with L1-MRL and A-learning due to worse variable selection performance. For L1-MRl and A-learning, L1-MRl tends to keep gender, BMI, and HbA1c level as important variables for both two stages, while A-learning tends to select only gender as an important variable for both two stages but consider BMI and HbA1c as important only during the first stage. Since the ideal method is expected to yield stable variable selection results when repeatedly implemented, to quantify the stability of variable selection results, we calculate the average pairwise Jaccard index between the sets of selected variables under 100 repeated analyses, where the Jaccard index between two repeated analyses is defined as

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}.$$

Here, S_1 and S_2 denote the set of indices among 9 candidate feature variables selected as important under two repeatedly analyses and we let $J(S_1, S_2) = 1$ if both S_1 and S_2 are both empty set. By definition, a larger Jaccard index closer to 1 indicates higher similarity between the sets of selected variables and more stable variable selection results. By checking the average Jaccard index between L1-MRL and A-learning also reported in Table 4.3, we can notice that L1-MRL and A-learning have similar average Jaccard index during the first stage but the Jaccard index of L1-MRL during the second stage will be significantly higher than A-learning. This suggests that the variable selection of L1-MRL is more stable and consistent across both two stages compared with A-learning under repeated analyses. From the clinical application perspective, HbA1c is known and adopted as one of the most important health biomarkers to make the treatment decision for T2D patients (American Diabetes Association, 2022*b*), and studies have also unveiled that overweight will also affect the response of insulin therapy to T2D patients (Yki-Järvinen et al., 1997). Thus, compared with A-learning, L1-MRL also produces more meaningful variable selection results, which tends to select both HbA1c and BMI as important variables for both two stages and is consistent with clinical guidance and evidence. The better variable selection performance of MRL also explains the result in Table 4.2 that L1-MRL can lead to higher reward gain by producing more stable variable selection and including known important biomarkers to tailor treatments to patients, where the lack of sufficient variable selection of A-learning matches the observation in simulation studies that A-learning can be overconservative under certain scenarios. To sum up, the real data example shows that L1-MRL remains to have overall the best performance compared

with the other 4 competing methods when applied to real observational data, with both higher reward, more stable and clinically meaningful variable selection results.

4.6 Discussion

Incorporating variable selection in learning optimal DTRs has drawn increasing attention in recent years driven by the clinical demand for treating chronic diseases. To respond to the challenge, in this chapter, we propose a new machine learning-based approach, namely L1-MRL, to estimate the optimal decision rules and identify important variables that contribute to treatment optimization at the same time. As one of the main distinctness, due to the simultaneous property of the MRL, the proposed framework is able to estimate the decision rules and conduct variable selection contingently across all stages, where existing methods, up to our best knowledge, can only estimate the decision rules and conduct variable selection stage by stage separately. The simulation studies and the real data example indicate that L1-MRL has overall better performance than the compared existing DTRs methods with variable selection capability.

Apart from the variable selection, incorporating additional restrictions over tolerable adverse risk or allowable treatment budget and learning the optimal treatment rules under restriction has also been studied in recent years. As one of the possible extensions, L1-MRL can also be extended to substantiate variable selection in learning optimal DTRs under consideration of additional restriction by including corresponding constraints to the optimization problem. However, since the objective of L1-MRL is non-convex, the penalty term needs to be carefully designed to facilitate numerical efficiency while preserving theoretical property to guarantee that the estimation will lead to the optimal sparse rules. Future extensions over L1-MRL or novel new methods are still expected to address the variable selection challenge when additional restrictions over the rules need to be satisfied.

Moreover, it is worth noting that apart from the cross-stage variable selection, because of the simultaneous property of the MRL framework, the MRL framework can also be extended to address other cross-stage restrictions over the decision rules in learning optimal DTRs via adding different penalty terms. One possible extension is to incorporate appropriate penalty terms over the similarity between recommended treatments over time for each patient to impose smoothness restriction over the treatment trajectory at the individual level. Such extension can be used to address the real application when treatment rules need to be designed to avoid frequency switching of treatment in a short period of time to reduce potential risk and medical burden

caused to patients due to treatment change. Further studies of different penalty terms are expected to reflect such as the variation of treatment over time or other cross-stage restrictions over the decision rules to tackle real problems with additional restrictions over decision rules across multiple stages.

4.7 Details of Coordinate Decent DC Algorithm for Solving L1-MRL

For convenience, we let

$$O_i = \frac{1}{n}\frac{Y_i}{\prod_{t=1}^{T} p(A_{it}|H_{it})} \quad \text{or} \quad O_i = \frac{1}{n}\frac{\widehat{Y}_i}{\prod_{t=1}^{T} p(A_{it}|H_{it})}.$$

By subtracting and adding an additional term

$$\sum_{i=1}^{n} O_i^- \min(\psi(A_{i1}H_{i1}^T\boldsymbol{\theta}_1/\eta_n), ..., \psi(A_{iT}H_{iT}^T\boldsymbol{\theta}_n/\eta_n))$$

in (4.4) or (4.5), one can obtain that L1-MRL problem is equivalent to maximizing

$$\begin{aligned}
L(\boldsymbol{\theta}) =& \sum_{i=1}^{n} O_i \min(\psi(A_{i1}H_{i1}^T\boldsymbol{\theta}_1/\eta_n), ..., \psi(A_{iT}H_{iT}^T\boldsymbol{\theta}_n/\eta_n)) \\
&+ \sum_{i=1}^{n} O_i^- \min(\psi(|H_{i1}^T\boldsymbol{\theta}_1|/\eta_n), ..., \psi(|H_{iT}^T\boldsymbol{\theta}_n|/\eta_n)) - \lambda_n \sum_{p=1}^{P}\sum_{t=1}^{T} \frac{|\theta_{t,p}|}{\sqrt{\sum_{s=1}^{T}\widetilde{\theta}_{s,p}^2}} \\
=& \sum_{i=1}^{n} |O_i| \min(A_{i1}H_{i1}^T\boldsymbol{\theta}_1/\eta_n, ..., A_{iT}H_{iT}^T\boldsymbol{\theta}_n/\eta_n, d_i) \\
&- \sum_{i=1}^{n} |O_i| \min(A_{i1}H_{i1}^T\boldsymbol{\theta}_1/\eta_n, ..., A_{iT}H_{iT}^T\boldsymbol{\theta}_n/\eta_n, 1-d_i) \\
&+ \sum_{i=1}^{n} O_i^- \min(|H_{i1}^T\boldsymbol{\theta}_1|/\eta_n, ..., |H_{iT}^T\boldsymbol{\theta}_n|/\eta_n) - \lambda_n \sum_{p=1}^{P}\sum_{t=1}^{T} \frac{|\theta_{t,p}|}{\sqrt{\sum_{s=1}^{T}\widetilde{\theta}_{s,p}^2}},
\end{aligned}$$

where we have abused the notation and used $\theta_{t,p}$ to denote all unknown parameters and assume that the first P coefficient needs to be penalized. In the previous equations, $d_i = \mathbb{I}(O_i \geq 0)$ and to obtain the second equality, we have used the fact that

$$\min(\psi(x_1), ..., \psi(x_T)) = \min(x_1, ..., x_T, 1) - \min(x_1, ..., x_T, 0).$$

By reorganizing the objective function, we can equivalently minimize the new objective function

$$\begin{aligned}
L'(\boldsymbol{\theta}) =& \sum_{i=1}^{n} |O_i| \max(-A_{i1} H_{i1}^T \boldsymbol{\theta}_1/\eta_n, ..., -A_{iT} H_{iT}^T \boldsymbol{\theta}_n/\eta_n, -d_i) \\
& - \sum_{i=1}^{n} O_i^- \min(|H_{i1}^T \boldsymbol{\theta}_1|/\eta_n, ..., |H_{iT}^T \boldsymbol{\theta}_n|/\eta_n) \\
& + \lambda_n \sum_{p=1}^{P} \sum_{t=1}^{T} \frac{|\theta_{t,p}|}{\sqrt{\sum_{s=1}^{T} \widetilde{\theta}_{s,p}^2}} \\
& - \sum_{i=1}^{n} |O_i| \max(-A_{i1} H_{i1}^T \boldsymbol{\theta}_1/\eta_n, ..., -A_{iT} H_{iT}^T \boldsymbol{\theta}_n/\eta_n, -(1-d_i)).
\end{aligned}$$

To obtain an estimated solution for the L1-MRL problem, we consider solving the minimization above via the coordinate decent. For convenience, we let

$$\boldsymbol{\theta}_t^{(k)} = (\theta_{t,1}^{(k)}, ..., \theta_{t,K_t}^{(k)}),$$

$$\boldsymbol{\theta}_{t,-p}^{(k+1,k)} = (\theta_{t,1}^{(k+1)}, ..., \theta_{t,p-1}^{(k+1)}, \theta_{t,p+1}^{(k)}, ..., \theta_{t,K_t}^{(k)}),$$

$$H_{it,-p} = (H_{it,1}, ..., H_{it,p-1}, H_{it,p+1}, ..., H_{it,K_t}),$$

where we use K_t to denote the number of unknown parameters at stage t. Given current parameter vector $(\theta_{1,1}^{(k+1)}, ..., \theta_{t,p-1}^{(k+1)}, \theta_{t,p}^{(k)}, ..., \theta_{T,K_T}^{(k)})$, we update $\theta_{t,p}^{(k)}$ by fixing remaining parameters as constants and solving the optimization problem

$$\theta_{t,p}^{(k+1)} = \arg\min_{\theta} S_1(\theta) - S_2(\theta), \tag{4.7}$$

where

$$\begin{aligned}
S_1(\theta) =& \sum_{i=1}^{n} |O_i| \max(-A_{it}(H_{it,p}\theta + H_{it,-p}^T \boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n, c_{it}) \\
& + \sum_{i=1}^{n} O_i^- g_2((H_{it,p}\theta + H_{it,-p}^T \boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n, c_{it}'') + \gamma_{t,p}\lambda_n \frac{|\theta|}{\sqrt{\sum_{s=1}^{T} \widetilde{\theta}_{s,p}^2}} \\
S_2(\theta) =& \sum_{i=1}^{n} |O_i| \max(-A_{it}(H_{it,p}\theta + H_{it,-p}^T \boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n, c_{it'}) \\
& + \sum_{i=1}^{n} O_i^- g_1((H_{it,p}\theta + H_{it,-p}^T \boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n),
\end{aligned}$$

with $\gamma_{t,p} = 1$ if $\theta_{t,p}$ needs to be penalized and $\gamma_{t,p} = 0$ otherwise,

$$g_1(x) = |x|, \quad g_2(x,c) = \max(-x-c,0) + \max(x-c,0),$$

$$\begin{aligned} c_{it} = \max(&- A_{i1}H_{i1}^T\boldsymbol{\theta}_1^{(k+1)}/\eta_n, ..., -A_{i,t-1}H_{i,t-1}^T\boldsymbol{\theta}_{t-1}^{(k+1)}/\eta_n, \\ &- A_{i,t+1}H_{i,t+1}^T\boldsymbol{\theta}_{t+1}^{(k)}/\eta_n, ..., -A_{iT}H_{iT}^T\boldsymbol{\theta}_T^{(k)}/\eta_n, -d_i), \end{aligned}$$

$$\begin{aligned} c'_{it} = \max(&- A_{i1}H_{i1}^T\boldsymbol{\theta}_1^{(k+1)}/\eta_n, ..., -A_{i,t-1}H_{i,t-1}^T\boldsymbol{\theta}_{t-1}^{(k+1)}/\eta_n, \\ &- A_{i,t+1}H_{i,t+1}^T\boldsymbol{\theta}_{t+1}^{(k)}/\eta_n, ..., -A_{iT}H_{iT}^T\boldsymbol{\theta}_T^{(k)}/\eta_n, -(1-d_i)), \end{aligned}$$

$$c'_{it} = \max(|H_{i1}^T\boldsymbol{\theta}_1^{(k+1)}|/\eta_n, ..., |H_{i,t-1}^T\boldsymbol{\theta}_{t-1}^{(k+1)}|/\eta_n, |H_{i,t+1}^T\boldsymbol{\theta}_{t+1}^{(k)}|/\eta_n, ..., |H_{iT}^T\boldsymbol{\theta}_T^{(k)}|/\eta_n).$$

Note that both S_1 and S_2 are convex function of θ, therefore (4.7) can be solved by applying the DC-algorithm (Tao and An, 1997) where we iteratively solve

$$\theta^{(s+1)} = \min_{\boldsymbol{\theta}} \; S_1(\theta) - \frac{\partial S_2}{\partial \theta}(\theta^{(s)})(\theta - \theta^{(s)}) \tag{4.8}$$

until converging starting from $\theta^{(0)} = \theta_{t,p}^{(k)}$. To further reduce the computational complexity, we approximate the subgradient $\frac{\partial S_2}{\partial \theta}(\theta)$ by

$$\begin{aligned} \frac{\partial \widetilde{S}_2}{\partial \theta}(\theta) = &\sum_{i=1}^n \frac{1}{\eta_n}|O_i|(-A_{it}H_{it,p})\frac{e^{-A_{it}(H_{it,p}\theta + H_{it,-p}^T\boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n}}{e^{-A_{it}(H_{it,p}\theta + H_{it,-p}^T\boldsymbol{\theta}_{t,-p}^{(k+1,k)})/\eta_n} + e^{c'_{it}}} \\ &+ \sum_{i=1}^n \frac{1}{\eta_n}O_i^- H_{it,p}\mathbb{I}(H_{it,p}\theta + H_{it,-p}^T\boldsymbol{\theta}_{t,-p}^{(k+1,k)} > 0) \\ &- \sum_{i=1}^n \frac{1}{\eta_n}O_i^- H_{it,p}\mathbb{I}(H_{it,p}\theta + H_{it,-p}^T\boldsymbol{\theta}_{t,-p}^{(k+1,k)} < 0) \end{aligned}$$

using the smoothing technique from Nesterov (2005). For fixed t and p, the optimization of (4.8) w.r.t. θ becomes a minimization problem w.r.t. a piecewise linear function which can be efficiently solved by calculating the derivatives at each ending point. For each t and p, we update $\theta_{t,p}^{(k)}$ until the DC procedure (4.8) converges, and the coordinate decent algorithm terminates until $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\theta}_1^{(k)}, ..., \boldsymbol{\theta}_T^{(k)}\}$ converges.

4.8 Proof of Theorem 4.1

We complete the proof by first verifying the lemma below:

Lemma 4.1 *For any compact subset $\mathcal{X}_t \subset \mathbb{R}^{d_t}$ and $B > 0$, let $\{X_t\}_{t=1}^T$ be T random vectors defined on spaces $\{\mathcal{X}_t\}_{t=1}^T$. Define*

$$\mathcal{W} = \left\{ \min(\psi(X_1^T\boldsymbol{\theta}_1), ..., \psi(X_T^T\boldsymbol{\theta}_T)) : \|\boldsymbol{\theta}_t\|_\infty \leq B,\ \boldsymbol{\beta}_t \in \mathbb{R}^{d_t},\ t = 1, ..., T \right\},$$

then for any $\delta > 0$, we have

$$P\left(\sup_{w \in \mathcal{W}} |\mathbb{P}_n[w] - E[w]| \geq c\frac{\sqrt{\sum_t d_t \log(BTd_t)}}{\sqrt{n}} + \delta \right) \leq c' e^{-n\delta^2}$$

holds for constants c and c' which do not depend on sample size n.

Proof: Without loss of generality, we assume that $\mathcal{X}_t$ is the unit ball of $\mathbb{R}^{d_t}$ for $t = 1, ..., T$. Let $\mathcal{B}$ denote the center of a $\frac{\epsilon}{d_t T}$ covering of interval $[-B, B]$ under Euclidean distance. For arbitrary $w \in \mathcal{W}$ associated with coefficient $\{\boldsymbol{\theta}_t\}_{t=1}^T$, we can find centers $\boldsymbol{b}_t \in \mathcal{B}^{d_t}$ for $t = 1, ..., T$ such that

$$\|\boldsymbol{b}_t - \boldsymbol{\theta}_t\|_\infty \leq \frac{\epsilon}{d_t T}.$$

Using the fact from Proposition 3.1 that for any $(x_1, .., x_T)$ and $(x'_1, ..., x'_T)$, we have

$$|\min(\psi(x_1), ..., \psi(x_T)) - \min(\psi(x'_1), ..., \psi(x'_T))| \leq \sum_{t=1}^T |x_t - x'_t|,$$

we can obtain that

$$\begin{aligned}
&\sup_{(X_1, ..., X_T) \in \mathcal{X}^T} \left| w - \min(\psi(X_1^T\boldsymbol{\theta}), ..., \psi(X_T^T\boldsymbol{\theta}_T)) \right| \\
\leq &\sum_{t=1}^T d_t \|\boldsymbol{b}_t - \boldsymbol{\theta}_t\|_\infty \\
\leq &\epsilon,
\end{aligned}$$

where to obtain the first inequality we have used the fact that $\psi(x)$ is 1-Lipsitz function. Since

$$\mathcal{N}([-B, B], \epsilon; \|\cdot\|_\infty) \leq \frac{B}{\epsilon},$$

the previous inequality implies that

$$\mathcal{N}(\mathcal{W}, \epsilon; \|\cdot\|_\infty) \leq \prod_{t=1}^{T} \left(\frac{BTd_t}{\epsilon}\right)^{d_t}.$$

Now, we show that the concentration inequality stated holds. Using Theorem 4.10 from Wainwright, we can show that

$$P\bigg(\sup_{w\in\mathcal{W}} |\mathbb{P}_n[w] - E[w]| \geq 2\text{Rad}_n(\mathcal{W}) + \delta\bigg) \geq ce^{-n\delta^2}$$

holds for any $\delta > 0$ where $\text{Rad}_n(\mathcal{W})$ denotes the Rademacher complexity of $\mathcal{W}$ defined as

$$\text{Rad}_n(\mathcal{W}) = \sup_{w\in\mathcal{W}} E_X E_\epsilon \bigg|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i w(X_i)\bigg|, \quad P(\epsilon_i = \pm 1) = 0.5.$$

Here, we abuse the notation and use $w(X_i)$ to denote i.i.d. replication of function w evaluated at X_i. Using Example 5.24 from Wainwright we have

$$E_\epsilon \bigg|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i w(X_i)\bigg| \leq \frac{24}{\sqrt{n}} \int_0^2 \sqrt{\log \mathcal{N}(\mathcal{W}, \epsilon; \|\cdot\|_{\mathbb{P}_n})} d\epsilon,$$

where $\|\cdot\|_{\mathbb{P}_n}$ denotes the empirical L_2-norm defined as

$$\|w_1 - w_2\|_{\mathbb{P}_n}^2 = \frac{1}{n}\sum_{i=1}^{n} (w_1(x_i) - w_2(x_i))^2.$$

Since $\|w_1 - w_2\|_{\mathbb{P}_n} \leq \|w_1 - w_2\|_\infty$, we have

$$\mathcal{N}(\mathcal{W}, \epsilon; \|\cdot\|_{\mathbb{P}_n}) \leq \mathcal{N}(\mathcal{W}, \epsilon; \|\cdot\|_\infty)$$

and consequently

$$E_\epsilon \bigg|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i w(X_i)\bigg| \leq c\frac{\sqrt{\sum_t d_t \log BTd_t}}{\sqrt{n}}.$$

The inequality is verified by taking expectation w.r.t. X. □

We now complete the proof of Theorem 4.1 using Lemma 4.1.

Proof: For convenience, we define

$$\mathcal{L}_1(\boldsymbol{\theta}) = -E\left[Y\frac{\min(\psi(A_1H_1^T\boldsymbol{\theta}_1/\eta),...,\psi(A_TH_T^T\boldsymbol{\theta}_T/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right],$$

$$\mathcal{L}_2(\boldsymbol{\theta}) = -E\left[Y^-\frac{\min(\psi(|H_1^T\boldsymbol{\theta}_1|/\eta),...,\psi(|H_T^T\boldsymbol{\theta}_T|/\eta))}{\prod_{t=1}^T p(A_t|H_t)}\right]$$

and use $\mathcal{Q}_1(\boldsymbol{\theta})$ and $\mathcal{Q}_2(\boldsymbol{\theta})$ to denote the empirical version of $\mathcal{L}_1(\boldsymbol{\theta})$ and $\mathcal{L}_2(\boldsymbol{\theta})$ respectively. Hence, we have

$$\mathcal{Q}(\boldsymbol{\theta}) = \mathcal{Q}_1(\boldsymbol{\theta}) + \mathcal{Q}_2(\boldsymbol{\theta}) + \sum_{t,p}\lambda_{np}|\beta_{tp}|$$

Let $\alpha_n = a_n + n^{-\frac{1}{4}+\frac{\eta}{2}}$ where $a_n = \max\{\lambda_{np} : |\beta_{tp}^*| \neq 0\}$, our goal is to show that

$$Q(\boldsymbol{\theta} * + \alpha_n\boldsymbol{\delta}) > Q(\boldsymbol{\theta}^*)$$

holds for any $\|\boldsymbol{\delta}\|_2 = C$ for some sufficiently large C that does not depend on sample size n where $\boldsymbol{\delta}$ is a vector in $\mathbb{R}^{\sum_{t=1}^T d_t}$.

To show this, we first note that for any $\xi \in (0, \frac{1}{2})$ we have

$$\begin{aligned} Q(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - Q(\boldsymbol{\theta}*) \geq & -|\mathcal{L}_1(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - \mathcal{Q}_1(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta})| - |\mathcal{L}_2(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - \mathcal{Q}_2(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta})| \\ & - |\mathcal{L}_1(\boldsymbol{\theta}^*) - \mathcal{Q}_1(\boldsymbol{\theta}^*)| - |\mathcal{L}_2(\boldsymbol{\theta}^*) - \mathcal{Q}_2(\boldsymbol{\theta}^*)| \\ & \underbrace{+ \mathcal{L}(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - \mathcal{L}(\boldsymbol{\theta}^*) + \sum_{t,p}\lambda_{np}|\beta_{tp} + \alpha_n\delta_{tp}| - \sum_{t,p}\lambda_{np}|\beta_{tp}|}_{I}. \end{aligned}$$

By doing Taylor expansion of $\mathcal{L}(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta})$ at $\boldsymbol{\theta}^*$ and use Assumption 4.5, the term I on the left-hand side of the inequality above is lower bounded by

$$I \geq \frac{1}{2}\eta^{-2}\alpha_n^2\boldsymbol{\delta}^T\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)\boldsymbol{\delta} + o_p(1) - s\eta^{-3}Ca_n\alpha_n, \tag{4.9}$$

for some constant s that does not depend on sample size n and constant C.

Since it is assumed that $a_n n^{\frac{1}{4}-\frac{\xi}{2}} = O(1)$ and $\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)$ is positive definite, we can choose sufficiently large C such that $\eta^{-2}\alpha_n^2\boldsymbol{\delta}^T\nabla^2\mathcal{L}(\beta^*)\boldsymbol{\delta}$ dominates $s\eta^{-3}Ca_n\alpha_n$ for any $\|\boldsymbol{\delta}\|_2 = C$. Hence, term I has order $\alpha_n^2 = n^{-\frac{1}{2}+\xi}$ up to a positive constant that does not depend on n. On the other hand, fixing C to be sufficiently large, then Assumption 4.4 and Lemma 4.1 imply that

$$|\mathcal{L}_1(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - \mathcal{Q}_1(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta})| = O_p(n^{-\frac{1}{2}}), \quad |\mathcal{L}_2(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - \mathcal{Q}_2(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta})| = O_p(n^{-\frac{1}{2}}),$$

$$|\mathcal{L}_1(\boldsymbol{\theta}^*) - \mathcal{Q}_1(\boldsymbol{\theta}^*)| = O_p(n^{-\frac{1}{2}}), \quad |\mathcal{L}_2(\boldsymbol{\theta}^*) - \mathcal{Q}_2(\boldsymbol{\theta}^*)| = O_p(n^{-\frac{1}{2}}).$$

Combine with (4.9), we can obtain that

$$Q(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) - Q(\boldsymbol{\theta}^*) > O_p(n^{-\frac{1}{2}}) + I = I(1 + +o_p(1)) > 0$$

for a sufficiently large n.

The discussion above indicates that for any $\epsilon > 0$

$$P\bigg(\inf_{\|\boldsymbol{\delta}\|_2=C} Q(\boldsymbol{\theta}^* + \alpha_n\boldsymbol{\delta}) \geq Q(\boldsymbol{\theta}^*)\bigg) \geq 1 - \epsilon,$$

for sufficient large n, which further indicates that there always exists a local minimum $\widehat{\boldsymbol{\theta}}_n$ such that $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 = O_p(n^{-\frac{1}{4}+\frac{\xi}{2}} + a_n)$. This completes the proof of the theorem □

CHAPTER 5: EXTENSIONS AND FUTURE WORK

As discussed in previous chapters, in real application especially for treating chronic diseases, aggressive treatment may induce multiple negative impacts on patients including but not limited to adverse medical risks, excessive medical cost or deteriorated life quality. Though our proposed methods in this work can be easily refined to accommodate problems with multiple constraints, it lacks theoretical justification to ensure that the choice of surrogate functions used in this work can still lead to consistent estimated rules under mild conditions. Moreover, due to the non-convexity of the surrogate functions adopted, finding a valid initial point and efficiently solving the optimization problem can be hard when the number of constraints is more than one. Hence, future extensions and studies are still expected to improve or develop new efficient methods to tackle the multiple constraints problems in learning optimal DTRs. Also, to our best knowledge, no existing method has ever been proposed to address the variable selection challenge when constraints are imposed in learning optimal DTRs. Future research is also expected to provide statistical tools with variable selection capability under restrictions.

For our work, we focus on handling the problem when the number of available treatments is finite and can be encoded as discrete random variables. For some applications such as drug dosage optimization, the treatment intervention will be a continuous variable and can no longer be fitted into the framework studied in this work. To handle the optimal DTRs problem with continuous treatments, when no additional constraints are considered, a flourishing number of methods have been proposed to learn the optimal rules, the works including Thall, Nguyen and Estey (2008); Chen, Zeng and Kosorok (2016); Li et al. (2020); Zhu et al. (2020); Zhou, Zhu and Zeng (2021); Chen, Li and Yu (2022); Ding, Li and Song (2022); Park, Chen and Yu (2023). In opposed, only a few methods have also been proposed to tackle the optimal DTRs problem with both continuous treatments and additional constraint. These methods include Bayesian approaches such as Thall and Cook (2004); Thall (2012); Lee et al. (2015) and policy learning approach Laber et al. (2018). However, the implementation of the methods mentioned requires strong model assumptions and are lack theoretical justification to guarantee that the learned rules are nearly optimal. New methods with solid theoretical justification and numerical efficiency are still expected for optimal DTRs problems with

continuous treatment under consideration of additional constraints. Also, throughout this work, we have always assumed that the three standard causal assumptions - SUTV, NUC and positivity assumption - hold for the problem studied. However, such assumptions can be violated in real applications, particularly for observational data. As listed at the end of Section 1.2.2, a series of works have been done to tackle the policy learning problem when one of three standard causal assumptions is violated without consideration of additional constraint over the optimal policy. Hence, future work can focus on developing new methods to learn the optimal DTRs under consideration of additional constraints when either one or multiple standard causal assumptions are violated.

Lastly, our work focuses on the problem when treatments are imposed at a fixed and finite number of time points. When the time horizon is closed to be infinite, a number of studies have been completed to learn the optimal treatment rules with infinite decision points (Luckett et al., 2020; Hu et al., 2021; Liao, Klasnja and Murphy, 2021; Shi et al., 2022; Zhou, Zhu and Qu, 2022; Gao, Shi and Song, 2023), and the methods are particularly useful when data is collected from mobile health devices. More recently, studies also explore the problem when the treatment intervention time can be adjusted based on the patient's health condition (Xu et al., 2016; Nahum-Shani et al., 2018; Nie, Brunskill and Wager, 2021; Hua et al., 2021; Chen et al., 2022) and recent research has also started developing new methods when medical surveillance time/method can also be optimized for patients at personal level stimulated by the concept of precision surveillance. With the rapid growth of the concept of precision medicine, how to unify all concerns in precision medicine and personalize the optimal treatment regimens for every patient still remains an open question. From a broad view, novel methods are expected to substantiate treatment personalization with finite/infinite treatment stages, discrete/continuous interventions, unevenly spread predictors, violation of standard causal assumptions, possible multiple constraints, and capability of variable selection particular for the case when patient's gene or long-term longitudinal health information data is available.

BIBLIOGRAPHY

Almirall, Daniel, Beth Ann Griffin, Daniel F. McCaffrey, Rajeev Ramchand, Robert A. Yuen and Susan A. Murphy. 2014. "Time-varying Effect Moderation Using the Structural Nested Mean Model: Estimation Using Inverse-weighted Regression with Residuals." *Statistics in Medicine* 33(20):3466–3487.

Almirall, Daniel, Thomas Ten Have and Susan A. Murphy. 2010. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation." *Biometrics* 66(1):131–139.

American Diabetes Association. 2022*a*. "Glycemic Targets: *Standards of Medical Care in Diabetes—2022*." *Diabetes Care* 45(Supplement_1):S83–S96.

American Diabetes Association. 2022*b*. "Pharmacologic Approaches to Glycemic Treatment: *Standards of Medical Care in Diabetes—2022*." *Diabetes Care* 44(Supplement 1):S111–S124.

Apovian, Caroline M., Jennifer Okemah and Patrick M. O'Neil. 2019. "Body Weight Considerations in the Management of Type 2 Diabetes." *Advances in Therapy* 36(1):44–58.

Athey, Susan and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica* 89(1):133–161.

Badanidiyuru, Ashwinkumar, Robert Kleinberg and Aleksandrs Slivkins. 2018. "Bandits with Knapsacks." *Journal of the ACM* 65(3):1–55.

Bellman, Richard. 1966. "Dynamic Programming." *Science* 153(3731):34–37.

Bennett, Andrew and Nathan Kallus. 2019. "Policy Evaluation with Latent Confounders via Optimal Balance.". arXiv preprint arXiv:1908.01920.

Bhatnagar, Shalabh and K. Lakshmanan. 2012. "An Online Actor–Critic Algorithm with Function Approximation for Constrained Markov Decision Processes." *Journal of Optimization Theory and Applications* 153(3):688–708.

Bian, Zeyu, Erica E. M. Moodie, Susan M. Shortreed and Sahir Bhatnagar. 2021. "Variable Selection in Regression-based Estimation of Dynamic Treatment Regimes." *Biometrics* p. biom.13608.

Blatt, Doron, Susan Allbritton Murphy and Ji Zhu. 2004. A-Learning for Approximate Planning. Technical Report 04-63 The Methology Center, Pennsylvania State University.

Bozdogan, Hamparsum. 1987. "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions." *Psychometrika* 52(3):345–370.

Buchman, Alan L. 2001. "Side Effects of Corticosteroid Therapy." *Journal of Clinical Gastroenterology* 33(4).

Butler, Emily L., Eric B. Laber, Sonia M. Davis and Michael R. Kosorok. 2018. "Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules: Incorporating Patient Preferences for Optimal Treatment." *Biometrics* 74(1):18–26.

Candes, Emmanuel and Terence Tao. 2007. "The Dantzig Selector: Statistical Estimation when p is Much Larger than n." *The Annals of Statistics* 35(6).

Cao, Xuanyu, Junshan Zhang and H. Vincent Poor. 2021. "Constrained Online Convex Optimization With Feedback Delays." *IEEE Transactions on Automatic Control* 66(11):5049–5064.

Cayci, Semih, Atilla Eryilmaz and Rayadurgam Srikant. 2020. Budget-Constrained Bandits over General Cost and Reward Distributions. PMLR pp. 4388–4398.

Chakraborty, Bibhas and Erica E.M. Moodie. 2013. *Statistical Methods for Dynamic Treatment Regimes*. Statistics for Biology and Health New York, NY: Springer New York.

Chen, Guanhua, Donglin Zeng and Michael R. Kosorok. 2016. "Personalized Dose Finding Using Outcome Weighted Learning." *Journal of the American Statistical Association* 111(516):1509–1521.

Chen, Guanhua, Xiaomao Li and Menggang Yu. 2022. Policy Learning for Optimal Individualized Dose Intervals. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ed. Gustau Camps-Valls, Francisco J. R. Ruiz and Isabel Valera. Vol. 151 Proceedings of Machine Learning Research: PMLR pp. 1671–1693.

Chen, Shuxiao and Bo Zhang. 2021. "Estimating and Improving Dynamic Treatment Regimes With a Time-Varying Instrumental Variable.". arXiv preprint arXiv:2104.07822.

Chen, Xin, Rui Song, Jiajia Zhang, Swann Arp Adams, Liuquan Sun and Wenbin Lu. 2022. "On estimating Optimal Regime for Treatment Initiation Time Based on Restricted Mean Residual Lifetime." *Biometrics* 78(4):1377–1389.

Chen, Yuan, Donglin Zeng and Yuanjia Wang. 2021. "Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes." *Journal of the American Statistical Association* 116(533):269–282.

Chow, Yinlam, Mohammad Ghavamzadeh, Lucas Janson and Marco Pavone. 2017. "Risk-Constrained Reinforcement Learning with Percentile Risk Criteria." *The Journal of Machine Learning Research* 18(1):6070–6120. Publisher: JMLR. org.

Chung, Wendy K., Karel Erion, Jose C. Florez, Andrew T. Hattersley, Marie-France Hivert, Christine G. Lee, Mark I. McCarthy, John J. Nolan, Jill M. Norris, Ewan R. Pearson, Louis Philipson, Allison T. McElvaine, William T. Cefalu, Stephen S. Rich and Paul W. Franks. 2020. "Precision Medicine in Diabetes: A Consensus Report From the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)." *Diabetes Care* 43(7):1617–1635.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector Networks." *Machine Learning* 20(3):273–297.

Cox, D. R. 1992. *Planning of Experiments*. Wiley classics library wiley classics library ed. New York: Wiley.

Cryer, P. E., S. N. Davis and H. Shamoon. 2003. "Hypoglycemia in Diabetes." *Diabetes Care* 26(6):1902–1912.

Cui, Yifan and Eric Tchetgen Tchetgen. 2021. "On a Necessary and Sufficient Identification Condition of Optimal Treatment Regimes with an Instrumental Variable." *Statistics & Probability Letters* 178:109180.

Ding, Dongsheng, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang and Mihailo Jovanovic. 2021. Provably Efficient Safe Exploration via Primal-Dual Policy Optimization. PMLR pp. 3304–3312.

Ding, Yi, Yingying Li and Rui Song. 2022. "Statistical Learning for Individualized Asset Allocation." *Journal of the American Statistical Association* pp. 1–11.

Dudik, Miroslav, John Langford and Lihong Li. 2011. "Doubly Robust Policy Evaluation and Learning.". arXiv preprint arXiv:1103.4601.

Ernst, Damien, Pierre Geurts and Louis Wehenkel. 2005. "Tree-Based Batch Mode Reinforcement Learning." *Journal of Machine Learning Research* 6(18):503–556.

Ertefaie, Ashkan, James R. McKay, David Oslin and Robert L. Strawderman. 2021. "Robust Q-Learning." *Journal of the American Statistical Association* 116(533):368–381.

Fahrbach, Jessie, Scott Jacober, Honghua Jiang and Sherry Martin. 2008. "The DURABLE Trial Study Design: Comparing the Safety, Efficacy, and Durability of Insulin Glargine to Insulin Lispro Mix 75/25 Added to Oral Antihyperglycemic Agents in Patients with Type 2 Diabetes." *Journal of Diabetes Science and Technology* 2(5):831–838.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96(456):1348–1360.

Fang, Ethan X., Zhaoran Wang and Lan Wang. 2022. "Fairness-Oriented Learning for Optimal Individualized Treatment Rules." *Journal of the American Statistical Association* pp. 1–14.

Fu, Zuyue, Zhengling Qi, Zhaoran Wang, Zhuoran Yang, Yanxun Xu and Michael R. Kosorok. 2022. "Offline Reinforcement Learning with Instrumental Variables in Confounded Markov Decision Processes.". arXiv:2209.08666 [cs, stat].

Gao, Yuhe, Chengchun Shi and Rui Song. 2023. "Deep Spectral Q-learning with Application to Mobile Health.". arXiv:2301.00927 [cs, stat].

Ghosh, Palash, Trikay Nalamada, Shruti Agarwal, Maria Jahja and Bibhas Chakraborty. 2022. "A Penalized Shared-parameter Algorithm for Estimating Optimal Dynamic Treatment Regimens.". arXiv:2107.07875 [cs, stat].

Gunter, Lacey, Ji Zhu and Susan Murphy. 2011. "Variable Selection for Qualitative Interactions in Personalized Medicine While Controlling the Family-Wise Error Rate." *Journal of Biopharmaceutical Statistics* 21(6):1063–1078.

Guo, Jeff J., Swapnil Pandey, John Doyle, Boyang Bian, Yvonne Lis and Dennis W. Raisch. 2010. "A Review of Quantitative Risk–Benefit Methodologies for Assessing Drug Safety and Efficacy—Report of the ISPOR Risk–Benefit Management Working Group." *Value in Health* 13(5):657–666.

Guo, Wenchuan, Xiao-Hua Zhou and Shujie Ma. 2021. "Estimation of Optimal Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve With High-Dimensional Covariates." *Journal of the American Statistical Association* 116(533):309–321.

Hastie, Trevor, Robert Tibshirani and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics 2nd ed ed. New York, NY: Springer.

He, Yizeng, Soyoung Kim, Mi-Ok Kim, Wael Saber and Kwang Woo Ahn. 2021. "Optimal Treatment Regimes for Competing Risk Data Using Doubly Robust Outcome Weighted Learning with Bi-level Variable Selection." *Computational Statistics & Data Analysis* 158:107167.

Henderson, Robin, Phil Ansell and Deyadeen Alshibani. 2010. "Regret-Regression for Optimal Dynamic Treatment Regimes." *Biometrics* 66(4):1192–1201.

Houede, Nadine, Peter F. Thall, Hoang Nguyen, Xavier Paoletti and Andrew Kramar. 2010. "Utility-Based Optimization of Combination Therapy Using Ordinal Toxicity and Efficacy in Phase I/II Trials." *Biometrics* 66(2):532–540.

Hu, Xinyu, Min Qian, Bin Cheng and Ying Kuen Cheung. 2021. "Personalized Policy Learning Using Longitudinal Mobile Health Data." *Journal of the American Statistical Association* 116(533):410–420.

Hua, William, Hongyuan Mei, Sarah Zohar, Magali Giral and Yanxun Xu. 2021. "Personalized Dynamic Treatment Regimes in Continuous Time: A Bayesian Approach for Optimizing Clinical Decisions with Timing." *Bayesian Analysis* .

Huang, Xiaolin, Lei Shi and Johan AK Suykens. 2014. "Ramp Loss Linear Programming Support Vector Machine." *The Journal of Machine Learning Research* 15(1):2185–2211.

Huang, Xuelin, Jing Ning and Abdus S. Wahed. 2014. "Optimization of Individualized Dynamic Treatment Regimes for Recurrent Diseases." *Statistics in Medicine* 33(14):2363–2378.

Huang, Ying. 2015. "Identifying optimal biomarker combinations for treatment selection through randomized controlled trials." *Clinical Trials* 12(4):348–356.

Huang, Ying and Youyi Fong. 2014. "Identifying Optimal Biomarker Combinations for Treatment Selection via a Robust Kernel Method." *Biometrics* 70(4):891–901.

Illenberger, Nicholas, Andrew J. Spieker and Nandita Mitra. 2021. "Identifying Optimally Cost-effective Dynamic Treatment Regimes with a Q-learning Approach.". arXiv preprint arXiv:2107.03441.

Jiang, Binyan, Rui Song, Jialiang Li and Donglin Zeng. 2020. "Entropy Learning for Dynamic Treatment Regimes." *Statistica Sinica* .

Jiang, Cong, Michael P. Wallace and Mary E. Thompson. 2022. "Dynamic Treatment Regimes with Interference." *Canadian Journal of Statistics* p. cjs.11702.

Jiang, Nan and Lihong Li. 2016. "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning.". arXiv preprint arXiv:1511.03722.

Kallus, Nathan. 2017. Recursive Partitioning for Personalization using Observational Data. In *Proceedings of the 34th International Conference on Machine Learning*, ed. Doina Precup and Yee Whye Teh. Vol. 70 of *Proceedings of Machine Learning Research* PMLR pp. 1789–1798.

Kallus, Nathan and Angela Zhou. 2019. "Confounding-Robust Policy Improvement.". arXiv preprint arXiv:1805.08593.

Kosorok, Michael R. and Eric B. Laber. 2019. "Precision Medicine." *Annual Review of Statistics and Its Application* 6(1):263–286.

Kosorok, Michael R. and Erica E. M. Moodie, eds. 2015. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Krzyszczyk, Paulina, Alison Acevedo, Erika J. Davidoff, Lauren M. Timmins, Ileana Marrero-Berrios, Misaal Patel, Corina White, Christopher Lowe, Joseph J. Sherba, Clara Hartmanshenn, Kate M. O'Neill, Max L. Balter, Zachary R. Fritz, Ioannis P. Androulakis, Rene S. Schloss and Martin L. Yarmush. 2018. "The Growing Role of Precision and Personalized Medicine for Cancer Treatment." *TECHNOLOGY* 06(03n04):79–100.

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

Laber, E. B. and Y. Zhao. 2015. "Tree-based Methods for Individualized Treatment Regimes." *Biometrika* 102(3):501–514.

Laber, Eric B., Daniel J. Lizotte and Bradley Ferguson. 2014. "Set-Valued Dynamic Treatment Regimes for Competing Outcomes." *Biometrics* 70(1):53–61.

Laber, Eric B., Daniel J. Lizotte, Min Qian, William E. Pelham and Susan A. Murphy. 2014. "Dynamic treatment regimes: Technical challenges and applications." *Electronic Journal of Statistics* 8(1).

Laber, Eric B., Fan Wu, Catherine Munera, Ilya Lipkovich, Salvatore Colucci and Steve Ripa. 2018. "Identifying Optimal Dosage Regimes under Safety Constraints: An Application to Long Term Opioid Treatment of Chronic Pain." *Statistics in Medicine* 37(9):1407–1418.

Ledoux, Michel and Michel Talagrand. 1991. *Probability in Banach Spaces*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Lee, Juhee, Peter F. Thall, Yuan Ji and Peter Müller. 2015. "Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity." *Journal of the American Statistical Association* 110(510):711–722.

Li, Fan and Fan Li. 2019. "Propensity Score Weighting for Causal Inference with Multiple Treatments." *The Annals of Applied Statistics* 13(4).

Li, Pin, Jeremy M.G. Taylor, Spring Kong, Shruti Jolly and Matthew J. Schipper. 2020. "A Utility Approach to Individualized Optimal Dose Selection Using Biomarkers." *Biometrical Journal* 62(2):386–397.

Li, Xiang, Shanghong Xie, Donglin Zeng and Yuanjia Wang. 2018. "Efficient L0-norm feature selection based on augmented and penalized minimization." *Statistics in Medicine* 37(3):473–486.

Liang, Muxuan, Young-Geun Choi, Yang Ning, Maureen A. Smith and Yingqi Zhao. 2021. "Estimation and Inference on High-dimensional Individualized Treatment Rule in Observational Data Using Split-and-pooled De-correlated Score.". arXiv preprint arXiv:2007.04445.

Liang, Shuhan, Wenbin Lu and Rui Song. 2018. "Deep Advantage Learning for Optimal Dynamic Treatment Regime." *Statistical Theory and Related Fields* 2(1):80–88.

Liao, Peng, Predrag Klasnja and Susan Murphy. 2021. "Off-Policy Estimation of Long-Term Average Outcomes With Applications to Mobile Health." *Journal of the American Statistical Association* 116(533):382–391.

Liu, Dora, Alexandra Ahmet, Leanne Ward, Preetha Krishnamoorthy, Efrem D Mandelcorn, Richard Leigh, Jacques P Brown, Albert Cohen and Harold Kim. 2013. "A Practical Guide to the Monitoring and Management of the Complications of Systemic Corticosteroid Therapy." *Allergy, Asthma & Clinical Immunology* 9(1):30.

Liu, Ying, Yuanjia Wang, Michael R. Kosorok, Yingqi Zhao and Donglin Zeng. 2018. "Augmented Outcome-weighted Learning for Estimating Optimal Dynamic Treatment Regimens." *Statistics in Medicine* 37(26):3776–3788.

Lizotte, Daniel J., Michael Bowling and Susan A. Murphy. 2012. "Linear Fitted-Q Iteration with Multiple Reward Functions." *Journal of machine learning research: JMLR* 13(Nov):3253–3295.

Longford, Nicholas T. and John A. Nelder. 1999. "Statistics Versus Statistical Science in the Regulatory Process." *Statistics in Medicine* 18(17-18):2311–2320.

Lu, Wenbin, Hao Helen Zhang and Donglin Zeng. 2013. "Variable Selection for Optimal Treatment Decision." *Statistical Methods in Medical Research* 22(5):493–504.

Luckett, Daniel J., Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis and Michael R. Kosorok. 2020. "Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning." *Journal of the American Statistical Association* 115(530):692–706.

Luckett, Daniel J., Eric B. Laber, Siyeon Kim and Michael R. Kosorok. 2021. "Estimation and Optimization of Composite Outcomes." *Journal of machine learning research: JMLR* 22:167.

Mahdavi, Mehrdad, Rong Jin and Tianbao Yang. 2012. "Trading Regret for Efficiency: Online Convex Optimization with Long Term Constraints." *The Journal of Machine Learning Research* 13(1):2503–2528. Publisher: JMLR. org.

McFarlane, Samy I. 2009. "Insulin Therapy and Type 2 Diabetes: Management of Weight Gain." *The Journal of Clinical Hypertension* 11(10):601–607.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg and Demis Hassabis. 2015. "Human-level Control Through Deep Reinforcement Learning." *Nature* 518(7540):529–533.

Moodie, Erica E. M., Nema Dean and Yue Ru Sun. 2014. "Q-Learning: Flexible Learning About Useful Utilities." *Statistics in Biosciences* 6(2):223–243.

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

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

Murphy, Susan A. 2005*b*. "A Generalization Error for Q-Learning." *Journal of machine learning research : JMLR* 6:1073–1097.

Murray, Thomas A., Ying Yuan and Peter F. Thall. 2018. "A Bayesian Machine Learning Approach for Optimizing Dynamic Treatment Regimes." *Journal of the American Statistical Association* 113(523):1255–1267.

Nahum-Shani, Inbal, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari and Susan A Murphy. 2018. "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support." *Annals of Behavioral Medicine* 52(6):446–462.

Nesterov, Yu. 2005. "Smooth Minimization of Non-smooth Functions." *Mathematical Programming* 103(1):127–152.

Nie, Xinkun, Emma Brunskill and Stefan Wager. 2021. "Learning When-to-Treat Policies." *Journal of the American Statistical Association* 116(533):392–409.

Pan, Yinghao, Eric B. Laber, Maureen A. Smith and Yingqi Zhao. 2021. "Reinforced Risk Prediction With Budget Constraint Using Irregularly Measured Data From Electronic Health Records." *Journal of the American Statistical Association* pp. 1–12.

Park, Chan, Guanhua Chen and Menggang Yu. 2023. "Personalized Two-sided Dose Interval.". arXiv:2302.12479 [stat].

Park, Chan Soon, You-Jung Choi, Tae-Min Rhee, Hyun Jung Lee, Hee-Sun Lee, Jun-Bean Park, Yong-Jin Kim, Kyung-Do Han and Hyung-Kwan Kim. 2022. "U-Shaped Associations Between Body Weight Changes and Major Cardiovascular Events in Type 2 Diabetes Mellitus: A Longitudinal Follow-up Study of a Nationwide Cohort of Over 1.5 Million." *Diabetes Care* 45(5):1239–1246.

Qi, Zhengling, Dacheng Liu, Haoda Fu and Yufeng Liu. 2020. "Multi-Armed Angle-Based Direct Learning for Estimating Optimal Individualized Treatment Rules With Various Outcomes." *Journal of the American Statistical Association* 115(530):678–691.

Qi, Zhengling and Yufeng Liu. 2018. "D-learning to Estimate Optimal Individual Treatment Rules." *Electronic Journal of Statistics* 12(2).

Qian, Min and Susan A. Murphy. 2011. "Performance Guarantees for Individualized Treatment Rules." *The Annals of Statistics* 39(2):1180–1210.

Qiu, Xin, Donglin Zeng and Yuanjia Wang. 2018. "Estimation and Evaluation of Linear Individualized Treatment Rules to Guarantee Performance." *Biometrics* 74(2):517–528.

Roberts, Kirk, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick and William R. Hersh. 2020. "Overview of the TREC 2020 Precision Medicine Track." *Text REtrieval Conference (TREC)* 1266.

Robins, James M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality*, ed. Maia Berkane. New York, NY: Springer New York pp. 69–117.

Robins, James M. 2004. Optimal Structural Nested Models for Optimal Sequential Decisions. Springer pp. 189–326.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors are not Always Observed." *Journal of the American Statistical Association* 89(427):846–866.

Rose, Eric J., Erica E. M. Moodie and Susan Shortreed. 2022. "Monte Carlo Sensitivity Analysis for Unmeasured Confounding in Dynamic Treatment Regimes.". arXiv preprint arXiv:2202.09448.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75(371):591.

Saghafian, Soroush. 2022. "Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach.". arXiv preprint arXiv:2112.04571.

Saha, Sarama and Peter E. H. Schwarz. 2017. "Impact of Glycated Hemoglobin (HbA1c) on Identifying Insulin Resistance among Apparently Healthy Individuals." *Journal of Public Health* 25(5):505–512.

Shah, Kushal S., Haoda Fu and Michael R. Kosorok. 2021. "Stabilized Direct Learning for Efficient Estimation of Individualized Treatment Rules.". arXiv preprint arXiv:2112.03981.

Shi, Chengchun, Ailin Fan, Rui Song and Wenbin Lu. 2018. "High-dimensional A-learning for Optimal Dynamic Treatment Regimes." *The Annals of Statistics* 46(3).

Shi, Chengchun, Shikai Luo, Yuan Le, Hongtu Zhu and Rui Song. 2022. "Statistically Efficient Advantage Learning for Offline Reinforcement Learning in Infinite Horizons." *Journal of the American Statistical Association* pp. 1–14.

Song, Rui, Michael Kosorok, Donglin Zeng, Yingqi Zhao, Eric Laber and Ming Yuan. 2015. "On sparse representation for optimal Individualized Treatment Selection with Penalized Outcome Weighted Learning." *Stat* 4(1):59–68.

Song, Rui, Weiwei Wang, Donglin Zeng and Michael R. Kosorok. 2015. "Penalized Q-Learning for Dynamic Treatment Regimens." *Statistica Sinica* pp. 901–920.

Steinwart, I., D. Hush and C. Scovel. 2006. "An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels." *IEEE Transactions on Information Theory* 52(10):4635–4643.

Steinwart, Ingo and Clint Scovel. 2007. "Fast Rates for Support Vector Machines Using Gaussian Kernels." *The Annals of Statistics* 35(2):575–607.

Stellato, Bartolomeo, Goran Banjac, Paul Goulart, Alberto Bemporad and Stephen Boyd. 2020. "OSQP: An Operator Splitting Solver for Quadratic Programs." *Mathematical Programming Computation* 12(4):637–672.

Sutton, Richard S. and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. Adaptive computation and machine learning Cambridge, Mass: MIT Press.

Sørensen, Thorkild IA. 1996. "Which Patients May be Harmed by Good Treatments?" *The Lancet* 348(9024):351–352.

Tao, Pham Dinh and Le Thi Hoai An. 1997. "Convex Analysis Approach to DC Programming: Theory, Algorithms and Applications." *Acta mathematica vietnamica* 22(1):289–355.

Tao, Yebin, Lu Wang and Daniel Almirall. 2018. "Tree-based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes." *The Annals of Applied Statistics* 12(3):1914–1938.

Thall, Peter F. 2012. "Bayesian Adaptive Dose-finding Based on Efficacy and Toxicity." *Journal of Statistical Research* 46(2):187–202.

Thall, Peter F., Hoang Q. Nguyen and Elihu H. Estey. 2008. "Patient-Specific Dose Finding Based on Bivariate Outcomes and Covariates." *Biometrics* 64(4):1126–1136.

Thall, Peter F. and John D. Cook. 2004. "Dose-Finding Based on Efficacy-Toxicity Trade-Offs." *Biometrics* 60(3):684–693.

UKPDS Group. 1998. "Intensive Blood-glucose Control with Sulphonylureas or Insulin Compared With Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes (UKPDS 33)." *The Lancet* 352(9131):837–853.

van der Laan, Mark J. and Maya L Petersen. 2007. "Causal Effect Models for Realistic Individualized Treatment and Intention to Treat Rules." *The International Journal of Biostatistics* 3(1).

van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer Series in Statistics New York, NY: Springer New York.

Wainwright, Martin J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. 1 ed. Cambridge University Press.

Wallace, Michael P. and Erica E. M. Moodie. 2015. "Doubly-Robust Dynamic Treatment Regimen Estimation Via Weighted Least Squares." *Biometrics* 71(3):636–644.

Wang, Yanqing, Ying-Qi Zhao and Yingye Zheng. 2020. "Learning-based Biomarker-assisted Rules for Optimized Clinical Benefit under a Risk-constraint." *Biometrics* 76(3):853–862.

Wang, Yuanjia, Haoda Fu and Donglin Zeng. 2018. "Learning Optimal Personalized Treatment Rules in Consideration of Benefit and Risk: With an Application to Treating Type 2 Diabetes Patients With Insulin Therapies." *Journal of the American Statistical Association* 113(521):1–13.

Watkins, Christopher John Cornish Hellaby. 1989. Learning from Delayed Rewards PhD thesis King's College Cambridge UK: .

Wing, Rena R., Wei Lang, Thomas A. Wadden, Monika Safford, William C. Knowler, Alain G. Bertoni, James O. Hill, Frederick L. Brancati, Anne Peters, Lynne Wagenknecht and the Look AHEAD Research Group. 2011. "Benefits of Modest Weight Loss in Improving Cardiovascular Risk Factors in Overweight and Obese Individuals With Type 2 Diabetes." *Diabetes Care* 34(7):1481–1486.

Wu, Yichao, Hao Helen Zhang and Yufeng Liu. 2010. "Robust Model-Free Multiclass Probability Estimation." *Journal of the American Statistical Association* 105(489):424–436.

Xu, Yanxun, Peter Müller, Abdus S. Wahed and Peter F. Thall. 2016. "Bayesian Nonparametric Estimation for Dynamic Treatment Regimes With Sequential Transition Times." *Journal of the American Statistical Association* 111(515):921–950.

Yki-Järvinen, Hannele, Leena Ryysy, Marjut Kauppila, Eila Kujansuu, Jorma Lahti, Tapani Marjanen, Leo Niskanen, Sulo Rajala, Seppo Salo, Pentti Seppälä, Timo Tulokas, Jorma Viikari and Marja-Riitta Taskinen. 1997. "Effect of Obesity on the Response to Insulin Therapy in Noninsulin-Dependent Diabetes Mellitus." *The Journal of Clinical Endocrinology & Metabolism* 82(12):4037–4043.

Yu, Ming, Zhuoran Yang, Mladen Kolar and Zhaoran Wang. 2019. "Convergent Policy Optimization for Safe Reinforcement Learning." *Advances in Neural Information Processing Systems* 32.

Zhang, B., A. A. Tsiatis, E. B. Laber and M. Davidian. 2013. "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions." *Biometrika* 100(3):681–694.

Zhang, Baqun, Anastasios A. Tsiatis, Eric B. Laber and Marie Davidian. 2012. "A Robust Method for Estimating Optimal Treatment Regimes." *Biometrics* 68(4):1010–1018.

Zhang, Baqun, Anastasios A. Tsiatis, Marie Davidian, Min Zhang and Eric Laber. 2012. "Estimating Optimal Treatment Regimes from a Classification Perspective." *Stat* 1(1):103–114.

Zhang, Baqun and Min Zhang. 2018. "C-learning: A New Classification Framework to Estimate Optimal Dynamic Treatment Regimes." *Biometrics* 74(3):891–899.

Zhang, Yichi, Eric B. Laber, Anastasios Tsiatis and Marie Davidian. 2015. "Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes." *Biometrics* 71(4):895–904.

Zhang, Yichi, Eric B. Laber, Marie Davidian and Anastasios A. Tsiatis. 2018. "Interpretable Dynamic Treatment Regimes." *Journal of the American Statistical Association* 113(524):1541–1549.

Zhao, Yingqi, Donglin Zeng, A. John Rush and Michael R. Kosorok. 2012. "Estimating Individualized Treatment Rules Using Outcome Weighted Learning." *Journal of the American Statistical Association* 107(499):1106–1118.

Zhao, Yingqi, Donglin Zeng, Eric B. Laber and Michael R. Kosorok. 2015. "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes." *Journal of the American Statistical Association* 110(510):583–598.

Zhao, Yingqi, Eric B. Laber, Yang Ning, Sumona Saha and Bruce E. Sands. 2019. "Efficient Augmentation and Relaxation Learning for Individualized Treatment Rules Using Observational Data." *Journal of machine learning research: JMLR* 20:48.

Zhao, Yufan, Michael R. Kosorok and Donglin Zeng. 2009. "Reinforcement Learning Design for Cancer Clinical Trials." *Statistics in Medicine* 28(26):3294–3315.

Zhou, Wenzhuo, Ruoqing Zhu and Annie Qu. 2022. "Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pT-Learning." *Journal of the American Statistical Association* pp. 1–14.

Zhou, Wenzhuo, Ruoqing Zhu and Donglin Zeng. 2021. "A parsimonious personalized Dose-finding Model via Dimension Reduction." *Biometrika* 108(3):643–659.

Zhou, Yunzhe, Zhengling Qi, Chengchun Shi and Lexin Li. 2023. "Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach.". arXiv:2210.14420 [cs, stat].

Zhou, Zhengyuan, Susan Athey and Stefan Wager. 2022. "Offline Multi-Action Policy Learning: Generalization and Optimization." *Operations Research* p. opre.2022.2271.

Zhu, Liangyu, Wenbin Lu, Michael R. Kosorok and Rui Song. 2020. Kernel Assisted Learning for Personalized Dose Finding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM pp. 56–65.

Zhu, Ruoqing, Yingqi Zhao, Guanhua Chen, Shuangge Ma and Hongyu Zhao. 2017. "Greedy Outcome Weighted Tree Learning of Optimal Personalized Treatment Rules: Greedy Outcome Weighted Tree." *Biometrics* 73(2):391–400.