LEARNING INDIVIDUALIZED TREATMENT RULES WITH SEQUENTIAL AND MULTI-OUTCOME DATA

Daiqi Gao

A dissertation submitted to the faculty of 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 Statistics and Operations Research.

Chapel Hill
2023

Approved by:

Yufeng Liu

Donglin Zeng

Nilay T. Argon

Yao Li

Quoc Tran-Dinh

ABSTRACT

Daiqi Gao: Learning Individualized Treatment Rules with Sequential and Multi-Outcome Data
(Under the direction of Yufeng Liu and Donglin Zeng)

Learning optimal individualized treatment rules (ITRs) has become increasingly important in the modern era of precision medicine. Many statistical and machine learning methods for learning optimal ITRs have been developed in the literature. In this dissertation, we propose several approaches to solve some important problems regarding the data generating process and the learning algorithm for estimating ITRs.

In the first project, we improve the outcome of interest in a clinical trial using a sequentially rule-adaptive design. Each entering patient will be allocated with a high probability to the current best treatment for this patient, which is estimated using the past data based on machine learning algorithm. We discuss the tradeoff between the training and test performance of the learnt ITR in the framework of contextual bandits. We also develop a tool that combines martingale with empirical process for sequentially generated data to tackle the theoretical problem with dependent data that cannot be solved by existing techniques for i.i.d. data.

In the second project, we focus on the multi-stage stationary treatment policy (MSTP), which prescribes treatment assignment probabilities using the same decision function over stages. We estimate and conduct statistical inference for the parameters of the MSTP in high-dimensional settings. We propose to estimate the MSTP based on a penalized doubly robust estimator of the value function, and construct confidence intervals of the low-dimensional parameters that we are interested in using a one-step estimator. The proposed method allows for a slow convergence rate of the nuisance parameters in the model with a guarantee of the $1/\sqrt{n}$ convergence rate of our interested parameters.

In the third project, we estimate the ITR that maximizes the primary outcome and causes little harm to auxiliary outcomes in the meanwhile. We propose a fusion penalty to encourage

ITRs based on the primary outcome and auxiliary outcomes to yield similar recommendations, and optimize a surrogate loss function for estimation. We derive the non-asymptotic properties for the proposed method and show that the agreement rate between the estimated ITRs for primary and auxiliary outcomes converges faster to the true rate compared to methods without using auxiliary outcomes.

To my parents,

Shiji Gao and Yan Wan,

for instilling in me a passion for life and a curiosity about the world.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my wonderful advisors Dr. Yufeng Liu and Dr. Donglin Zeng. I am extremely fortunate to be able to work with both of them. Dr. Liu has been very supportive throughout my entire Ph.D. life. In addition to providing insightful domain suggestions, he is especially helpful in teaching me soft skills and providing career advice. I have learnt a lot from him about how to thoroughly explore a new research problem and clearly present it. Dr. Zeng is remarkably knowledgeable in various statistics areas and highly efficient in collaborations. I am deeply grateful for his help in forming research ideas and completing technical details. His broad scientific vision, view for research and passion for learning has influenced me greatly.

Next I would like to thank my collaborators. Dr. Yuanjia Wang is kind, patient and supportive. I am grateful for her insightful ideas and comments on the projects that we have worked on together. I have learnt from her the way of thinking from the perspective of real biomedical applications. I would also like to thank Dr. Tony Cai, Dr. Kaibo Wang, Dr. Sheng Yu and Dr. Linjun Zhang, who have worked with me during my undergraduate study. They led me into the statistics field and build a great starting point for me to do rigorous research.

I would also thank my committee members Dr. Nilay Argon, Dr. Yao Li and Dr. Quoc Tran-Dinh for their time and effort in guiding my dissertation. They have provided useful comments and suggestions from their professional viewpoint, which have improved the completeness and clarity of this dissertation.

Finally, I feel incredibly fortunate to have been accompanied by my beloved friends, Qian Cheng, Jiaying Li, Stephanie Lin, Yiyun Luo, Haixu Ma, Weibin Mo, Xinyuan Niu, Taylor Petty, Zhengling Qi, Jose Sanchez Gomez, Hui Shen, Haodong Wang, Mingyi Wang, Peiyao Wang, Siqi Xiang, Wenyi Xie, Yuhan Xie, and Xiaofei Zhou, who have talked with me about interesting academic and life topics, shared the ups and downs of my life, and created colorful memories during my time as a

Ph.D. student. My very special thanks go to my parents Shiji Gao, Yan Wan and my boyfriend Haonan Chen, for their unconditional love, support and encouragement throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

For many diseases, patients respond heterogeneously to treatments and a one-size-for-all strategy is often not effective. Recent technology advances allow personalized treatment strategy by tailoring the treatment to patient characteristics, including demographics, medical histories or genetic information (Hamburg and Collins, 2010). The personalized policy is often referred as the Individualized Treatment Rule (ITR), which aims to maximize a predefined reward such as the patient's health status. In this dissertation, we investigate some machine learning (ML) methods, including reinforcement learning (RL), to learn ITR in both single-stage and multi-stage problems.

In this chapter, we introduce the background and some related literature in the field of personalized medicine and reinforcement learning. In Section 1.1, we introduce the ITR and the literature with various goals in single-stage and multi-stage settings. In Section 1.2, we introduce the basic concepts and algorithms of RL and its recent developments in estimating ITRs, including the single-stage special case of RL, the bandit algorithms. In Section 1.3, we outline the rest of the dissertation and briefly discuss several projects.

1.1 Individualized Treatment Rules

An ITR consists of a sequence of decision rules that determine which treatment to take for a patient at each stage based on the covariates and treatment history. A common goal is to find the optimal rule that can generate the largest mean reward across all stages and the whole population, with potential constraints on the form of the rules, the side effects, etc. For single-stage decision problems, only one decision rule is needed. For multi-stage decision problems, decision rules can be the same or different at every stage, depending on the problem and the practical need.

1.1.1 Single-Stage Decision Problems

In single-stage decision problem, a prognostic variable vector $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ is observed for each patient. Based on the covariates, we need to decide which treatment $A \in \mathcal{A}$ to take for the patient. A reward $R \in \mathbb{R}$ is then observed, usually as a function of the clinical outcome, with higher values desirable. An ITR is a map $\mathcal{D} : \mathcal{X} \mapsto \mathcal{A}$ that assigns the patient of covariates $\boldsymbol{X}$ to a treatment A. If the mean reward is used as the criterion, the optimal ITR

$$\mathcal{D}^* \in \arg\max_{\mathcal{D}:\mathcal{X}\mapsto\mathcal{A}} \mathbb{E}[R|A = \mathcal{D}(\boldsymbol{X})] \tag{1.1}$$

can generate the largest mean reward for the population. The quantity on the right hand side of (1.1) is called the value function of an ITR.

Methods for learning single-stage ITRs can be generally categorized as regression-based or classification-based methods. The Q-function is defined as the conditional expectation of rewards, i.e. $Q(\boldsymbol{x}, a) = E(R|\boldsymbol{X} = \boldsymbol{x}, A = a)$. A regression-based method fits a regression model for the Q-function and finds the best treatment $\mathcal{D}(\boldsymbol{x}) = \arg\max_{a\in\mathcal{A}} Q(\boldsymbol{x}, a)$. Qian and Murphy (2011) proposed to fit a parametric model of the covariates and the treatment against the reward using penalized least squares and find the treatment that maximizes the estimated reward. This method fits into the general framework of Q-learning. An alternative method called A-learning directly models the conditional average treatment effect (CATE), $Q(\boldsymbol{x}, 1) - Q(\boldsymbol{x}, 0)$, in the binary treatment case. Lu et al. (2013); Shi et al. (2016) used A-learning to select important variables in potentially high dimensional settings. Another related topic is subgroup identification, which finds the target patients with enhanced treatment effect (Tian et al., 2014; Chen et al., 2017).

In contrast to regression-based methods, classification-based methods try to find the best ITR by maximizing the average reward without relying on an estimation of the rewards. A common method is to use inverse probability weighted (IPW) estimator to consistently estimate the value function of a rule. A shown in Qian and Murphy (2011), the mean reward under an ITR $\mathcal{D}$ can be expressed as

$$\mathbb{E}[R|A = \mathcal{D}(\boldsymbol{X})] = \mathbb{E}\left[\frac{R\mathbb{1}(A = \mathcal{D}(\boldsymbol{X}))}{\pi(A; \boldsymbol{X})}\right]. \tag{1.2}$$

Zhao et al. (2012) proposed a weighted classification algorithm for binary treatments called outcome weighted learning (OWL), which transforms (1.2) into a convex optimization problem and solves it under the framework of support vector machine (SVM). Residual weighted learning (Zhou et al., 2017) and augmented outcome weighted learning (Liu et al., 2018b) are proposed to reduce the variance by removing the estimated main effect from the rewards. The doubly robust (DR) structure is also utilized in finding ITRs so that we can still get a consistent estimator of the value function when either the propensity score or the Q-function is misspecified (Liu et al., 2018b; Zhao et al., 2019). Other variations of OWL include multicategory outcome-weighted margin-based learning for multi-armed treatments (Zhang et al., 2020a), generalized outcome weighted learning for ordinal treatments (Chen et al., 2018) and personalized dose finding for continuous treatments (Chen et al., 2016).

Zhang et al. (2012b) proposed to use the augmented inverse probability weighted estimator (AIPWE) of the value function, which is also doubly robust, to find the parameters of ITR. However, the objective function is nonsmooth as opposed to the convex function in OWL. Zhang et al. (2012a) proposed C-learning that minimizes the CATE-weighted misclassification error rate. Athey and Wager (2021) maximized the cross-fitting AIPWE to optimize either binary treatments or infinitesimal nudges to continuous treatments using observational data, and established guarantees for the asymptotic utilitarian regret. Some tree-based methods have been proposed for generating interpretable rules which are important in clinical practice (Laber and Zhao, 2015; Zhu et al., 2017). There are also several other methods that considers the ITR problem from a different perspective. For example, angle-based direct learning considers the angle between the ITR of a patient and the treatments for multi-treatment settings (Qi et al., 2020). E-learning deals with misspecified treatment-free effect model and heteroscedasticity by designing the objective function using semiparametric efficient estimate (Mo et al., 2022). Another important topic is the statistical inference for the estimated rule (Song et al., 2017; Liang et al., 2022).

1.1.2 Multi-Stage Decision Problems

Multi-stage ITRs are often referred to as the dynamic treatment regime (DTR) (Murphy et al., 2001; Murphy, 2003) to reflect the possible change of treatment rules through time. Multi-stage decision problems are usually treated differently for finite stages and indefinite/infinite stages.

While problems with indefinite/infinite horizon are usually modeled with the Markov assumption and the treatment rule is homogeneous across stages, in finite-stage problems, each stage can be treated differently. We focus on the finite-stage problems in this section.

When there are T decision stages in total, a trajectory $(\boldsymbol{X}_t, A_t, R_t)_{t=1}^T$ is observed for each patient, where $\boldsymbol{X}_t \in \mathcal{X}_t$ is the time-varying covariates that can depend on covariates and treatments in previous stages, $A_t \in \mathcal{A}_t$ is the treatment, and $R_t \in \mathbb{R}$ is the outcome observed. A DTR is a sequence of maps $\boldsymbol{\mathcal{D}}_{1:T} = (\mathcal{D}_1, \ldots, \mathcal{D}_T)$, where $\mathcal{D}_t : \mathcal{X}_t \mapsto \mathcal{A}_t$ is the treatment rule at stage t. In this case, an outcome is observed at the end of each stage (Murphy, 2005b; Zhao et al., 2015; Liu et al., 2018b; Zhu et al., 2019) and the optimal DTR is defined as

$$\boldsymbol{\mathcal{D}}^*_{1:T} \in \underset{(\mathcal{D}_t:\mathcal{X}_t\mapsto\mathcal{A}_t)_{t=1}^T}{\arg\max} \sum_{t=1}^{T} \mathbb{E}[R_t | A_t = \mathcal{D}_t(\boldsymbol{X}_t)]. \tag{1.3}$$

There are also cases when an outcome is observed at the end of the study (Murphy, 2003; Zhang et al., 2013), so the observed trajectory is $(\boldsymbol{X}_1, A_1, \ldots, \boldsymbol{X}_T, A_T, R)$. The DTR is then defined as

$$\boldsymbol{\mathcal{D}}^*_{1:T} \in \underset{(\mathcal{D}_t:\mathcal{X}_t\mapsto\mathcal{A}_t)_{t=1}^T}{\arg\max} \mathbb{E}[R | A_1 = \mathcal{D}_1(\boldsymbol{X}_1), \ldots, A_T = \mathcal{D}_T(\boldsymbol{X}_T)]. \tag{1.4}$$

The quantity on the right hand side of (1.3) or (1.4) is called the value function of a DTR.

In either (1.3) or (1.4), the value function depends on the DTR at all stages. In order to estimate the value of the sequence of DTR, we need to estimate the effect of the rule at each stage in a backward fashion. Q-learning (Murphy, 2005b) and A-learning (Murphy, 2003) can be generalized to multi-stage settings. The main difference is that the Q-functions should be calculated in a backward fashion to incorporate the future rules and outcomes as delayed results of the current stage. Zhao et al. (2011) proposed an adaptive design for clinical trials based on Q-learning to select optimal compounds of treatments and the optimal time of the second treatments. Shi et al. (2018) proposed penalized A-learning in the high-dimensional setting and established oracle inequalities and error bounds. Sun and Wang (2021) proposed a stochastic tree search method based on the backward estimated mean of counterfactual outcomes.

Classification-based methods can also be generalized to multi-stage problems. Based on AIPWE, Zhang et al. (2013) proposed to directly maximize the value function over a paramet-

ric class. Backward outcome weighted learning and simultaneous outcome weighted learning are extensions of OWL for multi-stage problems, which backward estimate the value function and find the best DTR by minimizing the weighted surrogate loss (Zhao et al., 2015). Augmented outcome weighted learning removes the estimated main effect from the Q-function in a backward fashion to reduce the variance (Liu et al., 2018b). C-learning minimizes the CATE-weighted misclassification error rate at each stage (Zhang and Zhang, 2018a).

Statistical inference of the value of a given DTR provides information about the impact of implementing such a policy. Inference of the parameters of the learned DTR informs us the important variables in guidance of treatment assignment. Shi et al. (2020a) proposed to use subsample aggregating (subagging) to deal with nonregularity problems, when the treatment is neither beneficial nor harmful for a subpopulation. Laber et al. (2014) dealt with nonregularity under the framework of Q-learning and Zhu et al. (2019) discussed the problem under high-dimensional settings. Zhang and Zhang (2018b) considered variable selection for making treatment decisions under the framework of C-learning, by forward sequentially minimizing the weighted misclassification error rate.

1.2 Reinforcement Learning and Applications in ITR

Reinforcement learning is a class of algorithms aiming at making sequential decisions by trial and error. It can improve the performance on training data as the trial develops. Its wide applications include chess playing, robotics, adaptive controller and beyond (Sutton and Barto, 2018). Algorithms on single-stage problems are called multi-armed bandit, and those using context information in addition to environment information are called contextual bandits specifically.

1.2.1 Introduction to Reinforcement Learning

Assume there are T stages in a decision problem. The number of stages or the horizon T can be finite or infinite, random or nonrandom, fixed or different for each subject. Let $\boldsymbol{X}_t \in \mathcal{X}$ be a d-dimensional state vector and $A_t \in \mathcal{A}$ be the action at stage $t = 1, \ldots, T$. The state and action spaces $\mathcal{X}$ and $\mathcal{A}$ can also be extended to time-variant cases, but we discuss the simple case here for illustration. Assume $\boldsymbol{\pi} := \{\pi_1, \ldots, \pi_T\}$ is a policy such that each action A_t is taken following π_t.

Then a trajectory

$$\boldsymbol{D} = \{\boldsymbol{X}_1, A_1, \boldsymbol{X}_2, A_2, \ldots, \boldsymbol{X}_T, A_T, \boldsymbol{X}_{T+1}\}$$

is observed for each subject i. The initial state $\boldsymbol{X}_1$ are usually assumed to be independent and identically distributed across all trajectories. The reward $R_t \in \mathbb{R}$ can be a known or unknown function of the history prior to time t, that is, $\boldsymbol{H}_t = \{\boldsymbol{X}_1, A_1, \ldots, \boldsymbol{X}_t, A_t\}$. The return is defined as the total discounted reward

$$G(\boldsymbol{D}) = \frac{1}{\sum_{t=1}^{T} \gamma^{t-1}} \sum_{t=1}^{T} \gamma^{t-1} R_t,$$

where $\gamma \in (0, 1]$ is a discount factor. When $\gamma = 1$, the value function is actually the average reward across all time points. Then the value function of a policy $\boldsymbol{\pi}_t$ is the average return in the population

$$\mathcal{V}(\boldsymbol{\pi}) = \mathbb{E}\left[\frac{1}{\sum_{t=1}^{T} \gamma^{t-1}} \sum_{t=1}^{T} \gamma^{t-1} R_t \middle| A_t = \pi_t(\boldsymbol{X}_t, \boldsymbol{H}_{t-1}), t = 1, \ldots, T \right]$$

if all the actions follow this policy. The state-action value function at stage t, also called the Q-function, is defined to be

$$Q_t(\boldsymbol{x}_t, a_t) = \mathbb{E}\left[\frac{1}{\sum_{t=1}^{T} \gamma^{t-1}} \sum_{k=t}^{T} \gamma^{k-1} R_k \middle| \boldsymbol{X}_t = \boldsymbol{x}_t, A_t = a_t \right]$$

and the state value function is defined as

$$V_t(\boldsymbol{x}_t) = \mathbb{E}[Q_t(\boldsymbol{x}_t, a_t) | \boldsymbol{X}_t = \boldsymbol{x}_t].$$

These two functions are frequently used in RL literature.

Online algorithms update the policy after each trajectory. For each new stage t for the ith trajectory, we need to decide which action $A_{i,t}$ to take at each stage based on the previous states and actions $\boldsymbol{X}_{i,1}, A_{i,1}, \ldots, \boldsymbol{X}_{i,t-1}, A_{i,t-1}$ in this trajectory and all previous trajectories $\mathcal{D}_1, \ldots, \mathcal{D}_{i-1}$. These problems deal with two types of dependence, the dependence of states and actions within each trajectory, and the dependence between trajectories. On the contrary, batch or offline algorithms do not update until we get a batch of trajectories.

All reinforcement learning problems encapsulate an "exploration-exploitation" dilemma. While we want to follow the current policy given by previous information since they are more trustworthy than random guess, it is also possible that our estimation is not accurate. Consequently, we also want to take our opportunity to explore more action options. There is a trade-off between exploiting the current rule and exploring the true optimal rule.

Various approaches of exploration have been proposed. ϵ-greedy (Watkins, 1989; Yang and Zhu, 2002; Sutton and Barto, 2018) is one of the most widely used exploration methods. It chooses the current optimal policy with probability $1-\epsilon$ and performs pure randomization with probability ϵ. Boltzmann exploration assigns probabilities of whether to follow the current optimal policy using soft-max function (Sutton and Barto, 2018). The randomization probability is tailored to different characteristics. Another category of commonly used exploration method is Upper-Confidence Bound (UCB), which chooses the arm with the largest upper confidence bound (Li et al., 2010; Srinivas et al., 2010; Krause and Ong, 2011). UCB also allows the algorithm to choose the current optimal rule with a higher probability if we have more confidence based on the covariates. In fact, Chu et al. (2011) shows that a variant of LinUCB, SupLinUCB, has an optimal convergence rate up to logarithmic factors. Bayesian methods assign a treatment to a future patient according to the posterior distribution of reward parameters (Chapelle and Li, 2011; Liao et al., 2020). Action elimination is another branch that ignores the inferior arms gradually (Perchet and Rigollet, 2013). A special case of exploration is active learning, which is an efficient approach in estimating the final policy. Generally, active learning is a classification algorithm that can interactively query labels of specific data points.

The literature in RL can be generally categorized as policy evaluation and policy optimization. The former focuses on evaluating the value of a given target policy or evaluation policy $\boldsymbol{\pi}^e$, while the latter targets at finding the optimal policy. The policy used for generating the data is usually called behavior policy or sampling policy $\boldsymbol{\pi}^b$. When the target policy is the same as the behavior policy, the problem is called an on-policy problem. On the contrary, when the target policy and the behavior policy are different, the problem is called an off-policy problem.

1.2.2 Contextual Bandit

In contextual bandits, each trajectory i only contains a triplet $(\boldsymbol{X}_i, A_i, R_i)$, where the reward R_i is a function of $\boldsymbol{X}_i$ and A_i. For each new subject, we decide which action A_i to take based on $\boldsymbol{X}_i$ and the history information $\boldsymbol{X}_1, A_1, R_1, \ldots, \boldsymbol{X}_{i-1}, A_{i-1}, R_{i-1}$. While the actions may depend on the states and actions of other subjects due to the sequential decision procedure, the states of subjects are usually assumed to be independent. Some works focus on finding the optimal policy that can minimize the cumulative regret,

$$\sum_{i=1}^{n} \mathbb{E}[R_i(a_i^*)] - \mathbb{E}[R_i(A_i)],$$

where $R_i(a_i^*)$ is the reward under the optimal action a_i^* and $R_i(A_i)$ is the reward under the actual action A_i. The optimal policy can be estimated by regression-based methods to fit a model for the Q-function. Parametric methods includes linear regression (Li et al., 2010), LASSO (Bastani and Bayati, 2020), generalized linear models (Filippi et al., 2010); and nonparametric methods includes nearest neighbor (Yang and Zhu, 2002), Gaussian process regression (Krause and Ong, 2011), binning of the covariate space (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013), and local polynomial regression estimators that adjust to any smoothness level (Hu et al., 2020). Dudík et al. (2011) applies the doubly robust technique to the problems to avoid the drawbacks of potential misspecified reward models. Similar as in Section 1.1.1, the optimal action is the one that maximizes the estimated Q-function.

There is other literature that focus on various issues. Zhang et al. (2020b) derived inferential results about the parameters in a model of the batched bandits. Zhang et al. (2021) extended the results to single trajectory, infinite horizon contextual bandit problems using adaptively-weighted least squares. Chen et al. (2020) conducted inference on the model parameter and the value function under a linear reward model and ϵ-greedy policy for correctly specified or misspecified model. Chen et al. (2021a) proposed an efficient algorithm for online update of the rule based on stochastic gradient descent and conducted inference on the model parameter and the value function. In addition, Bastani et al. (2021) showed that a simple greedy algorithm can be rate

optimal when the contexts are sufficiently diverse, and proposed an algorithm that used observed data to determine whether to take a greedy policy or explore.

Instead of focusing on the training performance, pure exploration with fixed budget in multi-armed bandits (Lattimore and Szepesvári, 2020) tries to minimize the simple regret of the test set. Interested readers are referred to Tewari and Murphy (2017); Lattimore and Szepesvári (2020) for a comprehensive review of bandit problems.

1.2.3 General Reinforcement Learning

As discussed in Section 1.1.2, for finite-stage problems, the value functions and the optimal policies can be estimated backward from the last stage to the first stage. However, for indefinite or infinite stage problems, it is hard or impossible to find the last stage or to conduct backward estimation for a large number of stages. It is usually assumed that the decision process is non-Markov decision process (NMDP), time-varying Markov decision process (TMDP), or Markov decision process (MDP). NMDP does not make any structural assumption on the transition probability or the policy, and the joint distribution of the observed data can be written as

$$\mathbb{P}(\boldsymbol{d}) = \mathbb{P}_{\boldsymbol{X}_1}(\boldsymbol{x}_1) \prod_{t=1}^{T} \pi_t^b(a_t|\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \mathbb{P}_{\boldsymbol{X}_{t+1},R_t}(\boldsymbol{x}_{t+1}, r_t|\boldsymbol{h}_t).$$

TMDP is a special case of NMDP which assumes that the observed data satisfy the Markov property and that the behavior policy is a Markov policy. The joint distribution in TMDP can be written as

$$\mathbb{P}(\boldsymbol{D}) = \mathbb{P}_{\boldsymbol{X}_1}(\boldsymbol{x}_1) \prod_{t=1}^{T} \pi_t^b(a_t|\boldsymbol{x}_t) \mathbb{P}_{\boldsymbol{X}_{t+1},R_t}(\boldsymbol{x}_{t+1}, r_t|\boldsymbol{x}_t, a_t).$$

MDP is a special case of TMDP which assumes that the transition probability does not depend on t, so that we can treat every stage homogeneously (Sutton and Barto, 2018). The joint distribution in MDP can be written as

$$\mathbb{P}(\boldsymbol{D}) = \mathbb{P}_{\boldsymbol{X}_1}(\boldsymbol{x}_1) \prod_{t=1}^{T} \pi_t^b(a_t|\boldsymbol{x}_t) \mathbb{P}_{\boldsymbol{X},R}(\boldsymbol{x}_{t+1}, r_t|\boldsymbol{x}_t, a_t).$$

Shi et al. (2020b) developed a test for the Markov assumption in sequential decision making. There are also tests developed to ensure the stationarity of the data by detecting the change point (Li et al., 2022).

In offline policy evaluation, when the data is generated according to the target policy, the value can simply be estimated using the sample average. However, when the data is generated according to some behavior policy different from the target policy, we need to adjust for the bias coming from different distributions of data. Since the data generating process can be expensive, especially in clinical trials, it is often not realistic to perform on-policy evaluation. Not to mention that sometimes we are simultaneously interested in multiple policies, and the expense might be even higher. Therefore, off-policy evaluation (OPE) is an important topic in RL.

The first approach to OPE is called direct method (DM), where the value function

$$\widehat{\mathcal{V}}_{\mathrm{DM}}(\boldsymbol{\pi}^e) = \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_1(\boldsymbol{X}_{it}) \qquad \text{or} \qquad \widehat{\mathcal{V}}_{\mathrm{DM}}(\boldsymbol{\pi}^e) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \widehat{Q}_1(\boldsymbol{X}_{it}, A_{it})\pi_t^e(a|\boldsymbol{X}_{it})$$

is directly estimated by taking the average estimated Q-function or state value function. The Q-function or state value function is usually fitted using the Bellman equation (Sugiyama, 2015). Shi et al. (2021b) proposed such an OPE method by specifying a model for Q-function based on linear sieves and transforming the Bellman equation into a linear regression. The target policies can either be fixed or dependent on the observed data, and the proposed method can be extended to on-policy evaluation. Liao et al. (2021) proposed to minimize the Bellman error and use a coupled estimator to directly solve for the long-term average outcomes. Luckett et al. (2020) also used such method based on the state value function, although their final goal is policy optimization. All the above three papers provide results about the asymptotic distribution of the value function. For finite horizon problems, the value functions can be estimated backward as discussed in Section 1.1.2. However, DM may suffer from bias if the model of value functions is misspecified.

The second approach is importance sampling (IS), which uses the weight between the distribution of actions in the target policy and the behavior policy to adjust the bias in sample average. The weight can be calculated based on the whole trajectory or each decision point (Precup, 2000a),

such that

$$\widehat{\mathcal{V}}_{\mathrm{DM}}(\boldsymbol{\pi}^e) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{T}\frac{\pi_k^e(A_{ik}|\boldsymbol{X}_k,\boldsymbol{H}_{k-1})}{\prod_{k=1}^{T}\pi_k^b(A_{ik}|\boldsymbol{X}_k,\boldsymbol{H}_{k-1})}\left(\sum_{t=1}^{T}R_{it}\right)\right]$$

$$\text{or} \qquad \widehat{\mathcal{V}}_{\mathrm{DM}}(\boldsymbol{\pi}^e) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{t=1}^{T}\frac{\prod_{k=1}^{t}\pi_t^e(A_{ik}|\boldsymbol{X}_k,\boldsymbol{H}_{k-1})}{\prod_{k=1}^{t}\pi_k^b(A_{ik}|\boldsymbol{X}_k,\boldsymbol{H}_{k-1})}R_{it}\right].$$

It does not require the Markov assumption. However, since the total weight is a product of the weights at each stage, the estimate may suffer from a large variance, especially when the horizon is long. To improve the stability, a flattening parameter can be introduced to flatten the importance weight towards one (Sugiyama et al., 2007; Sugiyama, 2015). This flattening parameter can be tuned to balance the bias and variance according to the data. Liu et al. (2018a) proposed to use the stationary state-visitation distributions instead of the product of a long series of weights. The proposed method can efficiently avoid the exploding variance, but may need an accurate estimation of the stationary distribution, which might be challenging especially when the dimension of the states is high.

The doubly robust (DR) estimator

$$\widehat{\mathcal{V}}_{\mathrm{DM}}(\boldsymbol{\pi}^e) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\frac{\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(A_{ik}|\boldsymbol{X}_{ik})}{\prod_{k=1}^{t}\mu_k(A_{ik}|\boldsymbol{H}_{ik})}[R_{it}-\widehat{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{X}_{it},A_{it})] + \frac{\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{ik}|\boldsymbol{X}_{ik})}{\prod_{k=1}^{t-1}\mu_k(A_{ik}|\boldsymbol{H}_{ik})}\widehat{V}_t^{\boldsymbol{\theta}}(\boldsymbol{X}_{it})\right\}.$$

is a combination of DM and IS to solve the potential problems in these methods, where $\widehat{Q}_t$ and $\widehat{V}_t$ are estimated Q-function and state value function. Jiang and Li (2016) proposed to extend the DR estimator of contextual bandits (Dudík et al., 2011) to the multi-stage setting, and Thomas and Brunskill (2016) write the estimator in the form where the augmentation term is constructed based on Q-function and state value function. Although the value functions is model-based and can be misspecified, the DR property ensure that the estimator is unbiased and do not need Markov assumption. Farajtabar et al. (2018) further improve the method by minimizing the variance of the DR estimator to learning the model parameter. Kallus and Uehara (2020) proposed a new estimator double reinforcement learning (DRL), which is the cross-fold version of the DR estimator in non-MDP and uses the marginalized density ratios in MDP. They prove that DRL is semiparametrically

efficient and doubly robust. Shi et al. (2021a) improved DRL using a deeply-debiasing procedure so that the value estimator is consistent when either the Q-function, marginalized density ratio, or conditional density ratio estimator is consistent, and the convergence rate of nuisance parameters can be even slower. Zhang et al. (2022) focused on the inference based on adaptively sampled data from RL algorithms. See Uehara et al. (2022) for a full review of the OPE literature.

Policy optimization relies on an accurate evaluation of the policies. Several papers proposed to use batch or offline data for learning optimal policies. Murphy et al. (2016) estimated the long-term average outcome based on Bellman equation, and find the policy within a policy class that maximizes the long-term average outcome with a quadratic penalty. Liao et al. (2020) proposed a Bayesian methods to update the posterior distribution for the reward every night and generate a probability for sending an activity suggestion based on the posterior distribution. Luckett et al. (2020) estimated the state-value function of a given policy using DM based on Bellman equation, and use gradient descent to find the policy within a policy class. Zhou et al. (2022) proposed a Proximal Temporal consistency Learning (pT-Learning) framework to adaptively adjust between deterministic and stochastic sparse policy when a large number of treatment options existed. Liao et al. (2022) used the DR estimator of the value function based on the marginalized density ratio and proposed a coupled estimation framework to solve for the optimal policy within a policy class. Shi et al. (2022a) proposed an advantage learning framework that can improve the efficiency of any Q-learning type estimator, including deep Q-learning. Nie et al. (2021) developed an "advantage doubly robust" estimator to learn when to start which treatment without Markov assumption. Gao et al. (2023b) developed the deep spectral Q-learning algorithm to handle the mixed frequency data with high dimension. To address the problem of over-estimation of the value for the out-of-distribution actions in off-policy problems, the pessimism principle is deployed to restrict the learned policy to be close to the behavior policy (Yu et al., 2020; Kidambi et al., 2020; Jin et al., 2021; Xie et al., 2021).

Another line of research is about the unobserved confounders in RL. While all the previously mentioned literature assumes sequential ignorobility or no unmeasured confounding, this assumption can be violated in practice. See Shi et al. (2022c) for examples in ride-hailing platform and predictive policing. Some literature focuses on the confounded MDP model (Zhang and Bareinboim, 2016; Bennett et al., 2021). Fu et al. (2022); Xu et al. (2022) studied policy learning and

policy evaluation respectively with the help of instrumental variables in confounded MDP. Shi et al. (2022c) estimated the confidence interval of a target policy's value using some auxiliary variables that mediate the action effects. Wang et al. (2022) took in the expert recommendation that contain information of the unobserved confounders. Another approach is to utilize the partially observable MDP (POMDP) model, where the agent cannot observe the underlying states directly, and the policy can only be estimated from the history data Bennett and Kallus (2021); Shi et al. (2022b). Miao et al. (2022) provided the finite-sample error bound for OPE in POMDPs under non-parametric models with time-dependent proxy variables.

1.2.4 Applications of RL for Estimating ITRs

RL is a useful tool for estimating ITRs in sequential problems. This sequentiality can either lies between patients not enrolled in a clinical trial simultaneously, or between the treatment stages of the same patient.

Contextual bandits can be used to improve personalized treatment suggestion along the treatment process in single-stage treatment decision problems. For example, Bastani and Bayati (2020) applied their proposed LASSO Bandit algorithm on a simplified version of a medication dosing problem, where they sequentially improve the optimal discretized dosing level of Warfarin based on the patients' demographics, diagnoses data, etc. Lei et al. (2017) proposed an actor critic contextual bandit algorithm for personalizing mobile health interventions with linear reward assumption. Hu et al. (2021) proposed a contextual bandit method based on generalized linear mixed model with group lasso type penalty to develop personalized push schedules in mobile health, and the current contextual factor is allowed to be endogenous. Tewari and Murphy (2017) reviews contextual bandits and the challenges in mobile health application. Bandit and RL algorithms have also been widely applied to develop new clinical trial schemes. Minsker et al. (2016) proposed to enroll more patients around the decision boundary to increase the efficiency of estimating ITR based on active learning, which can be viewed as a special case of RL. Durand et al. (2018) designed an adaptive treatment allocation strategy within the contextual bandit framework in a mouse model to increase the data collection efficiency by assigning more samples to promising treatments.

In multi-stage decision problems, Luckett et al. (2020) applied their proposed V-learning on a study of type-I diabetes to decide whether to inject insulin based on the data collected by mobile

devices of patients. Nie et al. (2021) proposed a new estimator based on the doubly robust estimator in RL to decide which treatment to take and when to start the treatment. Liao et al. (2021) used OPE technique to construct confidence intervals of the value of given ITR in a mobile health study, HeartSteps. Liao et al. (2020) developed a Bayesian RL algorithm to generate physical activity suggestions in HeartSteps. Trella et al. (2022) studied the reward design for delayed effects to encourage oral hygiene behaviors with mobile devices. Shi et al. (2021a,b) proposed RL-based methods to construct confidence intervals of the value of a target ITR for a simulated diabetes study, the OhioT1DM dataset.

1.3 Outline of the Dissertation

Motivated by the problems in estimating ITRs and the recent development in RL techniques, we identify a few unaddressed problems in current literature. To address these problems, we propose several new approaches to fill in the gap.

In Chapter 2, we propose a sequentially adaptive trial and discuss the tradeoff between training and test performance of contextual bandits in learning ITRs. Many statistical and ML methods for learning optimal ITRs have been developed in the literature. However, most existing methods are based on data collected from traditional randomized controlled trials and thus cannot take advantage of the accumulative evidence when patients enter the trials sequentially. It is also ethically important that future patients should have a high probability to be treated optimally based on the updated knowledge so far. In this work, we propose a new design called sequentially rule-adaptive trials to learn optimal ITRs based on the contextual bandit framework, in contrast to the response-adaptive design in traditional adaptive trials. In our design, each entering patient will be allocated with a high probability to the current best treatment for this patient, which is estimated using the past data based on some ML algorithm (for example, outcome weighted learning in our implementation). We explore the tradeoff between training and test values of the estimated ITR in single-stage problems by proving theoretically that for a higher probability of following the estimated ITR, the training value converges to the optimal value at a faster rate, while the test value converges at a slower rate. This problem is different from traditional decision problems in the sense that the training data are generated sequentially and are dependent. We also develop a

tool that combines martingale with empirical process to tackle the problem that cannot be solved by previous techniques for i.i.d. data. We show by numerical examples that without much loss of the test value, our proposed algorithm can improve the training value significantly as compared to existing methods. Finally, we use a real data study to illustrate the performance of the proposed method. This work has been published on the *Journal of Machine Learning Research* (Gao et al., 2022).

In Chapter 3, we work on the statistical inference of the parameter of high dimensional ITRs in multi-stage problems (Gao et al., 2023a). One important class of DTR in practice, namely multi-stage stationary treatment policies, prescribe treatment assignment probabilities using the same decision function over stages, where the decision is based on the same set of features consisting of both baseline variables (e.g., demographics) and time-evolving variables (e.g., routinely collected disease biomarkers). Although there has been extensive literature to construct valid inference for the value function associated with the dynamic treatment policies, little work has been done for the policies themselves, especially in the presence of high dimensional feature variables. We aim to fill in the gap in this work. Specifically, we first estimate the multistage stationary treatment policy based on an augmented inverse probability weighted estimator for the value function to increase the asymptotic efficiency, and further apply a penalty to select important feature variables. We then construct one-step improvement of the policy parameter estimators. Theoretically, we show that the improved estimators are asymptotically normal, even if nuisance parameters are estimated at a slow convergence rate and the dimension of the feature variables increases exponentially with the sample size. Our numerical studies demonstrate that the proposed method has satisfactory performance in small samples, and that the performance can be improved with a choice of the augmentation term that approximates the rewards or minimizes the variance of the value function.

In Chapter 4, we consider the case where a clinical trial contains multiple outcomes. In practice, the optimal ITR that maximizes its associated value function is also expected to cause little harm on other non-primary outcomes. For example, when treating the major depressive disorder, the primary goal is to reduce depressive symptoms, but the overall clinical improvement should not be negatively affected. Hence, one goal is to learn the ITR that not only maximizes the value function for the primary outcome, but also approximates the optimal rule for the other auxiliary outcomes as close as possible. In this work, we propose a fusion penalty to encourage ITRs based on the primary

outcome and auxiliary outcomes to yield similar recommendations. We then optimize a surrogate loss function using empirical data for estimation. We derive the non-asymptotic properties for the proposed method and show that the agreement rate between the estimated ITRs for both primary and auxiliary outcomes converges to the true agreement rate at a faster rate as compared to methods without using auxiliary outcomes. Finally, simulation studies and a real data example are used to demonstrate the finite-sample performance of the proposed method.

CHAPTER 2

Non-asymptotic Properties of Individualized Treatment Rules from Sequentially Rule-Adaptive Trials[1]

2.1 Introduction

For many diseases, patients respond heterogeneously to treatments and a one-size-for-all strategy is often not effective. Recent technology advances allow personalized treatment suggestions by tailoring it to patient characteristics, including demographics, medical histories or genetic information (Hamburg and Collins, 2010). The personalized policy is often referred to as the Individualized Treatment Rule (ITR), which aims to maximize a predefined reward such as the patient's health status.

The optimal ITR can be estimated through regression-based or classification-based methods. The former fits a regression model for the rewards and finds the treatment with the maximum estimated reward (Qian and Murphy, 2011). The latter obtains the optimal ITR directly by maximizing the average reward. For example, Zhao et al. (2012) proposed a weighted classification algorithm called outcome weighted learning (OWL), which is based on the support vector machine (SVM) and equipped with various kernels. There are also variations of OWL designed for ITR estimation in single-stage problems (Zhou et al., 2017; Chen et al., 2018) and multi-stage problems (Zhao et al., 2015; Liu et al., 2018b).

For all the above methods, to avoid unobserved confounding bias as present in observational studies, data used to learn optimal ITRs are typically obtained from randomized controlled trials (RCTs), where patients receive treatments based on a prefixed probability rule. RCTs are conducted primarily to compare the efficacy of new treatments. However, in the case when the control drug is not beneficial or is even harmful, patients may have to switch treatments or withdraw from

[1]This chapter previously appeared as an article in the Journal of Machine Learning Research. The original citation is as follows: Gao, D., Liu, Y., and Zeng, D. (2022). Non-asymptotic properties of individualized treatment rules from sequentially rule-adaptive trials. *Journal of Machine Learning Research*, 23(250):1–42.

the study due to little benefit or adverse events under the assigned treatments. This may cause violation of the randomization and result in bias in estimating clinical efficacy. In fact, as data are gathered during the process, we already have an inference about which treatment should be better for the next patient. A more effective design for the trial should be sequentially adaptive so that any new patients entering the trial are more likely to receive the best treatment learned from the past. This is especially important ethically since an inferior treatment may cause severe health issues to a patient. A sequentially adaptive trial has the advantage to better maintain randomization while keeping most of the study participants benefiting from their assigned treatments. As commented in Thall (2002), a clinical trial ideally should provide patients in the trial with the best treatment available, while also generate data for improving therapies. We will discuss the tradeoff between the two goals from a statistical viewpoint in this chapter. We refer to the clinical trial data as the training set and refer to an independent population as the test set for clarity.

The clinical trials that allow the trial protocol to be modified according to observed patient information as the trial continues are called adaptive clinical trials (ACTs) (Chow, 2014). A special class of ACTs is the response-adaptive randomization (Hu and Rosenberger, 2006), which is divided into four categories: restricted randomization, response-adaptive randomization, covariate-adaptive randomization, and covariate-adjusted response-adaptive (CARA) randomization. The latter three are adjusted for response, covariates, and response with covariates respectively. As an example of CARA, Zhang et al. (2007) proposed a framework for the treatment distribution to converge to a predefined distribution, which can be applied to generalized linear models. Hu et al. (2015) suggested to balance ethics in avoiding assigning patients to inferior arms and efficiency in the power of detecting treatment differences. ACTs sometimes also use Bayesian designs to find the optimal dose schedule based on efficacy and toxicity and maximize survival time by combining different phases (Thall et al., 2013; Riviere et al., 2018; Chapple and Thall, 2019). These methods mainly use adaptive designs to improve the efficiency, which refer to the power of estimating average treatment effects, and are not suitable for learning optimal ITRs. There are also a few papers for learning subgroup treatment effects through enrichments (Kim et al., 2011; Lai et al., 2012; Renfro et al., 2016), but they are not optimal for finding ITRs. Furthermore, theoretical justification is lacking for the estimated treatment effects for all subgroups.

There is a close connection between the sequentially adaptive design and the contextual bandit, which is a class of algorithm that deals with online decision problems. As a single-stage special case of reinforcement learning, it aims at making sequential decisions through trial and error. All reinforcement learning algorithms encapsulate an "exploration-exploitation" dilemma. Various exploration methods have been proposed in the contextual bandit literature. The ϵ-greedy methods assign the current optimal arm with a probability of $1-\epsilon$ or chooses from all arms randomly with a total probability of ϵ (Yang and Zhu, 2002; Chen et al., 2020). Boltzmann exploration assigns probabilities of whether to follow the current optimal policy using the soft-max function based on the estimated mean rewards of arms (Sutton and Barto, 2018). Upper-Confidence Bound (UCB) methods choose the arm with the largest upper confidence bound, which either has a large estimated mean reward or a large estimated variance (implying great uncertainty) (Li et al., 2010; Chu et al., 2011; Krause and Ong, 2011). Bayesian methods assign a treatment to a future patient according to the posterior distribution of reward parameters (Chapelle and Li, 2011; Liao et al., 2020). Action elimination is another branch that ignores the inferior arms gradually (Perchet and Rigollet, 2013). Different estimation methods have also been proposed in linear scenarios (Auer, 2002; Li et al., 2010; Chu et al., 2011; Chen et al., 2020; Bastani and Bayati, 2020) and nonlinear scenarios (Yang and Zhu, 2002; Krause and Ong, 2011; Zhou et al., 2020) under the contextual bandits framework. Interested readers are referred to Tewari and Murphy (2017); Lattimore and Szepesvári (2020) for a comprehensive review of bandit problems. However, most works in contextual bandits focus on the training phase and do not address the test performance theoretically. Pure exploration with a fixed budget in multi-armed bandits (MAB, Lattimore and Szepesvári (2020)) also tries to minimize the test regret (also called simple regret), but they generally do not require a small training regret (also called cumulative regret). Bubeck et al. (2009) illustrated the tradeoff between training and test performance in MAB algorithms without a context, that an asymptotically optimal policy for training regret will lead to a suboptimal policy for test regret. Lattimore and Szepesvári (2020) also discussed in Chapter 33 that algorithms with logarithmic cumulative regret in MAB settings (for example UCB) are not well suited for pure exploration. In contrast, we consider more complex settings with context and also provide a way to find the balance point.

To our best knowledge, the most relevant clinical trial design for learning ITR is the active clinical trial (Minsker et al., 2016), which is an active-learning based algorithm. In terms of data

collection, they focus on exploring patients close to the decision boundary and omit the trials on patients known to benefit from one of the treatments with a high probability. However, the actually conducted trials are still purely randomized, and will not benefit from previous information. Practically speaking, the patients omitted from the trial still need to be recruited to collect their basic information before deciding whether they are close to the boundary, which can still create a burden on the trial and the patients.

We propose a sequentially adaptive trial design named "rule-adaptive design" in contrast to "response-adaptive design". It updates the treatment assignment policy during the clinical trial using some statistical or machine learning methods, so that the outcomes in the clinical trial are improved. In the meantime, we also allow for some exploration probability in order to learn an efficient final ITR. In the current work, we consider estimating the two-armed ITR with OWL and explore with ϵ-greedy or a variation of Boltzmann exploration. Different from most contextual bandit methods which rely on a regression model of the rewards, our OWL-based algorithm is a weighted classification method which tries to maximize the rewards directly. Only a model for the treatment effect is specified and thus minimum assumption (for example, boundedness) is needed for the main effect, unlike in Li et al. (2010) and Chen et al. (2020) where a reward model is constructed for the total effect. While Chambaz et al. (2017) and Chen et al. (2020) focused on the inference of the parameters or value functions, we perform the regret analysis.

Specifically, we consider a trial with n sequentially enrolled patients with independent feature variables. Since some of the characteristics of a patient can only be observed after the patient is enrolled in the clinical trial and the process maybe expensive, we assume that we cannot choose which patients to enroll. After a pilot trial of some patients, we assign any incoming patient the estimated optimal treatment learnt from the available data with a probability of p and the other treatment with a probability of $1 - p$. We restrict that p is bounded by $1 - \epsilon$ and ϵ, where ϵ is a positive constant between 0 and 0.5. Furthermore, we let ϵ decay to zero as the ITR estimation gets more accurate over the trial. If the probability p is a constant that does not depend on the current context or the history information, including the characteristics, treatments and rewards of previous patients, the above method is actually ϵ-greedy. Note that the ϵ defined here is one half of that in the definition of ϵ-greedy in most reinforcement learning literature. However, we allow p to be dependent on the current status and the history in theory and in simulation. In this algorithm,

p governs the chance of exploration. Intuitively, a small p indicates a high tendency to follow the current estimated ITR. Future patients to enter the trial are likely to receive a favorable treatment when data accumulate. On the other hand, a small p limits the chance of exploring new treatments. This leads to a slow convergence of the learnt ITR to the optimal one, yielding a suboptimal ITR if the training sample size in the trial is not large enough. This suggests a tradeoff between training and test performance. Our proposed class of algorithms allows adaptive probabilities to depend on already collected data in a flexible way, and includes Boltzmann exploration and an approximate UCB algorithm as special cases.

In this chapter, to fully characterize the performance of the rule-adaptive design, we establish the convergence rate of both the test regret for the learnt ITR if implemented in an independent population, and the training regret for patients in the training set. The former concerns the expected reward loss as compared to the theoretically optimal ITR. The latter describes the cumulative reward loss between actually observed rewards and the hypothetical rewards if each patient would receive the learnt optimal ITR over time. The established bounds depend on the number of initial patients, the number of patients enrolled in the main trial, and the decay rate of the ϵ-sequence. The bounds clearly indicate a tradeoff between the training and test performance of the algorithm. This tradeoff can be useful for us to choose an ϵ-sequence that guarantees a small loss of rewards for the testing sample due to the reduction of exploration in the training process, while at the same time allowing a majority of the experiment patients to receive better than random treatments. To our knowledge, these are the first rigorous results for contextual bandits.

Our proofs for establishing bounds are substantially different from the ones that are based on i.i.d. training data, due to the challenge that the treatment assignment depends on the past data. In the proof, we derive a new concentration inequality for suprema of a martingale sequence by extending the results in Rakhlin et al. (2015). Particularly, to obtain the sequential Rademacher complexity of function classes needed in the inequality, we develop a new mathematical tool that applies the empirical process and bracketing number technique to martingale sequences. Bae and Levental (1995) showed that Freedman's inequality (Freedman, 1975) works well for ergodic Markov chains as a substitution for Bernstein's inequality in i.i.d. sequences. Van de Geer (1995), Nishiyama (1997) and Nishiyama et al. (2000) also took similar approaches in continuous-time martingales or some martingales with jumps. Rakhlin et al. (2015) created a scheme of extending empirical

process and symmetrization methods to martingale. Chambaz et al. (2017) derived a new maximal inequality for martingales based on the uniform entropy integral. However, to our knowledge, our work is the first one to make use of bracketing numbers in the test value bound of martingale sequences. As a remark, we note that Rakhlin and Sridharan (2014) provided a bound for sequential Rademacher complexity of linear functions on dual spaces of covariates and linear coefficients. In contrast, our method applies to any function class with bounded bracketing integral.

The rest of this chapter is organized as follows. In Section 2.2, we describe our proposed algorithm that uses the OWL algorithm for learning ITRs over time. Section 2.3 gives theoretical guarantees for the performance of our algorithm on the training and test sets. We describe the implementation details of our proposed algorithm, and discuss the connections and differences between our algorithm and existing methods in Section 2.4. In Section 2.5, we conduct extensive simulation studies to examine how parameters in our algorithm influence the empirical results, and compare our method with randomized controlled trials, LinUCB (Li et al., 2010) and active clinical trials (Minsker et al., 2016). We further use a real data example to illustrate the advantage of the proposed method in Section 2.6. The chapter is concluded with some remarks in Section 2.7.

2.2 Methodology

We consider the single-stage decision problem, the case where a single treatment recommendation is made for every patient. For each patient, the feature variables or covariates $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ are observed. We assume that the covariates $\{\boldsymbol{X}_i\}_{i=1}^{\infty}$ are drawn from a population independently and identically. Based on the covariates, we need to decide which treatment to take for the patient. We focus on a two-armed problem in this chapter. That is, the treatment A takes values in $\mathcal{A} = \{1, -1\}$. An outcome $R \in \mathbb{R}$ is then observed, which is also called the reward, with higher values desirable. An ITR is a map $\mathcal{D} : \mathcal{X} \mapsto \mathcal{A}$ that assigns the patient of covariates $\boldsymbol{X}$ to a treatment A. An optimal ITR can generate the largest mean reward for the test data. If there exists a measurable discriminant function $f : \mathcal{X} \mapsto \mathbb{R}$ such that $\mathcal{D} = \text{sign}\{f\}$, we only need to find such a function f.

2.2.1 Learning Algorithm for Updating ITRs

We propose to estimate the ITR using machine learning methods, OWL in particular, since it is shown to provide useful ITR recommendations in various scenarios (Zhao et al., 2012). We briefly describe the method of OWL below.

Let $\mathbb{P}$ be the joint distribution of $\boldsymbol{Z} := (\boldsymbol{X}, A, R)$ and $\mathbb{E}$ be the corresponding expectation. If the data are sampled according to the ITR $\mathcal{D}$, that is, given $A = \mathcal{D}(\boldsymbol{X})$, the distribution and expectation are denoted as $\mathbb{P}^{\mathcal{D}}$ and $\mathbb{E}^{\mathcal{D}}$ respectively. Then the optimal ITR can be defined as $\mathcal{D}^* := \arg\max_{\mathcal{D}} \mathbb{E}^{\mathcal{D}}(R)$ and the optimal decision function f^* satisfies $\text{sign}\{f^*\} = \mathcal{D}^*$. Qian and Murphy (2011) showed that the expected reward under policy $\mathcal{D}$ is given by

$$\mathbb{E}^{\mathcal{D}}(R) = \mathbb{E}\left[\frac{R\mathbb{1}(A = \mathcal{D}(\boldsymbol{X}))}{\pi(A; \boldsymbol{X})}\right], \tag{2.1}$$

where $\pi(A; \boldsymbol{X})$ is the probability of taking treatment A given covariates $\boldsymbol{X}$ of a patient. After transforming (2.1) to a loss function based on the 0-1 loss, Zhao et al. (2012) proposed OWL to instead minimize a surrogate loss, hinge loss $\phi(x) = [1 - x]^+$. That is, they try to find the function f that minimizes $\mathbb{E}[g^f(\boldsymbol{Z})]$, where $g^f(\boldsymbol{Z}) = R\phi(Af(\boldsymbol{X}))/\pi(A; \boldsymbol{X})$. If we obtain a total number of n observations, OWL tries to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi_i(A_i; \boldsymbol{X}_i)}\phi(A_i f(\boldsymbol{X}_i)).$$

A penalty term can be added to the loss function for high-dimensional settings to avoid overfitting. This is a weighted classification problem that can make use of the framework of SVM. The estimated ITR can be obtained by taking $\hat{D} = \text{sign}\{\hat{f}\}$. The resulting estimator of ITR generated by OWL is consistent (Zhao et al., 2012). Moreover, R_i can be replaced by $R_i - \mathbb{E}(R|\boldsymbol{X}_i)$ to further improve the learning performance (Liu et al., 2018b).

2.2.2 Sequentially Rule-Adaptive Trials (SRATs)

We describe the proposed algorithm to improve the clinical trial outcome and learn the optimal ITR as follows. Before the trial begins, assume we already have a pure randomized pilot trial of small size n_0, from which our first function $\hat{f}_0$ can be estimated. Then the first patient $i = 1$ can

choose to follow $\hat{\mathcal{D}}_0$ or not. The observations in initial samples all have a propensity score of 0.5. The function is updated after each patient has been treated. Denote the estimated function based on data before the ith patient coming as $\hat{f}_{i-1}$, and the corresponding ITR as $\hat{\mathcal{D}}_{i-1}$ for $i = 1, \ldots, n$. Assume p_i is a probability that can depend on the current feature variables $\boldsymbol{X}_i$ and the history information of previous patients, bounded away from 0 and 1 for all i. At each time point i, we choose to follow our current estimated ITR $\hat{\mathcal{D}}_{i-1}$ with a probability p_i or choose the other treatment with a probability $1 - p_i$. Let I_i be a binary variable such that the ith treatment follows $\hat{\mathcal{D}}_{i-1}$ if $I_i = 1$ and follows $-\hat{\mathcal{D}}_{i-1}$ if $I_i = -1$. That is, I_i takes the value 1 with a probability of p_i and the value -1 with a probability of $1 - p_i$. Then the treatment can be chosen as $A_i = I_i \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$.

When p_i only depends on the order i but not on the history and covariates, our algorithm actually follows the ϵ-greedy exploration method. Note that the randomization probability is sometimes described in another way. In most reinforcement learning literature, for $\tilde{\epsilon}_i \in (0, 1]$ at stage i, ϵ-greedy chooses the best arm with a probability of $1 - \tilde{\epsilon}_i$; and with a total probability of $\tilde{\epsilon}_i$, it chooses from all arms randomly with equal probability. Our definition coincides with this in the sense that $1 - p_i = \tilde{\epsilon}_i/2$. We use the slightly different notation here to describe the boundedness assumption of p_i in a more general way. In the special case when $p_i = 0.5$ for all $i = 1, \ldots, n$, the adaptive clinical trial degenerates into a purely randomized clinical trial. Besides, as a limiting case without truncation, Boltzmann exploration assumes $p_i = \text{logit}^{-1}(\text{benefit}_i)$, where benefit_i is the difference between the estimated rewards of two treatments for the patient i.

Although we choose the presumed best arm with a high probability, there is also some chance that we explore the other arm and observe consequences. When i is small, estimations are usually not accurate due to the large sampling bias and estimation bias. A large probability should be assigned to the inferior arm to allow for exploration and reduce variances. When data accumulate and the ITR estimation gets more accurate as i increases, we will take a higher probability for following the current estimated ITR. Therefore, a decreasing sequence of $\{p_i\}_{i=1}^n$ is desirable. As n goes to infinity, we want $p_n \to 0$ if our estimation method is consistent. The speed at which p_n decreases depends on the convergence rate of estimation.

Let $\boldsymbol{Z}_i^{(0)} = \{\boldsymbol{X}_i^{(0)}, A_i^{(0)}, R_i^{(0)}, I_i^{(0)}\}$ be the feature variables, treatments and rewards of the patient $i = 1, \ldots, n_0$ in the pilot trial. Here $I_j^{(0)}$ can take any value since we do not have an estimated ITR to follow in the pilot trial. Similarly, denote $\boldsymbol{Z}_i = \{\boldsymbol{X}_i, A_i, R_i, I_i\}$ to be all the

information about the patient $i = 1, \ldots, n$ in the main trial. Extend the definition of $\mathbb{P}, \mathbb{P}^{\mathcal{D}}$ and $\mathbb{E}, \mathbb{E}^{\mathcal{D}}$ to be the joint distributions and expectations of $\boldsymbol{Z}$ respectively. For simplicity, denote $\boldsymbol{H}_{i-1}$ as the history information for the ith patient, where $\boldsymbol{H}_0 := \{\boldsymbol{Z}_1^{(0)}, \boldsymbol{Z}_2^{(0)}, \ldots \boldsymbol{Z}_{n_0}^{(0)}\}$ and $\boldsymbol{H}_i := \{\boldsymbol{Z}_1^{(0)}, \boldsymbol{Z}_2^{(0)}, \ldots \boldsymbol{Z}_{n_0}^{(0)}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_i\}, i = 1, \ldots, n$. Then before we decide which treatment to take for the ith ($i = 1, \ldots, n$) patient, the data that we can base our decision on are $\{\boldsymbol{H}_{i-1}, \boldsymbol{X}_i\}$. The final ITR is estimated from the whole training sample $\boldsymbol{H}_n$.

To adapt the algorithm of estimating ITRs to our sequential setting, we will denote $\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ as the probability of taking treatment A_i at stage i to indicate that it depends on the history $\boldsymbol{H}_{i-1}$ and the covariates $\boldsymbol{X}_i$ for the main trial. The probability $p_i = p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ defined as $\mathbb{P}(I_i = 1 | \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ also depends on the history and covariates, and is a simplified notation for $\pi_i(\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i); \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$.

We make the following assumptions to quantify potential outcomes for both the pilot trial and the main trial. Although data are sequentially generated in the main trial, R_i still only depends on A_i and $\boldsymbol{X}_i$ for each $i = 1, \ldots, n$.

Assumption 1 (Ignorability)**.** The treatment A_i ($A_i^{(0)}$) is independent of the potential outcome $R_i^*(a)$ ($R_i^{(0)*}(a)$) given feature variables $\boldsymbol{X}_i$ ($\boldsymbol{X}_i^{(0)}$) for all $a \in \mathcal{A}$ and all $i = 1, \ldots, n$ ($i = 1, \ldots, n_0$).

Assumption 2 (Consistency)**.** The observed outcome R_i ($R_i^{(0)}$) under a treatment $A_i = a$ ($A_i^{(0)} = a$) equals the potential outcome $R_i^*(a)$ ($R_i^{(0)*}(a)$) for all $a \in \mathcal{A}$ and all $i = 1, \ldots, n$ ($i = 1, \ldots, n_0$).

For the pilot trial, we make an additional assumption on propensity scores.

Assumption 3 (Positivity)**.** There exists a constant $c_0 > 0$ such that $\pi_i(a; \boldsymbol{X}_i^{(0)}) \geq c_0$ for all $a \in \mathcal{A}$ and all $\boldsymbol{X}_i^{(0)} \in \mathcal{X}$ for all $i = 1, \ldots, n_0$.

We do not need to make the positivity assumption for the main trial since it is guaranteed by our data generating process when we require $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ to be bounded away from 0 and 1. We will formally quantify the assumptions on the probability $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ in Sections 2.3.1 and 2.3.2. Different choices of $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ will be discussed in Section 2.4 and their performances will be compared in Section 2.5.

Following the scheme of OWL, we propose to minimize the ϕ-risk using the hinge loss in a function class $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}$. Using OWL, we can obtain the first estimated function

$$\hat{f}_0 = \arg\min_{f \in \mathcal{F}} \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{R_j^{(0)}}{\pi_j(A_j^{(0)})} \phi(A_j^{(0)} f(\boldsymbol{X}_j^{(0)})) \tag{2.2}$$

using the pilot trial and update it to get

$$\hat{f}_i = \arg\min_{f \in \mathcal{F}} \frac{1}{n_0 + i} \left\{ \sum_{j=1}^{n_0} \frac{R_j^{(0)}}{\pi_j(A_j^{(0)})} \phi(A_j^{(0)} f(\boldsymbol{X}_j^{(0)})) + \sum_{j=1}^{i} \frac{R_j}{\pi_j(A_j; \boldsymbol{H}_{j-1}, \boldsymbol{X}_j)} \phi(A_j f(\boldsymbol{X}_j)) \right\} \tag{2.3}$$

for $i = 1, \ldots, n$ along with the main trial. For weighted SVM problems, the function class $\mathcal{F}$ is generally taken to be a linear space for linear decision rules or a reproducing kernel Hilbert space (RKHS) for nonlinear decision rules. The full algorithm is summarized in Algorithm 1.

Algorithm 1: Sequentially Rule-Adaptive Trial

Initialize. For n_0 number of patients, assign treatments randomly with equal probabilities and observe $\{\boldsymbol{Z}_j^{(0)}\}_{j=1}^{n_0}$;
Estimate $\hat{f}_0$ with $\{\boldsymbol{Z}_j^{(0)}\}_{j=1}^{n_0}$ by (2.2);
for $i = 1, \ldots, n$ **do**
  Observe feature variables $\boldsymbol{X}_i$;
  Estimate the best treatment $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i) = \text{sign}\{\hat{f}_{i-1}(\boldsymbol{X}_i)\}$;
  Sample I_i from $\{-1, 1\}$ with a probability $\{1 - p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i), p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)\}$ respectively;
  Take the treatment $A_i = I_i \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ and observe the reward R_i;
  Update the function $\hat{f}_i$ with $\{\boldsymbol{Z}_j^{(0)}\}_{j=1}^{n_0} \cup \{\boldsymbol{Z}_j\}_{j=1}^{i}$ by (2.3).
end
Let $\hat{\mathcal{D}}_n = \text{sign}\{\hat{f}_n\}$ be the final estimated ITR.

2.3 Theoretical Results for SRAT

In order to demonstrate a tradeoff between training and test performance of our algorithm, we need to bound the estimated value function on both sets. Previous work has shown a bound for the test value trained on i.i.d. data (Zhao et al., 2012). We will expand the bounds of OWL to dependent training samples.

2.3.1 Performance Guarantee for the Test Set

Define $\mathcal{V}(f) := \mathbb{E}^{\text{sign}\{f\}}(R)$ as the value function of f. We use value function $\mathcal{V}(\hat{f}_n)$ as an indicator of how well our algorithm performs on the test set, after training on n observations. We will call $\mathcal{V}(\hat{f}_n)$ the test value, and define $\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n)$ as the test regret. Remember that $\mathcal{V}(f^*) = \max_f \mathcal{V}(f)$ according to the definition of f^*. In this chapter we assume that the optimal function f^* belongs to the function class $\mathcal{F}$, in which we find the estimated ITR. As a consequence, Zhao et al. (2012, Theorem 3.2) implies that the excess risk satisfies $0 \leq E^{\text{sign}\{f^*\}}(R) - E^{\text{sign}\{f\}}(R) \leq E[g^f(Z)] - E[g^{f^*}(Z)]$. If f minimizes $E[g^f(Z)]$, the right-hand side cannot be larger than zero since f^* is also in the function class $\mathcal{F}$. Therefore, we have $E^{\text{sign}\{f\}}(R) = E^{\text{sign}\{f^*\}}(R)$, which suggests that f and f^* have the same Bayesian risk. For example, we would reasonably assume that f^* is linear in the covariates or in the basis function of covariates for a linear space $\mathcal{F}$. The following result shows that the test regret converges in probability and gives the convergence rate.

We first introduce some key notations. With $N_{[]}(\eta, \mathcal{F}, \|\cdot\|)$ being the bracketing number for the set $\mathcal{F}$ with respect to the semi-norm $\|\cdot\|$, define a bracketing integral of $\mathcal{F}$ as

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) := \int_0^\delta \sqrt{1 + \log N_{[]}(\eta, \mathcal{F}, \|\cdot\|)} \mathrm{d}\eta.$$

Let $L_2(\mathbb{P})$ norm be the L_2 norm with respect to measure $\mathbb{P}$. An envelope of function class $\mathcal{F}$ is any function $F : \mathcal{X} \mapsto \mathbb{R}$ such that $f(x) \leq F(x)$ for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. The minimal envelope function is $F(x) = \sup_{f \in \mathcal{F}} |f(x)|$, for all $x \in \mathcal{X}$. The $*$ symbols on the top right corner of $\mathbb{P}$ and $\mathbb{E}$ indicate outer probability and the corresponding outer expectation respectively in order to avoid measurability problems (Van der Vaart and Wellner, 1996).

Assumption 4. Suppose we have a nonincreasing sequence of $\{\epsilon_1, \ldots, \epsilon_n\}$ with $\epsilon_i \in (0, 0.5]$ for all $i = 1, \ldots, n$, where each ϵ_i can only depend on the order i. Assume $\epsilon_i \leq p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \leq 1 - \epsilon_i$ almost surely for all i.

Assumption 5. There exists a positive constant r such that $\|R_i\|_\infty \leq r$ for all i.

Assumption 6. Suppose $\mathcal{F}$ is a class of measurable functions satisfying

$$\int_0^\infty \sqrt{1 + \log N_{[]}(\eta, \mathcal{F}, L_2(\mathbb{P}))} d\eta < \infty. \tag{2.4}$$

Let F be the minimal envelope function of $\mathcal{F}$ and assume F has a weak second moment, that is, $x^2\mathbb{P}^*(F(\boldsymbol{X}) > x) \to 0$ as $x \to \infty$.

Theorem 2.3.1. *Assume the pilot trial satisfies Assumptions 1, 2, 3, 5 and the main trial satisfies Assumptions 1, 2, 4, 5. If we take $c_0 = 0.5$ and a function class satisfying Assumption 6 in Algorithm 1, then with a probability higher than $1 - e^{-\delta}$ for any $\delta > 0$,*

$$\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n) \leq \frac{C}{n_0 + n}\left[(J + \sqrt{\delta b})r\sqrt{n_0} + rb\delta + \frac{r^2 bJ}{\epsilon_n^2}\sqrt{\delta n \log^3 n}\right], \tag{2.5}$$

where $J := \sup_{\mathbb{P}} J_{[]}(\|F\|_{\mathbb{P},2}, \mathcal{F}, L_2(\mathbb{P}))$, $b := \sup_{f\in\mathcal{F}} \|f\|_\infty$, and C is a constant depending on δ, r, b, J and $\{\epsilon_i\}_{i=1}^n$.

Remark. The above bound shows that the terms containing n_0 are not dominant as long as the order of n_0 does not exceeds the order of n, since ϵ_n is nonincreasing and ϵ_n^{-2} has an order of $\Omega(1)$. In practice, n_0 can be taken as the minimum value that a stable initial rule $\hat{f}_0$ can be estimated with. For example, if the covariates $\boldsymbol{X}$ has a dimension d including an intercept, n_0 can be taken as $d+1$ for linear kernel. We choose n_0 to be a small constant in our simulation study in Section 2.5. For generality, we will assume that $n_0 = O(n)$ in the following analysis, which includes the constant n_0 as a special case.

Remark. Note that the bracketing number and covering number here are defined for i.i.d. data, since $\mathcal{F}$ is defined on $\mathcal{X}$ and the observed feature variables $\{\boldsymbol{X}_i\}_{i=1}^\infty$ are i.i.d. The constant J characterizes the complexity of the function class $\mathcal{F}$. It generally increases as the dimension d of covariates increases, and will result in a larger upper bound.

Remark. For the bound (2.5) to be non-trivial, we need the right-hand side to be $o_p(1)$. That is, when assuming $n_0 = O(n)$, we need ϵ_n to decay slower than $n^{-1/4}\log^{3/4}(n)$ and J to be finite. Intuitively speaking, if the ϵ sequence decays too fast and the algorithm is extremely greedy in the training process, then the data sample is biased and cannot be used to learn an efficient final ITR.

Remark. Theorem 2.3.1 holds when n is large enough but finite, so Assumption 4 ensures that the positivity assumption is satisfied for all n. The randomness parameter ϵ_n, which can be close to zero, is incorporated in the error bounds and accounts for the variance inflation in the value estimation. Simulation study in Section 2.5 shows that there is no significant variance inflation for

different choices of ϵ_n sequences in practice. The complexity of the function class containing π_i increases as the lower and upper bounds of propensity score get wider, but our proof only relies on the lower bound ϵ_i.

In our sequentially dependent algorithm, any constant sequence $\{\epsilon_1, \ldots, \epsilon_n\}$ can generate a convergence rate of $n^{-1/2}\log^{3/2}(n)$ as long as $n_0 = O(n)$. If we take $\epsilon_i = 0.5$ for all i, the algorithm degenerates to pure randomization. Therefore, the traditional RCT is actually a special case contained in our framework. Zhao et al. (2012) proved that the convergence rate of OWL with the Gaussian kernel almost achieves $n^{-1/2}$ under the Geometric noise assumption. The extra $\log^{3/2}(n)$ term comes from a martingale concentration inequality that we used, as shown in Section 2.8.3 in the supplementary material. This indicates that the efficiency of learning ITR is not significantly affected by using sequentially generated data.

Example 2.1. If $\mathcal{F}$ is a class of linear functions with bounded parameters $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^d$, the above assumptions are satisfied. Linear functions are Lipschitz in parameters in the sense that $|f_{\boldsymbol{\beta}_1}(\boldsymbol{x}) - f_{\boldsymbol{\beta}_2}(\boldsymbol{x})| \leq m(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)G(\boldsymbol{x})$ for Euclidean metric m on the index parameter set, $G(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$, and for every $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ by Cauchy–Schwarz inequality. By Theorem 2.7.11 of Van der Vaart and Wellner (1996), $N_{[]}(2\eta \|G\|, \mathcal{F}, \|\cdot\|)$ is bounded by $N(\eta, \mathcal{B}, m)$. Since $N(\eta, \mathcal{B}, m) \leq K/\eta^d$ for some constant K, $N_{[]}(\eta, \mathcal{F}, L_2(\mathbb{P}))$ can be bounded by $2^d K \|G\|^d_{\mathbb{P},2}/\eta^d$ for all measure $\mathbb{P}$. If we further assume that $\|G\|_{\mathbb{P},2} \leq u$ for all measure $\mathbb{P}$ and some constant $u > 0$, for example, when the covariate space $\mathcal{X}$ is bounded, then the constant J and the integral in (2.4) is finite. The assumptions in Theorem 2.3.1 are then satisfied. $\lozenge$

The general idea of proof is to find a classification risk bound for the weighted SVM on sequentially generated data. It is quite similar to the proof idea of Theorem 4 in Bartlett et al. (2006). However, the key step of their proof relies on a variant of Talagrand's inequality (Talagrand, 1994; Bousquet, 2002), which is a concentration inequality of suprema of empirical process on i.i.d data. On the contrary, our algorithm generates data that are adapted to a filtration.

We will define some new notations here. For any sequence $\{Y_i\}_{i\in\mathbb{N}}$ adapted to a filtration $\{\mathcal{G}_i\}_{i\in\mathbb{N}}$, observe that $\{\mathbb{E}_{i-1}f(Y_i) - f(Y_i)\}_{i\in\mathbb{N}}$ is a martingale difference sequence for any measurable function f, where $\mathbb{E}_{i-1}(\cdot) := \mathbb{E}(\cdot|\mathcal{G}_{i-1})$. Define a martingale process indexed by $f \in \mathcal{F}$ analogous to

an empirical process as

$$f \mapsto \mathbb{M}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{i-1}f(Y_i) - f(Y_i)\right\}.$$

In accordance with Rakhlin et al. (2015), the scaling factor $\sqrt{n}$ is not included in the definition.

In our setting, let $\mathcal{G}_0 = \sigma\{\boldsymbol{H}_0\}$ and $\mathcal{G}_i = \sigma\{\boldsymbol{H}_i\}, i \in \mathbb{N}$, so that $\{\boldsymbol{Z}_i\}_{i\in\mathbb{N}}$ is adapted to the filtration $\{\mathcal{G}_i\}_{i\in\mathbb{N}}$. Similar as the definition in Section 2.2.1, let the loss function on a single observation in a sequential experiment be

$$g^f(\boldsymbol{Z}_i) = \frac{R_i\phi(A_i f(\boldsymbol{X}_i))}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}.$$

Note that $\boldsymbol{Z}_i$ is implicitly dependent on the history $\boldsymbol{H}_{i-1}$ through A_i. Define $h^f(\boldsymbol{Z}_i) = g^f(\boldsymbol{Z}_i) - g^{f^*}(\boldsymbol{Z}_i)$ as the difference between the loss generated by any f and the optimal function f^*. Based on our weighted classification setting, we can further define a weighted version of the martingale process by

$$\mathbb{W}_n(f) := \mathbb{M}_n(h^f) = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\Big].$$

The key step is to bound the test regret by the conditional expectations of h^f. To extend the idea to a martingale sequence, we make use of sequential complexity techniques and a suprema concentration inequality presented in Rakhlin et al. (2015, Lemma 2.8.2). The inequality essentially relies on $\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{W}_n(f)$, so we first present the following lemma for the upper bound of the expectation of suprema. We use the symbol "$\lesssim$" to indicate that the left-hand side is no larger than the right-hand side for all n up to a universal constant.

Lemma 2.3.2. *Assume the main trial satisfies Assumptions 1, 2, 4, 5. If we take a function class satisfying Assumption 6 in Algorithm 1, then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{W}_n(f) \lesssim \frac{r}{\sqrt{n}\epsilon_n} J_{[]}(\|F\|_{\mathbb{P},2}, \mathcal{F}, L_2(\mathbb{P})). \tag{2.6}$$

The above lemma suggests that if some f performs well enough on the training set compared to f^*, then it should not be too bad on the test set as well. When ϵ_n does not depend on n, $\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{W}_n(f)$ converges at a rate of $n^{-1/2}$, which is the same as the rate for independent data.

2.3.2 Performance Guarantee for the Training Set

We propose to use $\bar{R}_n := \sum_{i=1}^n R_i/n$ as the measure of performance on the training set, which does not concern the pilot trial. We will call $\bar{R}_n$ the training value and it indicates what we really observe in n patients drawn out of the population. Furthermore, we define our regret on the training set as $\sum_{i=1}^n [\mathcal{V}(\hat{f}_{i-1}) - R_i]/n$. Each observed reward R_i is compared with the corresponding $\mathcal{V}(\hat{f}_{i-1})$, which is the value function based on previous $(i-1)$ data points, and the sum of differences is recorded.

A common metric in bandit problems for training data is the cumulative regret for n observations. It is defined as the difference between the expectation of the sum of rewards under the optimal ITR and that under the estimated ITR, that is, $\sum_{i=1}^n \mathbb{E}_{\boldsymbol{x}_i} R(\mathcal{D}^*(\boldsymbol{x}_i)) - \mathbb{E}_{\boldsymbol{x}_i} R(\hat{\mathcal{D}}_{i-1}(\boldsymbol{x}_i))$, where $\boldsymbol{x}_i$ is the instantiated tailoring variable vector for the ith $(i = 1, \ldots, n)$ patient. It mainly measures how much benefit the actual treatments generate compared with the optimal ones for fixed tailoring variables $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ regardless of the randomness in rewards. A bound on the expectation of regret or a probably approximately correct (PAC) bound is often derived. However, the true optimal rule $\mathcal{D}^*$ and the expectation of rewards are unknown in the training process. Furthermore, the cumulative regret does not include the intrinsic randomness in rewards.

Here we present the training regret bound in terms of our definition with an additional assumption on the randomization probability p_i.

Assumption 7. Suppose we have another nonincreasing sequence of $\{\epsilon_1', \ldots, \epsilon_n'\}$ with $\epsilon_i' \in (0, 1)$ for all $i = 1, \ldots, n$, where each ϵ_i' can only depend on the order i. Assume that $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 1 - \epsilon_i'$ almost surely for all i.

Under Assumptions 4 and 7, the two sequences $\{\epsilon_i\}_{i=1}^n$ and $\{\epsilon_i'\}_{i=1}^n$ actually help create upper and lower bounds of $1 - p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$, which are needed in the test and training regret bounds respectively.

Theorem 2.3.3. *Assume the main trial satisfies Assumptions 1, 2, 5, 7 and we have* $0 < p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) < 1$ *for all* i. *Then with a probability higher than* $1 - e^{-\delta}$ *for any* $\delta > 0$,

$$\left| \frac{1}{n} \sum_{i=1}^n \left[\mathcal{V}(\hat{f}_{i-1}) - R_i \right] \right| \leq C' r \left[\sqrt{\frac{\delta \wedge \delta^2}{n}} + \left(\frac{1}{n} \sum_{i=1}^n \epsilon_i' \right) \right], \tag{2.7}$$

where C' is a constant depending on δ, r and $\{\epsilon_i'\}_{i=1}^n$.

Remark. The concentration bound implies that the training regret is upper bounded by the average of the ϵ_i' sequence plus a term of order $O_p(1/\sqrt{n})$. ϵ_i' should be of order $o(1)$ if we need the training regret to converge to zero. Otherwise, if there is always some probability that the inferior treatment is taken, the training reward cannot be optimal. Specifically, when $\{\epsilon_i'\}_{i=1}^n$ is constant and does not rely on i, the above bound is a constant. The purely randomized clinical trial is a special case of this setting.

Remark. In most of the cases, the randomization probability of following the inferior treatment is $1-p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \leq \epsilon_i' \leq 0.5$ and is nonincreasing. For example, Assumption 7 is satisfied by ϵ-greedy with nonincreasing $\epsilon_i' = \epsilon_i$ for all i. However, in some special cases such as Boltzmann exploration, $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ can be less than 0.5 if the estimated benefit is negative. This can happen when the method for learning ITR (OWL in our design) and that for estimating benefit (for example, linear regression) are different. While OWL recommends $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$, the difference between the estimated rewards of $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ and $-\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ can be negative. In this case, we also need the probability of a negative benefit to converge to zero at a certain rate.

If we assume the true optimal value function $\mathcal{V}(f^*)$ is known and compare each R_i for $i = 1, \ldots, n$ with it, we have the following result. Note that $\left|\mathcal{V}(f^*) - \bar{R}_n\right|$ is a notion more similar to the cumulative regret. Except for the randomness in rewards, the difference only lies in the optimal value. While the cumulative regret considers maximum rewards for each individual, we still focus on the population value.

Corollary 2.3.4. *Let the assumptions in Theorems 2.3.1 and 2.3.3 hold. With a probability higher than $1 - e^{-\delta}$ for any $\delta > 0$,*

$$\left|\mathcal{V}(f^*) - \bar{R}_n\right| \leq C''\Bigg[r\sqrt{\frac{\delta \wedge \delta^2}{n}} + r\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i'\right) + \frac{1}{n(n_0+i)}\sum_{i=0}^{n-1}\Bigg((J+\sqrt{(\delta+\log n)b})r\sqrt{n_0} + rb(\delta+\log n) + \frac{r^2bJ}{\epsilon_i^2}\sqrt{i\log^3 i(\delta+\log n)}\Bigg)\Bigg],$$

where C'' is a constant depending on r, b, J, δ and the sequences $\{\epsilon_i\}_{i=1}^n$, $\{\epsilon_i'\}_{i=1}^n$, if we take $i\log^3 i = 0$ for $i = 0$.

The above corollary demonstrates the well-known exploration-exploitation tradeoff in contextual bandits when the observed reward is compared to the true optimal value. The first two terms on the left-hand side come from Theorem 2.3.3, which characterize the loss in the value due to exploration and increases as ϵ_i' increases. On the other hand, the last term, which comes from Theorem 2.3.1, describes the regret of exploiting the estimated ITR compared to the optimal ITR and decreases with more exploration. The optimal rate is achieved when the two components strike a balance.

2.3.3 Tradeoff Between Training and Test Values

In this section, we discuss the tradeoff between the training value and the test value. To better describe the convergence rates of training and test values, we can set a decreasing schedule for ϵ_n and ϵ_n'. Here we assume ϵ_n and ϵ_n' decreases polynomially with n since the upper bounds in (2.5) and (2.7) are dominated by polynomial terms of n.

Theorem 2.3.5. *Assume $\epsilon_n = \epsilon_0 n^{-(1-\theta)/4}$ with $\epsilon_0 \in (0, 0.5]$, $\theta \in (0, 1]$ and $\epsilon_n' = \epsilon_0' n^{-(1-\theta')/4}$ with $\epsilon_0' \in (0, 1)$, $\theta' \in (-\infty, 1]$. Let Assumptions 1-7 hold and assume $\epsilon_n \leq \epsilon_n'$ for all n. If $n_0 = O(n)$, then the test value $\mathcal{V}(\hat{f}_n)$ converges to $\mathcal{V}(f^*)$ at a rate of $O_p(n^{-\theta/2}(\log n)^{3/2})$, and the training value $\sum_{i=1}^n R_i/n$ converges to $\sum_{i=1}^n \mathcal{V}(\hat{f}_{i-1})/n$ at a rate of $O_p(n^{-(1-\theta')/4})$. If we further assume that $\theta = \theta'$ and $\epsilon_0 \leq \epsilon_0'$, then the two regrets converge at the same rate $O_p(n^{-1/6})$ when $\theta = 1/3$.*

The above results suggest that the convergence rate in the logarithmic scale is negative in θ for the test regret and positive in θ' for the training regret. When θ and θ' are close to 0, $\{\epsilon_i\}_{i=1}^n$ and $\{\epsilon_i'\}_{i=1}^n$ decay fast and the algorithm is greediest on the training set, leading to a fast convergence of the training value and a slow convergence of the test value. On the contrary, when $\theta = \theta' = 1$, $\{\epsilon_i\}_{i=1}^n$ and $\{\epsilon_i'\}_{i=1}^n$ are constant sequences that does not change with the order i. The test value converges quickly while the training value may not converge in this case. This demonstrates in theory why there is a tradeoff between training and test values. Where the "balance" point is can be defined differently in difference settings. Theorem 2.3.5 provides a balance point where the two rates match with each other. In Sections 2.5 and 2.6, we will further demonstrate the tradeoff between training and test values using numerical examples. Note that ϵ-greedy satisfies the assumptions with $\theta = \theta'$ and $\epsilon_0 = \epsilon_0'$.

2.4 Implementation

Recall that in theory we allow the randomization probability $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ to be a constant or be dependent on the current covariates $\boldsymbol{X}_i$ and the history $\boldsymbol{H}_{i-1}$. In implementation, when $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ is a constant that only depends on the order i, the exploration method becomes the special case ϵ-greedy. We call the full algorithm SRAT-E in this case.

To build a bridge between ϵ-greedy and UCB methods, for example LinUCB (Li et al., 2010) in linear cases, we propose to let $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ depend on the history in the following way. While OWL provide an estimation of ITR, we need a separate regression model to show how much benefit a patient will gain from one treatment against the other. In the case of a greatly positive benefit, we can assign the current patient $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ with a large probability since we are almost sure that this treatment is the better one. On the contrary, if the benefit is negative, we allow for more exploration. Specifically, let $\hat{\mu}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ and $\hat{\sigma}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ be the estimated mean and standard deviation of the reward of the ith patient given the treatment a, where $a \in \{\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i), -\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)\}$. Denote $\hat{U}_a(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \hat{\mu}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) + \alpha_i \hat{\sigma}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ as the upper confidence bound of the estimated reward, where α_i is a constant tuning parameter that does not depend on $\mathcal{G}_{i-1}$ or $\boldsymbol{X}_i$. Note that the estimations rely on the regression model completely, since OWL does not provide an estimation of rewards, but only provides a distance between the covariate point and the decision boundary. Further define

$$\hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \hat{U}_{\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) - \hat{U}_{-\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$$

as the UCB-based benefit, which is the difference between the estimated UCB of rewards given two treatments. Let the probability p_i be

$$p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \begin{cases} 1 - \epsilon_i & \text{if } \hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 0, \\ \max\left\{\epsilon_i, \text{logit}^{-1}\left\{\frac{\hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}{\gamma_i}\right\}\right\} & \text{if } \hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) < 0, \end{cases}$$

where γ_i is a constant tuning parameter that does not depend on $\mathcal{G}_{i-1}$ or $\boldsymbol{X}_i$. Recall that we truncate the probability by ϵ_i because we require that p_i is bounded away from 0 and 1. We

call this method SRAT-B since the randomization probability is partially based on Boltzmann exploration.

In practice, we can estimate $\hat{\mu}_a$ and $\hat{\sigma}_a$ for $a \in \mathcal{A}$ by $\hat{\mu}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_a(\boldsymbol{H}_{i-1})$ and $\hat{\sigma}_a(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = [\boldsymbol{X}_i^T \hat{\mathbf{W}}_a(\boldsymbol{H}_{i-1})^{-1} \boldsymbol{X}_i]^{1/2}$, where $\hat{\boldsymbol{\beta}}_a(\boldsymbol{H}_{i-1})$ and $\hat{\mathbf{W}}_a(\boldsymbol{H}_{i-1})$ are the estimated linear parameter and variance matrix before stage i. Following Li et al. (2010, Algorithm 1), the initial estimates can be obtained by

$$\hat{\mathbf{W}}_a(\boldsymbol{H}_0) = \mathbf{I}_d + (\mathbf{X}_a^{(0)})^T \mathbf{X}_a^{(0)}, \quad \hat{\boldsymbol{Y}}_a(\boldsymbol{H}_0) = (\mathbf{X}_a^{(0)})^T \boldsymbol{R}_a^{(0)}, \quad \text{and } \hat{\boldsymbol{\beta}}_a(\boldsymbol{H}_0) = \hat{\mathbf{W}}_a(\boldsymbol{H}_0)^{-1} \hat{\boldsymbol{Y}}_a(\boldsymbol{H}_0),$$

where

$$\mathbf{X}_a^{(0)} := [\boldsymbol{X}_j^{(0)}]_{j:A_j^{(0)}=a}^T \quad \text{and} \quad \boldsymbol{R}_a^{(0)} := [R_j^{(0)}]_{j:A_j^{(0)}=a}^T$$

for all $a \in \mathcal{A}$. The identity matrix $\mathbf{I}_d$ of dimension d is added to avoid the singularity of $\hat{\mathbf{W}}_a$ when the sample size is small. Then we iteratively update $\hat{\boldsymbol{\beta}}_a$ and $\hat{\mathbf{W}}_a$ for $a = A_i$ after each stage i by

$$\hat{\mathbf{W}}_a(\boldsymbol{H}_i) = \hat{\mathbf{W}}_a(\boldsymbol{H}_{i-1}) + \boldsymbol{X}_i \boldsymbol{X}_i^T, \qquad \hat{\boldsymbol{Y}}_a(\boldsymbol{H}_i) = \hat{\boldsymbol{Y}}_a(\boldsymbol{H}_{i-1}) + R_i \boldsymbol{X}_i$$

and let $\hat{\boldsymbol{\beta}}_a(\boldsymbol{H}_i) = \hat{\mathbf{W}}_a(\boldsymbol{H}_i)^{-1} \hat{\boldsymbol{Y}}_a(\boldsymbol{H}_i)$. The parameters for the treatment not selected at the stage i, which is $a = -A_i$, will not be updated at this stage.

When $\hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 0$, it means that the regression model prefers $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ than $-\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$. This is also the conclusion by the OWL model. Therefore, we are actually requiring the treatment to follow $\hat{\mathcal{D}}_{i-1}$ with high probability when the two models agree with each other. However, when the two models disagree, we assign treatment $-\hat{\mathcal{D}}_{i-1}$ with a soft probability based on the estimated benefit.

In LinUCB, the treatment is taken as $\text{sign}\{\hat{U}_1(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) - \hat{U}_{-1}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)\}$ with a probability 1, where the regression model is the ordinary least squares (OLS) model. This implies that $\mathbb{P}(I_i = 1) = \mathbb{P}(A_i = \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)) = \mathbb{1}[\hat{B}_i(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 0]$. The actual value of $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ does not really matter here since the probability is symmetric for the two treatments. If $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i) = 1$, then

$$\mathbb{P}(A_i = 1 | \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \mathbb{1}[\hat{U}_1(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) - \hat{U}_{-1}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 0];$$

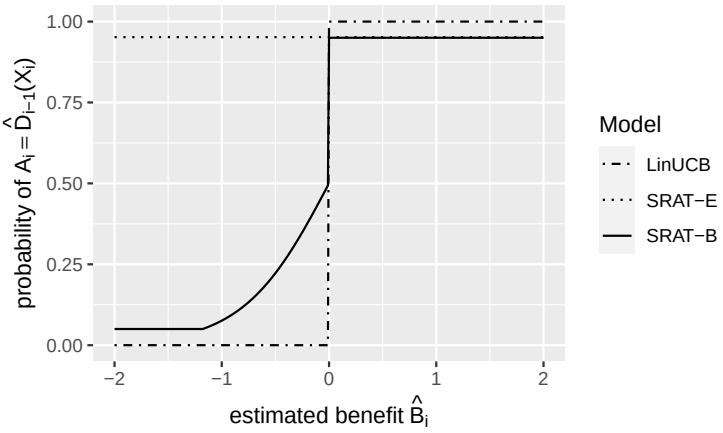

Figure 2.1: The randomization probability $\mathbb{P}(A_i = \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)|\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ of SRAT-E, SRAT-B and LinUCB when $\epsilon_i = 0.05$ and $\gamma_i = 0.4$.

otherwise, if $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i) = -1$, then

$$\begin{aligned}\mathbb{P}(A_i = -1|\boldsymbol{H}_{i-1}, \boldsymbol{X}_i) &= \mathbb{1}[\hat{U}_{-1}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) - \hat{U}_1(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) \geq 0] \\ &= 1 - \mathbb{P}(A_i = 1|\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)\end{aligned}$$

and they are equivalent if $\mathbb{P}(\hat{U}_1(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i) = \hat{U}_{-1}(\alpha_i, \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)) = 0$.

The relationship between SRAT-E, SRAT-B and LinUCB can be illustrated in Figure 2.1. While the randomization probability of SRAT-E is not affected by the estimated benefit of $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ over $-\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$, the probability of LinUCB is purely determined by this benefit. Note that the dot-dashed line is symmetric about zero for LinUCB, since the value of $\hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$ does not affect the probability of $A_i = 1$ as we discussed before. SRAT-B is a method that has an exploration probability in between, which actually approximates that of LinUCB when $\epsilon_i \to 0$ and $\gamma \to 0$. In this sense, our proposed variation of Boltzmann exploration is a soft version of UCB. If we also take OLS to be our model for estimating the benefit, we can view LinUCB as a limiting case of SRAT-B in the training process. However, since the treatment rules are learnt from OLS and OWL respectively, the test values are based on completely different final ITRs.

2.5 Simulation Study

We assess the empirical performance of SRAT on training and test samples using synthetic data. Here we examine two scenarios. In both scenarios, let $\boldsymbol{X}$ be a 10-dimensional vector $(X_1, X_2, \ldots, X_{10})$. Assume $\boldsymbol{X}$ has a joint distribution $N(0, \Sigma)$ truncated by $[-1, 1]$ for each dimension, where Σ is the covariance matrix with 1 on the diagonal and 0.1 off-diagonal. The treatment A is generated from $\{-1, 1\}$ according to the SRAT algorithm and other algorithms to be compared. Assume the reward R is normally distributed with mean $Q_0(\boldsymbol{X}, A) = m_0(\boldsymbol{X}) + T_0(\boldsymbol{X}, A)$ and variance $\nu_0(\boldsymbol{X}) = 0.2(X_1^2 X_3 + 1)$. Here m_0 is the main effect and T_0 is the treatment effect. The variance ν_0 is allowed to be a function of $\boldsymbol{X}$ to show that our proposed SRAT does not rely on the variance of rewards. We consider two scenarios as follows:

1. Linear treatment effect $T_0(\boldsymbol{X}, A) = 0.5(0.2 - X_1 - X_2)A$;

2. Nonlinear treatment effect $T_0(\boldsymbol{X}, A) = 0.5(0.2 - X_1^2 - X_2)A$.

In both scenarios, the main effect $m_0(\boldsymbol{X}) = 1 + 2X_1 + X_2^2 + 2X_2X_3$ is nonlinear. It can be easily seen that the optimal ITR is determined by $T_0(\boldsymbol{X}, A)$.

For our proposed SRAT algorithm, we first generate n_0 patients along with their purely randomized treatments and observed clinical outcomes. Then at each step i, we get a new sample of feature variables $\boldsymbol{X}_i$, and estimate its current optimal treatment by $\hat{\mathcal{D}}_{i-1}$. In accordance with our theoretical results in Example 2.1, we only use the linear kernel for OWL. The package `DTRlearn2` (Chen et al., 2019) is used to implement the OWL algorithm with L_2 penalty. It improves the learning performance by removing the main effect from the rewards and takes care of negative rewards by flipping the sign of the reward and the action simultaneously (Liu et al., 2018b). Next, we sample the binary indicator I_i with a probability $\{1 - p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i), p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)\}$ and take $A_i = I_i \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)$. Here $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ is defined in Section 2.4 for SRAT-E and SRAT-B differently, and the truncation parameter ϵ_n is defined as

$$\epsilon_0 n^{-(1-\theta)/4}, \quad \text{where } \epsilon_0 \in (0, 0.5], \theta \in (0, 1]$$

as in Theorem 2.3.5. The final estimated ITR decision function $\hat{f}_n$ will be evaluated using the test data. We also include RCT, estimated by OWL, as a special case of SRAT with $\epsilon_0 = 0.5$ and $\theta = 1$ in our simulation.

To compare our algorithm with existing bandit methods in linear scenario, we also implement LinUCB (Li et al., 2010) for demonstration. It is widely used in reinforcement learning and its variation SupLinUCB (Chu et al., 2011) is known to be rate optimal in contextual bandit problems with linear reward functions. LinUCB chooses the treatment with the largest upper confidence bound of reward, which is estimated by linear regression. It does not require a pilot trial for initialization, but we still generate one of size n_0 for it in consistency with our algorithm.

The active clinical trial (Minsker et al., 2016) is also compared here, which targets an effective ITR. Minsker et al. (2016) applied the active learning technique in the clinical trial and proposed to only conduct clinical trials on patients close to the decision boundary. In this way, patients that will benefit from one of the treatments with a high probability can be omitted from the trial and thus save experiment expenses and efforts. Minsker et al. (2016) considered two nonparametric methods, Gaussian process regression (AL-GP) and kernel smoothing (AL-BV) to construct a confidence interval around the decision boundary. The actually recruited patients are assigned to each treatment with equal probabilities. Since the two methods generally perform similarly in different scenarios, we only compare with AL-GP in our simulation study. AL-GP also requires a pilot trial, and we take n_0 as the initial sample size as well.

We fit each estimation model of the corresponding algorithm with linear terms of $X_1, \ldots, X_{10}$ for scenario 1, and with both linear and quadratic terms of $X_1, \ldots, X_{10}$ for scenario 2. While SRAT-E, RCT, LinUCB and AL-GP involve only one model, SRAT-B relies on both OWL and OLS models. Since at least 21 observations are needed to fit an initial model with 20 predictors and an intercept for OWL in scenario 2, we choose $n_0 = 30$ for both scenarios, which is almost the least possible for a reliable estimate of the initial rule. By comparing different values of n_0, we see that n_0 does not affect the results of SRAT-E and SRAT-B significantly. Larger n_0 reduces randomness but does not improve the average performance. This verifies the theoretical result that n_0 is not a dominating term in the theoretical bound as long as it has an order $O(n)$.

We have proved the properties of training and test regrets of SRAT in theoretical analysis and they will be used here as an indication of training and test performance of each algorithm. Each

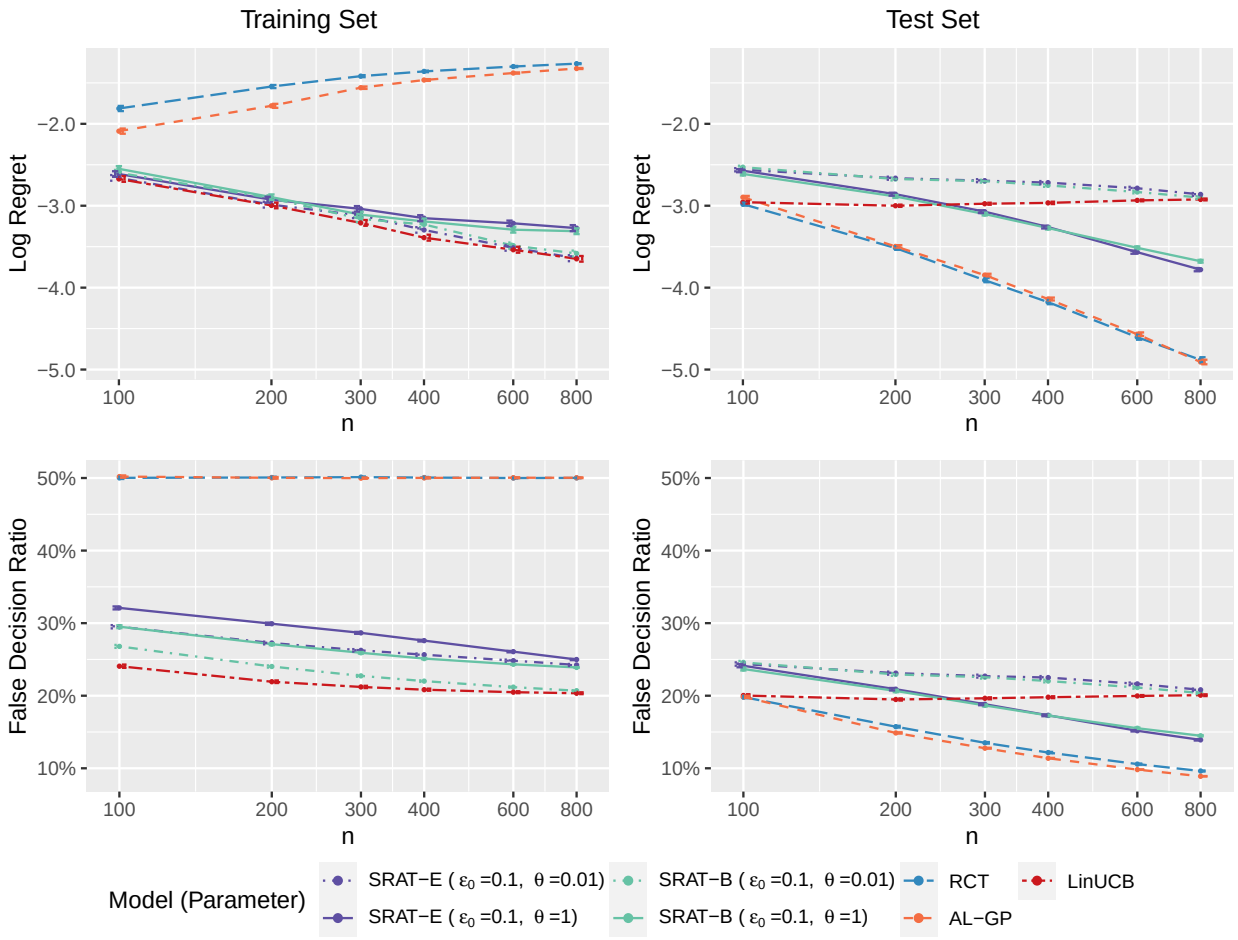

Figure 2.2: Scenario 1. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size n.

value function $\mathcal{V}$ is computed numerically using a sample of size $100,000$ randomly drew out of an independent population. The value function is estimated using the mean reward on this set.

We first compare the convergence rate of regret for different algorithms. SRAT-E and SRAT-B are implemented with $\epsilon_0 = 0.1$ and $\theta = 0.01$ or 1. As will be discussed later in Figure 2.4, the training and test regrets are monotone in the parameters ϵ_0 and θ. Therefore, to save space, we only show two possible combinations of parameters here. The scheduling parameter γ_i for SRAT-B is taken as 0.999^i so that it will not decay too fast to zero. RCT is a special case of SRAT with $\epsilon_0 = 0.5$ and $\theta = 1$. According to Li et al. (2010), the click-through rate (mean reward) of LinUCB in news article recommendation does not change much on the deployment bucket (test set) when $\alpha \geq 0.2$, while it decreases quickly on the learning bucket (training set) as α increases from 0.2. In our experiment settings, α does not affect training and test regrets significantly.

Therefore, we will fix $\alpha_i = 0.2$ for all i for LinUCB and SRAT-B in our following experiments. The process is repeated $1,000$ times and the resulting values are averaged across all iterations. To better illustrate the polynomial relationship between training or test regret and the sample size n, we plot the regret values and the sample sizes on the logarithmic scale. The false decision ratio, or $1 -$ accuracy in classification literature, is also displayed against n. One standard error of the mean regret or the mean false decision ratio across the $1,000$ iterations is reported on each point. The result of scenario 1 is plotted in Figure 2.2. The plot of scenario 2, Figure 2.7, is included in the supplementary material since it shows a similar conclusion as scenario 1.

According to Figure 2.2, LinUCB is the greediest on the training process, with the least regret and false decision ratio. As discussed in Section 2.4, LinUCB can actually be viewed as a limiting case of SRAT-B on the training set. Indeed, our proposed greediest algorithms, SRAT-E and SRAT-B with parameters $\epsilon_0 = 0.1$ and $\theta = 0.01$, perform similarly as LinUCB in terms of training regret. AL-GP and RCT take purely randomized treatments on the training set, so they have the largest training regret and a 50% training accuracy. Since the training regret is calculated based on $\mathcal{V}(\hat{f}_{n-1})$ which is increasing as n grows, the training regret actually increases for largely randomized methods. In theory, the training regret of RCT is bounded by a constant that does not rely on n when the ϵ-sequence is constant. SRAT-E and SRAT-B perform similarly in terms of regrets on both training and test sets, but SRAT-B has a lower false decision ratio on the training set. The logarithms of their training and test regrets are approximately linear in $\log n$, which is consistent with our theory.

On the test set, AL-GP and RCT perform the best due to their full exploration in the training process. LinUCB needs to fit the regression model of rewards and thus relies on both the main effect and the treatment effect model. In addition, to estimate the upper confidence bound, it needs an assumption on the inference model. With these limitations, the regret or false decision ratio of LinUCB on the test set does not decrease. When n is small, the final ITR estimated by LinUCB can sometimes be optimal since the true ITR is linear. However, the ITR converges to the projection onto that of the linear total reward space when n is large and thus the average regret gets pulled up. On the other hand, OWL tries to find the decision function that maximizes the reward directly. It only requires a correct model of the treatment effect for consistency, without

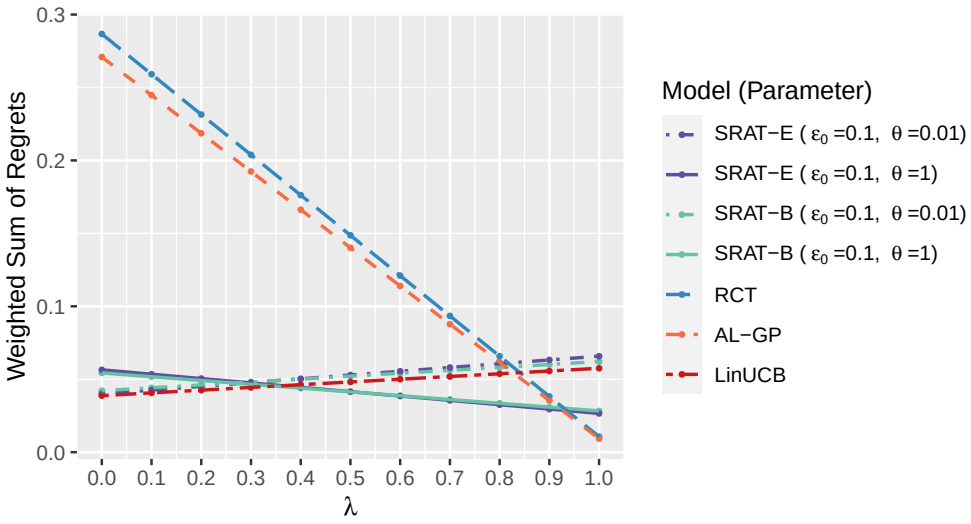

Figure 2.3: The weighted sum of training and test regrets in scenario 1 when $n = 800$.

any assumption on the main effect or the distribution of the error term. Therefore, SRATs with $\epsilon_0 = 0.1, \theta = 1$ outperform LinUCB on the test set when n is larger than 200.

We plot a weighted sum of training and test regrets in Figure 2.3 to show their balance. Specifically, the weighted sum is defined as

$$\lambda \text{Regret}_{test} + (1-\lambda)\text{Regret}_{train} = \lambda \frac{1}{n} \sum_{i=1}^{n} [\mathcal{V}(\hat{f}_{i-1}) - R_i] + (1-\lambda)[\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n)]$$

for $\lambda \in [0, 1]$, so that it equals the training regret when $\lambda = 0$ and equals the test regret when $\lambda = 1$. The sample size is fixed at 800. The initial value of truncation parameter ϵ_0 equals 0.1 and the decay parameter θ takes values in $0.01, 1$ for SRAT-E and SRAT-B. The plot shows that we should choose LinUCB when we consider the training regret only, and should choose AL-GP or RCT when we consider the test regret only. However, if we want to consider the performance on both the training and the test sets, we should choose SRAT-E or SRAT-B with $\theta = 1$.

The change of SRAT-E with different parameters θ and sample size n is demonstrated in Figure 2.4 for scenario 1. Since SRAT-B performs quite similarly to SRAT-E as shown in Figures 2.2 and 2.7, we omit it here to save space. The parameter θ can take values from $0.01, 0.1, 0.2, \ldots, 1$ and n can take values from $100, 200, 400, 800$. Note that only when $\epsilon_0 = 0.5$ and $\theta = 1$, our algorithm represents pure RCT. Thus we only illustrate our findings with $\epsilon_0 = 0.5$ here. Other ϵ_0's give similar conclusion, and smaller ϵ_0 means better training performance and worse test performance.

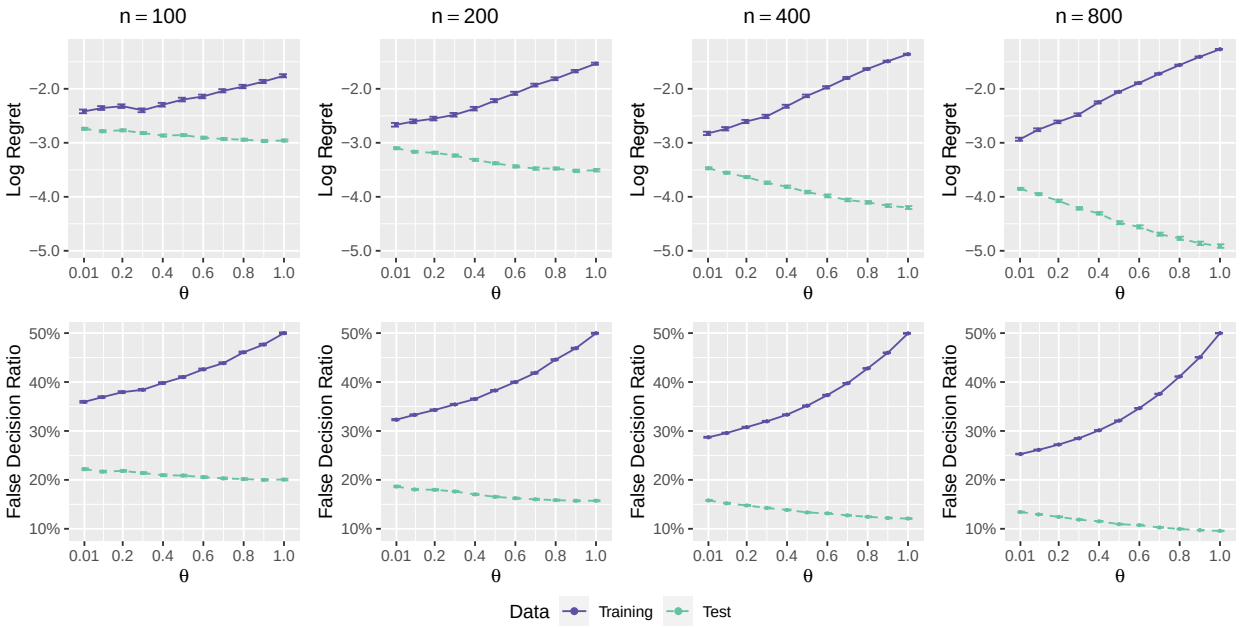

Figure 2.4: Scenario 1 with $\epsilon_0 = 0.5$. The regret (logarithmic scale) and the false decision ratio on the training or test set against parameter θ.

The values and standard errors of the mean regret and mean false decision ratio are shown. For all sample sizes, the plots clearly show the tradeoff between training and test performance. Note that when θ increases, ϵ_i increases for all i and the treatments are more randomized in the training process. While the training regret increases with more randomization, the test regret decreases. The false decision ratio shows a similar tendency. All the points with $\theta = 1$ have an accuracy of 50% on the training set, which indeed illustrates the pure randomization. In accordance with the theory, the logarithm of training and test regrets are approximately linear in θ. In practice, the training regret is more affected than the test regret by θ. As shown in Figure 2.4, when $n = 800$, the training regret increases by $e^{-1.27} - e^{-2.93} = 0.227$ while the test regret decreases by $e^{-3.85} - e^{-4.91} = 0.014$ when θ increases from 0.01 to 1.

Using this simulation example, we can also illustrate how to find the sample size needed for a clinical trial of certain purposes. Given different requirements for the trial and the population, we need different sample sizes. Here we illustrate the situation when the proportion of patients assigned the better treatments is required to reach a certain level in Figure 2.5 for SRAT-E in scenario 1. Note that the variation trends of correct decision ratios against θ are opposite for the training and test data. In particular, θ should be small enough so that the decision process is

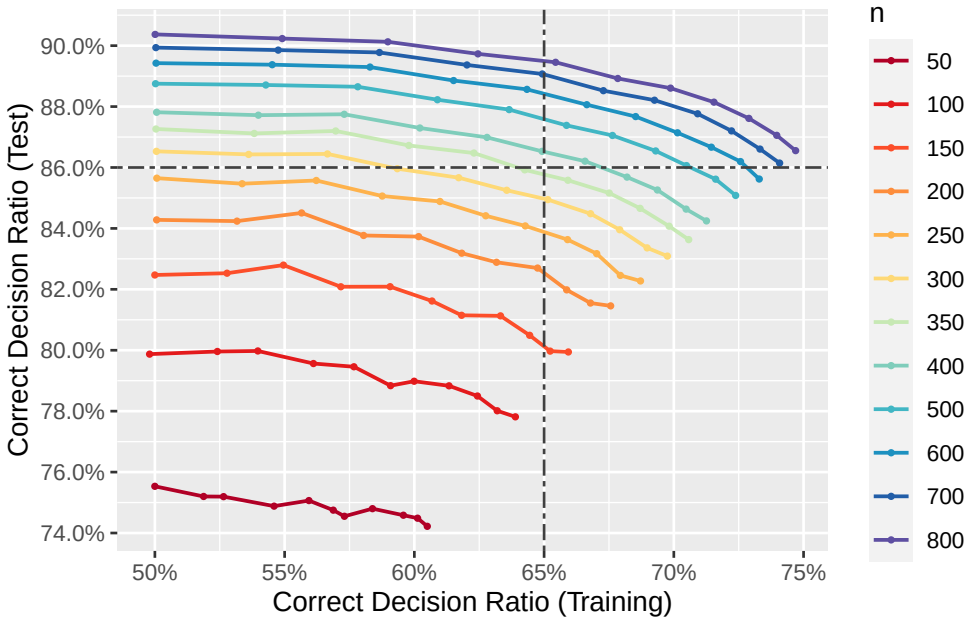

Figure 2.5: Sample size consideration for SRAT-E in scenario 1 with $\epsilon_0 = 0.5$. Correct decision ratios on the test set against that on the training set. Each line represents a sample size n and each point on the line represents a value of θ. Points to the right correspond to smaller θ, and thus lead to higher correct decision ratio on the training set and lower ratio on the test set.

greedy on the training set, and in the meanwhile it should be large enough so that the final ITR is efficient on the test set. It is clear that the two accuracies are negatively correlated. For example, when we need the training ratio to be greater than 65%, $\theta \leq 0.1$ for $n = 150$, $\theta \leq 0.2$ for $n = 200$, $\theta \leq 0.3$ for $n = 250$, $\theta \leq 0.4$ for $n = 300$, $\theta \leq 0.4$ for $n = 350$, $\theta \leq 0.4$ for $n = 400$, $\theta \leq 0.5$ for $n = 500$, $\theta \leq 0.5$ for $n = 600$, $\theta \leq 0.5$ for $n = 700$, and $\theta \leq 0.6$ for $n = 800$ will all do. When we need the test ratio to be greater than 86%, $\theta \geq 0.8$ for $n = 300$, $\theta \geq 0.6$ for $n = 350$, $\theta \geq 0.4$ for $n = 400$, $\theta \geq 0.2$ for $n = 500$, $\theta \geq 0.1$ for $n = 600$, any θ for $n = 700$, and any θ for $n = 800$ all satisfy the requirement. However, only points lie in the top right rectangle marked by the two dot-dashed lines meet the two requirements simultaneously. The smallest sample size among these points is $n = 400$, with $\theta = 0.4$. Other levels of the correct decision ratios and their required sample sizes are listed in Table 2.1. Since larger θ generates better ITR and ITR is our ultimate goal, we report the largest θ corresponding to the minimum sample size required.

Training	Test				
	0.74	0.78	0.82	0.86	0.90
0.49	50(1.0)	100(1.0)	150(1.0)	300(1.0)	800(1.0)
0.55	50(0.6)	100(0.7)	150(0.7)	300(0.8)	800(0.8)
0.60	50(0.1)	100(0.3)	200(0.6)	350(0.6)	
0.65	150(0.1)	150(0.1)	250(0.3)	400(0.4)	
0.70	350(0.01)	350(0.01)	350(0.01)	500(0.2)	

Table 2.1: Clinical trial sample sizes needed for different requirements of correct decision ratios on the training and test sets.

2.6 Real Data Analysis

We use a real study to illustrate the performance of the proposed method. The Nefazodone-CBASP trial was designed to compare the efficacy of several treatment options for patients with nonpsychotic chronic major depressive disorder (MDD) (Keller et al., 2000). Specifically, 681 outpatients were randomized to either Nefazodone, Cognitive Behavioral-Analysis System of Psychotherapy (CBASP), or the combination of Nefazodone and CBASP with equal probabilities. The primary outcome was the score on the 24-item Hamilton Rating Scale for Depression (HRSD). Lower HRSD scores indicate satisfactory therapeutic efficacy. T-tests have shown that the combination treatment generated significantly lower HRSD scores than the other two treatments, and there are no significant differences between the Nefazodone group and the CBASP group. However, CBASP requires two onsite visits to the clinic weekly, which burdens patients compared with Nefazodone alone. Consequently, we want to investigate whether CBASP is necessary for all patients. Here we compare Nefazodone with the combination treatment only. We consider three feature variables for treatment suggestions: the baseline HRSD scores, the alcohol dependence, and the HAMA somatic anxiety scores, following Minsker et al. (2016), which referred to Gunter et al. (2007). There were 436 patients with complete information on treatments, rewards and feature variables, among which 216 were randomized to Nefazodone and 220 belonged to the combined treatment group.

To simulate an adaptive clinical trial, we first generate a treatment suggestion based on the tailoring variables of the next patient using our algorithm. If the actual treatment taken is consistent with our suggestion, we take down the whole record of this patient, including feature variables, the treatment and the reward; otherwise, we drop this record and move on to the next. Note that the

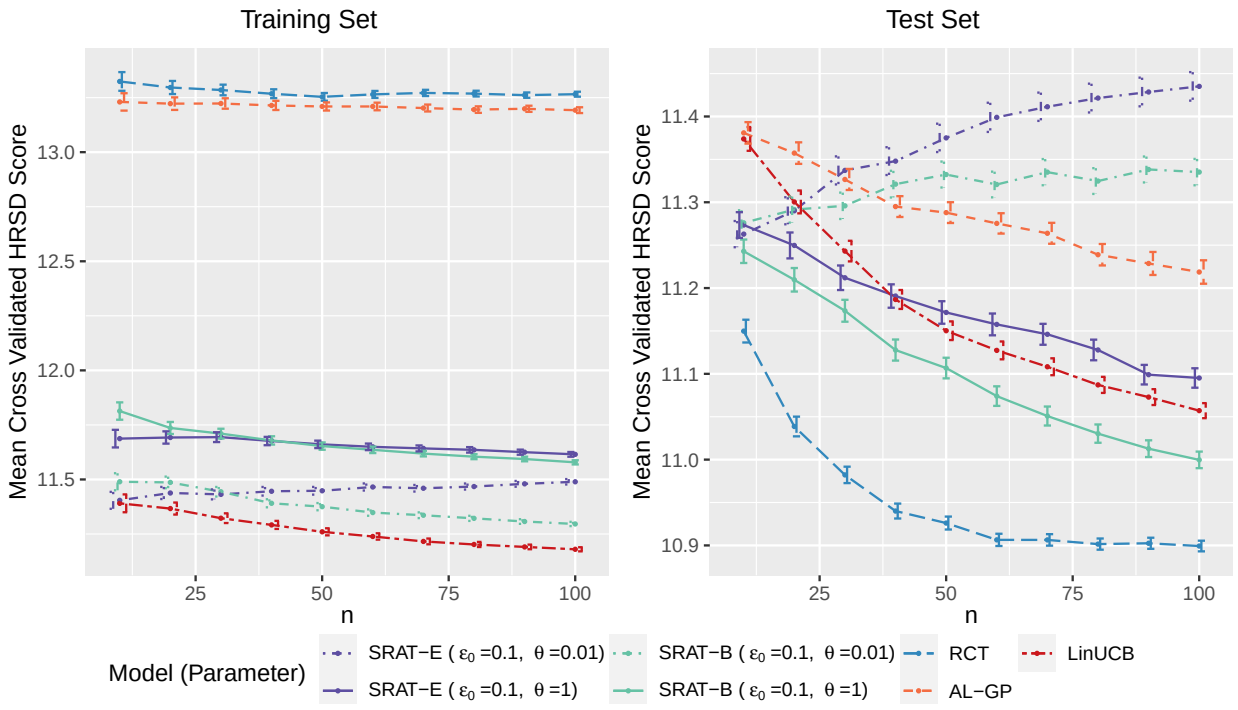

Figure 2.6: Mean cross-validated HRSD scores against the sample size n.

first n_0 suggestions are given with equal probabilities on each treatment. Five-fold cross validation is used here to avoid overfitting. Specifically, the data set is partitioned into five parts randomly. Four of the five parts are used iteratively as training data to apply our algorithm in generating the treatment suggestion. The last part is used as the test set to evaluate the ITR. The performance on the test data is evaluated using an unbiased estimator of the value function $\mathcal{V}(f)$ (Qian and Murphy, 2011; Minsker et al., 2016)

$$\sum_{i=1}^{n} \frac{R_i \mathbb{1}\left[A_i = \text{sign}\{f(\boldsymbol{X}_i)\}\right]}{\pi_i(A_i; X_i)} \Bigg/ \sum_{i=1}^{n} \frac{\mathbb{1}\left[A_i = \text{sign}\{f(\boldsymbol{X}_i)\}\right]}{\pi_i(A_i; X_i)} .$$

Here the rewards R_i's are defined as the negative HRSD scores.

The initial sample size n_0 is fixed at 50. The recruitment stops when the sample size n reaches 100, or the training data run out. We average the mean reward on each test fold for $n = 10, 20, \ldots, 100$. The process is repeated $1,000$ times. Finally, the means and standard errors of means across all iterations are reported. From Section 2.5, we know that the training and test values are monotone in ϵ_0 and θ. Therefore, we only demonstrate the situation when $\epsilon_0 = 0.1$ and $\theta = 0.01, 1$. The contextual bandit algorithm LinUCB and the active clinical trial method AL-GP

are also compared here. Figure 2.6 displays the negative mean rewards, that is, the mean cross validated HRSD scores, against the sample size n. Lower scores are more satisfactory.

On the training set, LinUCB produces the least HRSD scores on the training set and SRAT-B with $\epsilon = 0.1, \theta = 0.01$ is the second best. Note that LinUCB can be viewed as a limiting case of SRAT-B on the training set as discussed before, and is actually the greediest among the algorithm family. Patients taking purely randomized treatments suggested by RCT and AL-GP have higher HRSD scores. On the test set, RCT produces the most desirable HRSD score, followed by SRAT-B with $\theta = 1$. LinUCB is slightly worse due to its greediness. AL-GP is not competitive on both sets, maybe because the nonparametric method is not efficient when the sample size is small.

2.7 Discussion

Our goal is to construct an efficient ITR, and in the meantime make the data collection process, the clinical trial, as beneficial to the patients as possible. We propose a classification-based bandit algorithm, SRAT, that uses OWL to update the ITR and ϵ-greedy or a variation of Boltzmann exploration for exploration. This is a work of finding the tradeoff between the ethics of patients involved in the clinical trial and the general population. We also present a new theoretical analysis tool based on empirical process for estimating finite sample risk bound on martingale sequences. Given different requirements of training and test performance, the sample size needed is illustrated by simulation.

In this chapter, we assume that the true optimal decision function lies in the function class where we search for the estimated function, and proved a $n^{-1/2}$ convergence rate of test regret up to logarithmic factors for a constant $\{\epsilon_i\}_{i=1}^n$ sequence. If tailoring variables have high dimensions, a penalty term can be added in finding the optimal solution to avoid overfitting. For i.i.d. data, when Gaussian kernel is used with a penalty term and the optimal function need not be in the function class, Steinwart and Scovel (2007) proved a rate faster than $n^{-1/2}$ for SVM under Tsybakov's noise assumption and geometric noise assumption, and Zhao et al. (2012) proved a rate a little bit slower than $n^{-1/2}$ for OWL under the geometric noise assumption. How to extent these ideas to sequentially generated data is still an open question.

Currently, the estimated ITR is updated after each trial. However, it can be a burden on the computation resources and running time if the algorithm runs slowly or the sample size is too large. Batch sampling is an efficient approach that worths investigation. Apart from accelerating the training process, it also allows in-time evaluation of the current estimated optimal ITR. Part of the batch can be drew randomly as a test set. How the estimation improves through time can be recorded as well.

Another interesting question is how to set up an early stopping rule. We can stop enrolling new patients into a clinical trial if the learnt ITR is good enough. This can be done by constructing a confidence interval for the estimated value of the learnt ITR. If we have enough confidence that the estimated value is satisfactory in the clinical sense, we can stop the trial at this point. Future work is needed on constructing a confidence interval for sequentially generated data.

This article focuses on a single-stage problem. However, it is widely recognized that some diseases require multiple treatments throughout the therapeutic session. For example, the sequential multiple assignment randomized trial (SMART) is a way of connecting potential outcomes with observed data (Lavori and Dawson, 2000; Murphy, 2005a; Murphy et al., 2007). Patients are randomized at every decision point. An abundance of literature has discussed this issue on independent data (Zhao et al., 2015; Liu et al., 2018b). Problems on infinite horizon can be solved with additional Markovian assumptions and offline data (Luckett et al., 2020). However, multi-stage decision problems with slack constraints on the value function or with online data still worth investigation.

2.8 Supplementary Materials

2.8.1 Preliminaries

We provide useful lemmas used for proving our main theorems, among which Lemmas 2.8.1 to 2.8.4 are quoted from existing literature without proof.

Talagrand's inequality (Talagrand, 1994; Bousquet, 2002) below is used to prove the convergence rate in Theorem 2.3.1 for i.i.d. data in the pilot trial. The following version is taken from Steinwart and Scovel (2007, Theorem 5.3).

Lemma 2.8.1 (Talagrand's inequality)**.** *Assume* $\{\boldsymbol{X}_i\}_{i=1}^n$ *are independent* $\mathcal{X}$*-valued random variables on* $\mathbb{P}$*. Let* $\mathcal{F}$ *be a countable set of functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *and assume that all functions* f *in* $\mathcal{F}$

are $\mathbb{P}$-measurable, square-integrable such that $\|f\|_\infty \le U < \infty$ and $\mathbb{E}[f(\boldsymbol{X}_1)] = \cdots = \mathbb{E}[f(\boldsymbol{X}_n)] = 0$. Let $Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(\boldsymbol{X}_i)$, $\mu^* := \mathbb{E}Z$ and let σ^2 be a positive real number such that $\sigma^2 \ge \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathrm{Var}[f(\boldsymbol{X}_i)]$. Then for all $\delta \ge 0$,

$$\mathbb{P}\left(Z \ge 3\mu^* + \sqrt{2\delta\sigma^2 n} + U\delta\right) \le e^{-\delta}.$$

The following lemma gives an analogy of Talagrand's inequality on martingale processes. A $\mathcal{Z}$-valued tree $\boldsymbol{z}$ of depth n is a rooted complete binary tree with nodes generated by elements of $\mathcal{Z}$. The tree $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ is a sequence of labeling functions such that $\boldsymbol{z}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{Z}$. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ be a sequence of i.i.d. Rademacher random variables. Then the sequential Rademacher complexity of a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree $\boldsymbol{z}$ is defined as

$$\mathcal{R}_n(\mathcal{F}, \boldsymbol{z}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \eta_i f(\boldsymbol{z}_i(\boldsymbol{\eta}))\right].$$

Further, define

$$\mathcal{R}_n(\mathcal{F}) := \sup_{\boldsymbol{z}} \mathcal{R}_n(\mathcal{F}, \boldsymbol{z}),$$

and Rakhlin et al. (2015) showed that

$$\frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{M}_n(f) \le \mathcal{R}_n(\mathcal{F}) \le 2\sup_{\mathbb{P}} \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{M}_n(f) + \frac{D}{2\sqrt{n}}, \tag{2.8}$$

where $D = \inf_{z\in\mathcal{Z}} \sup_{f,f'\in\mathcal{F}} [f(z) - f'(z)] \ge 0$. This indicates that $\mathcal{R}_n(\mathcal{F})$ and the expectation of the martingale process suprema $\sup_{\mathbb{P}} \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{M}_n(f)$ are on the same scale.

The covering numbers are also extended to sequential data. A set V of $\mathbb{R}$-valued trees of depth n is a (sequential) ϵ-cover with respect to L_p-norm of $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ on a tree z of depth n if for any $f \in \mathcal{F}$ and any $\eta \in \{\pm 1\}^n$, there exists $\boldsymbol{v} \in V$ such that $\left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{v}_i(\eta) - f(\boldsymbol{z}_i(\eta))|^p\right)^{1/p} \le \epsilon$. The sequential covering number of a function class $\mathcal{F}$ on a given tree $\boldsymbol{z}$ is defined as

$$\mathcal{N}_p(\epsilon, \mathcal{F}, \boldsymbol{z}) = \min\left\{|V| : \text{ V is an } \epsilon\text{-cover with respect to } L_p\text{-norm of } \mathcal{F} \text{ on } \boldsymbol{z}\right\}.$$

Moreover, define the maximal L_p covering number of $\mathcal{F}$ over depth-n trees as $\mathcal{N}_p(\epsilon, \mathcal{F}, n) = \sup_{\boldsymbol{z}} \mathcal{N}_p(\epsilon, \mathcal{F}, \boldsymbol{z})$.

Lemma 2.8.2 (Lemma 15 in Rakhlin et al. 2015). *For $\mathcal{F} \subset [-1,1]^{\mathcal{Z}}$, for $n \geq 2$ and any $t > 0$, we have that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_{i-1}f(Z_i)\right| > t\right) \leq 8L\exp\left(-\frac{t^2}{c\log^3 n\mathcal{R}_n^2(\mathcal{F})}\right) \tag{2.9}$$

under the mild assumptions $\mathcal{R}_n(\mathcal{F}) \geq 1/n$ and $\mathcal{N}_\infty(2^{-1},\mathcal{F},n) \geq 4$. Here c is an absolute constant and $L > e^4$ is such that $L > \sum_{j=1}^{\infty}\mathcal{N}_\infty(2^{-j},\mathcal{F},n)^{-1}$.

From the above lemma, we can see that the concentration inequality essentially relies on the sequential Rademacher complexity $\mathcal{R}_n(\mathcal{F})$, which can be upper and lower bounded by functions of $\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$.

To obtain a bound on the suprema of martingale process over a finite set, we take use of a martingale inequality and a conclusion about L_ψ-Orlicz norm. Freedman's inequality is an extension of Bernstein's inequality to martingale difference sequences.

Lemma 2.8.3 (Freedman's inequality, Freedman 1975). *Suppose $\{X_i\}_{i\geq 1}$ is a $\mathcal{G}_i$-adapted martingale difference sequence and $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\mathbb{P}(S_n \geq t) \leq \exp\left\{-\frac{1}{2}\frac{t^2}{\|\langle S\rangle_n\|_\infty + \max_i\|X_i\|_\infty t/3}\right\},$$

where $\langle S\rangle_n = \sum_{i=1}^n \mathbb{E}\left(X_i^2|\mathcal{G}_{i-1}\right)$ is the quadratic variation of S.

The following lemma gives a bound on the expectation of suprema over a finite set using L_ψ-Orlicz norm.

Lemma 2.8.4 (Van der Vaart and Wellner 1996). *Suppose that $X_1,\ldots,X_n$ are arbitrary random variables satisfying the probability tail bound*

$$\mathbb{P}(|X_i| > t) \leq 2\exp\left\{-\frac{1}{2}\frac{t^2}{d+cx}\right\},$$

for all $t > 0$ and $i = 1,\ldots,n$ for fixed positive numbers c and d. Then there is a universal $K < \infty$ so that

$$\left\|\max_{1\leq i\leq n}|X_i|\right\|_{\psi_1} \leq K\left\{c\,\mathrm{Log}\,n + \sqrt{d}\sqrt{\mathrm{Log}\,n}\right\},$$

where the L_ψ-Orlicz norm is defined as $\|X\|_\psi = \inf\{c > 0 : \mathbb{E}\psi(|X|/c) \leq 1\}$ for any random variable X, and $\psi_p = e^{x^p} - 1$ is a Young modulus for each $p \geq 1$.

The following inequality is a key step in the proof of Lemma 2.3.2. It bounds the expectation of the suprema of a martingale process over a finite set, after which the bound on a general set can be derived.

Corollary 2.8.5. *Suppose $\{X_i\}_{i\geq 1}$ is a $\mathcal{X}$-valued, $\mathcal{G}_i$-adapted martingale difference sequence. For any finite set $\mathcal{F}: \mathcal{X} \mapsto \mathbb{R}$,*

$$\mathbb{E}\left\|\sqrt{n}\mathbb{M}_n\right\|_{\mathcal{F}} \lesssim \frac{1}{\sqrt{n}} \max_{f\in\mathcal{F}} \|f\|_\infty \operatorname{Log}|\mathcal{F}| + \frac{1}{\sqrt{n}} \max_{f\in\mathcal{F}} \sqrt{\|\langle M\rangle_n\|_\infty} \sqrt{\operatorname{Log}|\mathcal{F}|},$$

where $\langle M\rangle_n = \sum_{i=1}^n \operatorname{Var}[f(X_i)|\mathcal{G}_{i-1}]$.

Proof. First we rewrite Lemma 2.8.3 in the form of a martingale process. For a $\mathcal{G}_i$-adapted sequence $\{Y_i\}_{i\geq 1}$, take $\mathbb{E}[f(Y_i)|\mathcal{G}_{i-1}] - f(Y_i)$ as X_i in Lemma 2.8.3, which is a martingale difference sequence. The supremum term can be bounded as $\|\mathbb{E}[f(Y_i)|\mathcal{G}_{i-1}] - f(Y_i)\|_\infty \leq 2\|f\|_\infty$. Scale both sides by a factor of $\sqrt{n}$ and we get

$$\mathbb{P}(\left|\sqrt{n}\mathbb{M}_n(f)\right| > t) \leq 2\exp\left\{-\frac{1}{2}\frac{t^2}{\|\langle M\rangle_n\|_\infty / n + 2\|f\|_\infty t/(3\sqrt{n})}\right\}, \tag{2.10}$$

where $t > 0$ and $\langle M\rangle_n = \sum_{i=1}^n \operatorname{Var}[f(Y_i)|\mathcal{G}_{i-1}]$. The result follows by applying the inequality (2.10) to Lemma 2.8.4 and expand the L_ψ-Orlicz norm. □

The following lemma shows how the dependence of π_i on $\hat{f}_{i-1}$ can be canceled by the sampling probability so that it reduces to a constant term.

Lemma 2.8.6. *Under our problem settings, remember that the covariates $\boldsymbol{X}_i$ are i.i.d. Besides, $\mathcal{G}_i$ is defined as $\sigma\{\boldsymbol{H}_i\}, i \in \mathbb{N}$ and $\hat{f}_{i-1}$ is the estimated ITR based on $\boldsymbol{H}_{i-1}$. For any function $G: \mathcal{X} \mapsto \mathbb{R}$, we have*

$$\mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\right] = \mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right] = 2\mathbb{E}[G(\boldsymbol{X}_i)], \tag{2.11}$$

$$\mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\right] = \mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right] \leq \left(\frac{1}{\epsilon_i} + \frac{1}{0.5}\right)\mathbb{E}[G(\boldsymbol{X}_i)]. \tag{2.12}$$

Proof. For the first equation (2.11), notice that by tower property and the definition of I_i,

$$\begin{aligned}
&\mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right]\\
=&\mathbb{E}\left\{G(\boldsymbol{X}_i)\mathbb{E}\left[\frac{\mathbb{1}\left(I_i=1\right)}{\pi_i(\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}+\frac{\mathbb{1}\left(I_i=-1\right)}{\pi_i(-\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\middle|\boldsymbol{X}_i,\mathcal{G}_{i-1}\right]\middle|\mathcal{G}_{i-1}\right\}
\end{aligned}$$

We have that $\pi_i(\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)$ is a fixed function of $\boldsymbol{H}_{i-1},\boldsymbol{X}_i$ and that it equals $\mathbb{E}[\mathbb{1}\left(I_i=1\right)|\boldsymbol{H}_{i-1},\boldsymbol{X}_i]$. It is also true for the second term in the bracket. Therefore, the right-hand side equals $\mathbb{E}\left[2G(\boldsymbol{X}_i)|\mathcal{G}_{i-1}\right]$. The result follows from the assumption that $\boldsymbol{X}_i$ is independent of the history. The first equality in (2.11) can be proved by taking expectation of both sides of the equation.

Similarly, when G is divided by the square term of π_i in (2.12),

$$\begin{aligned}
&\mathbb{E}\left[\frac{G(\boldsymbol{X}_i)}{\pi_i^2(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right]\\
=&\mathbb{E}\left\{G(\boldsymbol{X}_i)\mathbb{E}\left[\frac{\mathbb{1}\left(I_i=1\right)}{\pi_i^2(\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}+\frac{\mathbb{1}\left(I_i=-1\right)}{\pi_i^2(-\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\middle|\boldsymbol{X}_i,\mathcal{G}_{i-1}\right]\middle|\mathcal{G}_{i-1}\right\}\\
=&\mathbb{E}\left\{G(\boldsymbol{X}_i)\left[\frac{1}{\pi_i(\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}+\frac{1}{\pi_i(-\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\right]\middle|\mathcal{G}_{i-1}\right\}.
\end{aligned}$$

Since one of $\pi_i(\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)$ and $\pi_i(-\hat{f}_{i-1}(\boldsymbol{X}_i);\boldsymbol{H}_{i-1},\boldsymbol{X}_i)$ must be lower bounded by 0.5 and the other one is lower bounded by ϵ_i, the inequality follows. Now for the first term in (2.12), its upper bound can be proved by taking expectation of both sides of the inequality. □

2.8.2 Proof of Lemma 2.3.2

The proof essentially follows the proof of Van der Vaart and Wellner (1996, Theorem 2.5.6), the bracketing entropy Donsker theorem, with an extension to martingale sequences.

In this proof, we assume that there exists a constant τ^2 such that the second conditional moment $\mathbb{E}(R^2|\boldsymbol{X},A)\leq\tau^2$. This also implies that the conditional variance $\mathrm{Var}(R|\boldsymbol{X},A)$ is bounded by τ^2. However, we will show that τ^2 does not appear in the dominating term of the final bound.

Proof. Define $L_{2,\infty}(\mathbb{P})$ norm as $\|f\|_{\mathbb{P},2,\infty}=\sup_{x>0}[x^2\mathbb{P}(|f(\boldsymbol{X})|>x)]^{1/2}$. Note that $L_{2,\infty}(\mathbb{P})$ norm is not actually a norm, but it can be shown that there is a norm equivalent to it up to a constant

multiple. The assumption (2.4) implies that

$$\int_0^\infty \sqrt{\log N_{[]}(\eta, \mathcal{F}, L_{2,\infty}(\mathbb{P}))}d\eta + \int_0^\infty \sqrt{\log N(\eta, \mathcal{F}, L_2(\mathbb{P}))}d\eta < \infty,$$

because $\|f\|_{\mathbb{P},2} \geq \|f\|_{\mathbb{P},2,\infty}$ for any measurable function f, and we have $N_{[]}(\eta, \mathcal{F}, L_2(\mathbb{P})) \geq N(\eta, \mathcal{F}, L_2(\mathbb{P}))$ for any function class $\mathcal{F}$.

For each positive integer q, define a bracketing number $N_q^1 := N_{[]}(2^{-q}, \mathcal{F}, L_{2,\infty}(\mathbb{P}))$ and a covering number $N_q^2 := N(2^{-q}, \mathcal{F}, L_2(\mathbb{P}))$. Then there are two partitions $\{\mathcal{F}_{qj}\}_{j=1}^{N_q^1}$ and $\{\mathcal{F}_{qk}\}_{k=1}^{N_q^2}$ of $\mathcal{F}$ into disjoint sets such that $\sum_q 2^{-q}\sqrt{\log N_q^1} < \infty$ and $\sum_q 2^{-q}\sqrt{\log N_q^2} < \infty$. Take intersection of the two partitions that correspond to the bracketing number and covering number respectively. The total number of sets will be $N_q := N_q^1 N_q^2$ and this joint partition $\{\mathcal{F}_{qj}\}_{j=1}^{N_q}$ satisfies the combined conditions:

$$\sum_q 2^{-q}\sqrt{\log N_q} < \infty, \tag{2.13}$$

$$\left\| (\sup_{f,g\in\mathcal{F}_{qj}} |f-g|)^* \right\|_{\mathbb{P},2,\infty} < 2^{-q}, \quad \forall j \in \{1, \ldots, N_q\}, \tag{2.14}$$

$$\sup_{f,g\in\mathcal{F}_{qj}} \|f-g\|_{\mathbb{P},2} < 2^{-q}, \quad \forall j \in \{1, \ldots, N_q\}. \tag{2.15}$$

Furthermore, the sequence of partitions can be chosen to be nested. To see this, consider a sequence of partitions $\{\bar{\mathcal{F}}_{qj}\}_{j=1}^{\bar{N}_q}$ that are possibly not nested. Take the partition at stage q to consist of all intersections of the form $\bigcap_{p=1}^q \bar{\mathcal{F}}_{p,i_p}$. Then this generates $N_q = \bar{N}_1 \ldots \bar{N}_q$ sets. Conditions (2.13) - (2.15) continue to hold since $(\log \prod_{p=1}^q \bar{N}_p)^{1/2} \leq \sum_{p=1}^q (\log \bar{N}_p)^{1/2}$.

Now for each q, fix a function $f_{qj} \in \mathcal{F}_{qj}$ to be the representative of the set $\mathcal{F}_{qj}$ and let ξ be the function of choosing the representative. In addition, let Δ be the function of finding the "size" of the set that a function belongs to. Then we have

$$\xi_q f := \sum_j I(f \in \mathcal{F}_{qj}) f_{qj}, \quad \text{and} \quad \Delta_q f := \sum_j I(f \in \mathcal{F}_{qj}) \sup_{f_1, f_2 \in \mathcal{F}_{qj}} |f_1 - f_2|^*.$$

For the weighted function h^f of f, define

$$\xi_q h^f := h^{\xi_q f}, \quad \text{and} \quad \Delta_q h^f := \sum_j I(f \in \mathcal{F}_{qj}) \sup_{f_1, f_2 \in \mathcal{F}_{qj}} \left| h^{f_1} - h^{f_2} \right|^*.$$

Note that $\Delta_q h^f = \sum_j I(f \in \mathcal{F}_{qj}) \sup_{f_1, f_2 \in \mathcal{F}_{qj}} \left| g^{f_1} - g^{f_2} \right|^*$ and $|\phi(Af_1) - \phi(Af_2)| \leq |f_1 - f_2|$ since $\phi(Af)$ is Lipschitz 1 with respect to f. Hence we obtain that $|R_i| \, \Delta_q h^f(\boldsymbol{Z}_i) \leq \Delta_q f(\boldsymbol{X}_i)/\pi_i(A_i; \hat{f}_{i-1})$. Note that $\xi_q h^f$ and $\Delta_q h^f$ form sets of only N_q functions when h^f ranges over $\mathcal{F}$. We will actually approximate each h^f with $\xi_q h^f$ and $\Delta_q h^f$. While $\mathcal{F}$ may be infinite, $\xi_q h^f$ and $\Delta_q h^f$ run over finite sets.

Let $\mathrm{Log}(x) := 1 + \log(x)$. For each fixed n and q_0, define truncation levels a_q and indicator functions A_q, B_q for $q \geq q_0$ as

$$\begin{aligned}
a_q &= 2^{-q}/\sqrt{\mathrm{Log}\, N_{q+1}}, \quad \forall q \geq q_0, \\
A_{q-1} f &= \mathbb{1}\left\{\Delta_{q_0} f \leq \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \leq \sqrt{n} a_{q-1}\right\}, \quad \forall q > q_0, \\
B_q f &= A_{q-1} f \mathbb{1}\left\{\Delta_q f > \sqrt{n} a_q\right\}, \quad \forall q > q_0, \\
B_{q_0} f &= \mathbb{1}\left\{\Delta_{q_0} f > \sqrt{n} a_{q_0}\right\}.
\end{aligned}$$

Since the partitions are nested, the functions A_q and B_q are constants in f on each set $\mathcal{F}_{qj}$ in level q. The key observation here is that

$$h^f - \xi_{q_0} h^f = (h^f - \xi_{q_0} h^f) B_{q_0} f + \sum_{q=q_0+1}^{\infty} (h^f - \xi_q h^f) B_q f + \sum_{q=q_0+1}^{\infty} (\xi_q h^f - \xi_{q-1} h^f) A_{q-1} f \qquad (2.16)$$

pointwise in x. To see this, note that either $B_q f = 0$ for all q or there is a unique q_1 such that $B_q f = 1$. In the former case, the first two terms are all zero and the third term has canceling components and converges to $f - \xi_{q_0} f$. In the latter case, the right-hand side of (2.16) is equivalent to $h^f - \xi_{q_1} h^f + \sum_{q=q_0+1}^{q_1} (\xi_q h^f - \xi_{q-1} h^f)$, and the result follows.

Write $\|\mathbb{M}_n(f)\|_{\mathcal{F}}$ as the supremum of $|\mathbb{M}_n(f)|$ as f ranges over $\mathcal{F}$. Then $\mathbb{E}^* \sup_{f\in\mathcal{F}} \mathbb{W}_n(f)$ can be bounded as

$$\begin{aligned}
\mathbb{E}^* \left\|\sqrt{n}\mathbb{W}_n(f)\right\|_{\mathcal{F}} \leq & \mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n(h^f - \xi_{q_0}h^f)B_{q_0}f\right\|_{\mathcal{F}} && (2.17)\\
& + \mathbb{E}^* \left\|\sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n(h^f - \xi_q h^f)B_q f\right\|_{\mathcal{F}} && (2.18)\\
& + \mathbb{E}^* \left\|\sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n(\xi_q h^f - \xi_{q-1}h^f)A_{q-1} f\right\|_{\mathcal{F}} && (2.19)\\
& + \mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n \xi_{q_0} h^f\right\|_{\mathcal{F}}. && (2.20)
\end{aligned}$$

To bound the first term (2.17), note that for any function class $\mathcal{H}$ with some envelope function H, $|\mathbb{M}_n(h)| \leq \frac{1}{n}\sum_{i=1}^n (\mathbb{E}_{i-1}H(\boldsymbol{Z}_i) + H(\boldsymbol{Z}_i))$ for all $h \in \mathcal{H}$. Then we have

$$\mathbb{E}^* \|\mathbb{M}_n(h)\|_{\mathcal{H}} \leq \mathbb{E}^* \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{i-1}H(\boldsymbol{Z}_i) + H(\boldsymbol{Z}_i)\right\|_{\mathcal{H}} = \frac{2}{n}\sum_{i=1}^n \mathbb{E}^* H(\boldsymbol{Z}_i).$$

An envelope function of $(h^f - \xi_{q_0}h^f)B_{q_0}f$ is

$$\frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \left|h^f - \xi_{q_0}h^f\right| B_{q_0}f \leq \frac{r}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} 2F \mathbb{1}\left\{2F > \sqrt{n}a_{q_0}\right\}$$

by the definitions of envelope function F and indicator function B_{q_0}. Therefore,

$$\begin{aligned}
\mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n(h^f - \xi_{q_0}h^f)B_{q_0}f\right\|_{\mathcal{F}} \leq & \frac{2}{\sqrt{n}}\sum_{i=1}^n \mathbb{E}^* \frac{r}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} 2F(\boldsymbol{X}_i)\mathbb{1}\left\{2F(\boldsymbol{X}_i) > \sqrt{n}a_{q_0}\right\}\\
= & \frac{4r}{\sqrt{n}}\sum_{i=1}^n \mathbb{E}^* 2F(\boldsymbol{X}_i)\mathbb{1}\left\{2F(\boldsymbol{X}_i) > \sqrt{n}a_{q_0}\right\}\\
\leq & \frac{4r}{a_{q_0}}\mathbb{E}^*\left[(2F)^2\mathbb{1}\left\{2F > \sqrt{n}a_{q_0}\right\}\right]\\
\lesssim & \frac{r}{a_{q_0}}\|F\|_{\mathbb{P},2}^2.
\end{aligned}$$

The equality comes from Lemma 2.8.6 and the third line is true since $\boldsymbol{X}_i$'s are i.i.d. Choose q_0 such that $2^{-q_0} = \delta \|F\|_{\mathbb{P},2}$ for some $\delta > 0$. Then

$$\mathbb{E}^* \left\| \sqrt{n}\mathbb{M}_n(h^f - \xi_{q_0} h^f) B_{q_0} f \right\|_{\mathcal{F}} \lesssim r 2^{-q_0} \sqrt{\operatorname{Log} N_{q_0}}. \tag{2.21}$$

For any function class $\mathcal{H}$ with some envelope function H, we can bound $|\mathbb{M}_n h|$ by $\frac{1}{n}\sum_{i=1}^n (\mathbb{E}_{i-1} H + H) = -\mathbb{M}_n H + \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{i-1} H$ for any $h \in \mathcal{H}$. Since $\left|(h^f - \xi_q h^f) B_q f\right| \le |R_i| \, \Delta_q f B_q f / \pi_i(A_i; \hat{f}_{i-1})$, the second term (2.18) can be bounded by

$$\begin{aligned}
&\mathbb{E}^* \left\| \sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n (h^f - \xi_q h^f) B_q f \right\|_{\mathcal{F}} \\
&= \sum_{q=q_0+1}^{\infty} \mathbb{E}^* \left\| \sqrt{n}\mathbb{M}_n \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\mathcal{F}} \\
&\quad + \sum_{q=q_0+1}^{\infty} 2\sqrt{n}\mathbb{E}^* \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{i-1} \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\mathcal{F}}.
\end{aligned} \tag{2.22}$$

By Corollary 2.8.5, for each q in the first term in (2.22), the expectation can be split into two parts:

$$\begin{aligned}
&\mathbb{E}^* \left\| \sqrt{n}\mathbb{M}_n \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\mathcal{F}} \\
\lesssim &\frac{1}{\sqrt{n}} \max_{f \in \mathcal{F}} \left\| \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\infty} \operatorname{Log} N_q \\
&+ \frac{1}{\sqrt{n}} \max_{f \in \mathcal{F}} \sqrt{\left\| \sum_{i=1}^n \operatorname{Var}\left[\frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \middle| \mathcal{G}_{i-1} \right] \right\|_{\infty}} \sqrt{\operatorname{Log} N_q}.
\end{aligned}$$

Since $\Delta_q f B_q f \le \Delta_{q-1} f A_{q-1} f \le \sqrt{n} a_{q-1}$, the L_∞ term in the first part can be bounded by

$$\left\| \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\infty} \le \frac{r}{\epsilon_n} \sqrt{n} a_{q-1} \tag{2.23}$$

for any $f \in \mathcal{F}$. For the second part, by the assumption of the second conditional moment,

$$\begin{aligned}
&\left\| \sum_{i=1}^{n} \mathrm{Var}\left[\frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \middle| \mathcal{G}_{i-1} \right] \right\|_\infty \\
\leq & \sum_{i=1}^{n} \left\| \mathbb{E}\left[\frac{\tau^2}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q^2 f B_q^2 f \middle| \mathcal{G}_{i-1} \right] \right\|_\infty \\
\leq & \sum_{i=1}^{n} \left\| \left(\frac{1}{\epsilon_i} + \frac{1}{0.5} \right) \mathbb{E}\left(\tau^2 \Delta_q^2 f B_q^2 f \right) \right\|_\infty .
\end{aligned} \tag{2.24}$$

The last inequality comes from Lemma 2.8.6. For any non-negative random variable X, we have the inequality $\|X\|_{2,\infty}^2 \leq \sup_{t>0} t\mathbb{E}\left[X \mathbb{1}(X > t)\right] \leq 2\|X\|_{2,\infty}^2$. Then

$$\sqrt{n} a_q \mathbb{E}\left(\Delta_q f B_q f\right) \leq \sqrt{n} a_q \mathbb{E}\left(\Delta_q f \mathbb{1}(\Delta_q f > \sqrt{n} a_q)\right) \leq 2\|\Delta_q f\|_{\mathbb{P},2,\infty}^2 \leq 2 \cdot 2^{-2q}. \tag{2.25}$$

Since $\Delta_q f B_q f$ is bounded by $\sqrt{n} a_{q-1}$ for $q > q_0$, it follows that

$$\mathbb{E}\left(\Delta_q^2 f B_q^2 f\right) \leq \sqrt{n} a_{q-1} \mathbb{E}\left(\Delta_q f B_q f\right) \leq 2 \frac{a_{q-1}}{a_q} 2^{-2q}.$$

Using Lemma 2.8.6 again and the inequality (2.25), the second term in (2.22) can be bounded as

$$\begin{aligned}
&\mathbb{E}^* \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{i-1} \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_q f B_q f \right\|_{\mathcal{F}} \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}^* \left\| 2r \mathbb{E}\left(\Delta_q f B_q f\right) \right\|_{\mathcal{F}} \leq 4r \frac{1}{\sqrt{n} a_{q-1}} 2^{-2q}.
\end{aligned} \tag{2.26}$$

Now apply the above bounds (2.23), (2.24), (2.26) on (2.18) to find

$$\begin{aligned}
&\mathbb{E}^* \left\| \sum_{q=q_0+1}^{\infty} \sqrt{n} \mathbb{M}_n (h^f - \xi_q h^f) B_q f \right\|_{\mathcal{F}} \\
\lesssim & \sum_{q=q_0+1}^{\infty} \left[\frac{1}{\sqrt{n}} \frac{r}{\epsilon_n} \sqrt{n} a_{q-1} \mathrm{Log}\, N_q + \frac{1}{\sqrt{n}} \sqrt{2\tau^2 \sum_{i=1}^{n} \left(\frac{1}{\epsilon_i} + \frac{1}{0.5} \right) \frac{a_{q-1}}{a_q} 2^{-2q}} \sqrt{\mathrm{Log}\, N_q} \right. \\
& \qquad \left. + \sqrt{n} 4r \frac{1}{\sqrt{n} a_{q-1}} 2^{-2q} \right] \\
\lesssim & \sum_{q=q_0+1}^{\infty} \left[\frac{r}{\epsilon_n} + \tau \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\epsilon_i}} + r \right] 2^{-q} \sqrt{\mathrm{Log}\, N_q}.
\end{aligned} \tag{2.27}$$

The last inequality comes from the fact that $a_{q-1}/a_q \leq (a_{q-1}/a_q)^2$ for decreasing a_q.

To handle the third term (2.19), first note that it is bounded by

$$\sum_{q=q_0+1}^{\infty} \mathbb{E}^* \left\| \sqrt{n}\mathbb{M}_n \frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_{q-1} f A_{q-1} f \right\|_{\mathcal{F}},$$

since the partition is nested. Then we can use Corollary 2.8.5 as in the bound of the second term (2.18). The maximum of L_∞ norm over $\mathcal{F}$ in the first part is upper bounded by $r\sqrt{n}a_{q-1}/\epsilon_n$. For the second part, use Lemma 2.8.6 and assumption (2.15) to find

$$\begin{aligned}
&\left\| \sum_{i=1}^{n} \mathrm{Var} \left[\frac{|R_i|}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_{q-1} f A_{q-1} f \middle| \mathcal{G}_{i-1} \right] \right\|_\infty \\
\leq &\left\| \sum_{i=1}^{n} \mathbb{E} \left[\frac{\tau^2}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \Delta_{q-1}^2 f \middle| \mathcal{G}_{i-1} \right] \right\|_\infty \\
\leq &\tau^2 \sum_{i=1}^{n} \left(\frac{1}{\epsilon_i} + \frac{1}{0.5} \right) \left\| \mathbb{E} \Delta_{q-1}^2 f \right\|_\infty \\
\leq &\tau^2 \sum_{i=1}^{n} \left(\frac{1}{\epsilon_i} + \frac{1}{0.5} \right) (2^{-q+1})^2.
\end{aligned}$$

Combining the two parts together, we have

$$\begin{aligned}
&\mathbb{E}^* \left\| \sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n (\xi_q h^f - \xi_{q-1} h^f) A_{q-1} f \right\|_{\mathcal{F}} \\
\lesssim &\sum_{q=q_0+1}^{\infty} \left[\frac{1}{\sqrt{n}} \frac{r}{\epsilon_n} \sqrt{n} a_{q-1} \operatorname{Log} N_q + \frac{1}{\sqrt{n}} \sqrt{2\tau^2 \sum_{i=1}^{n} \left(\frac{1}{\epsilon_i} + \frac{1}{0.5} \right) 2^{-2q+2}} \sqrt{\operatorname{Log} N_q} \right] \qquad (2.28) \\
\lesssim &\sum_{q=q_0+1}^{\infty} \left[\frac{r}{\epsilon_n} + \tau \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\epsilon_i}} \right] 2^{-q} \sqrt{\operatorname{Log} N_q}.
\end{aligned}$$

For the last term (2.20), consider two cases on whether the envelope function F is bounded by $\sqrt{n}a_{q_0}$ or not. Apply Corollary 2.8.5 in the first case. With the supremum part bounded by

$$\left\| \xi_{q_0} h^f \mathbb{1}\left(F \leq \sqrt{n}a_{q_0}\right) \right\|_\infty \leq \frac{r}{\epsilon_n} \left\| 2F \mathbb{1}\left(F \leq \sqrt{n}a_{q_0}\right) \right\|_\infty \leq \frac{2r}{\epsilon_n} \sqrt{n} a_{q_0}$$

and the conditional variance part bounded by

$$\left\|\sum_{i=1}^{n}\mathrm{Var}\left(\xi_{q_0}h\mathbb{1}\left(F\le\sqrt{n}a_{q_0}\right)\middle|\mathcal{G}_{i-1}\right)\right\|_{\infty}\le\tau^2\sum_{i=1}^{n}\left(\frac{1}{1-\epsilon_i}+\frac{1}{\epsilon_i}\right)\|F\|_{\mathbb{P},2}^2,$$

we have

$$\begin{aligned}
&\mathbb{E}^*\left\|\sqrt{n}\mathbb{M}_n\xi_{q_0}h^f\mathbb{1}\left(F\le\sqrt{n}a_{q_0}\right)\right\|_{\mathcal{F}}\\
\lesssim&\frac{1}{\sqrt{n}}\frac{r}{\epsilon_n}\sqrt{n}a_{q_0}\operatorname{Log}N_{q_0}+\frac{1}{\sqrt{n}}\sqrt{\tau^2\sum_{i=1}^{n}\left(\frac{1}{\epsilon_i}+\frac{1}{0.5}\right)\|F\|_{\mathbb{P},2}^2}\sqrt{\operatorname{Log}N_{q_0}}\\
\lesssim&\frac{r}{\epsilon_n}2^{-q_0}\sqrt{\operatorname{Log}N_{q_0}}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}\|F\|_{\mathbb{P},2}\sqrt{\operatorname{Log}N_{q_0}}.
\end{aligned}$$

In the second case, since $\xi_{q_0}h^f\mathbb{1}\left(F>\sqrt{n}a_{q_0}\right)$ is bounded by $\frac{2r}{\epsilon_n}F\mathbb{1}\left(F>\sqrt{n}a_{q_0}\right)$,

$$\mathbb{E}^*\left\|\sqrt{n}\mathbb{M}_n\xi_{q_0}h^f\mathbb{1}\left(F>\sqrt{n}a_{q_0}\right)\right\|_{\mathcal{F}}\lesssim\frac{r}{a_{q_0}}\|F\|_{\mathbb{P},2}^2$$

by the same argument in the bounds for (2.21). Therefore, by applying the triangle inequality and choosing q_0 so that $2^{-q_0}=\delta\|F\|_{\mathbb{P},2}$ for some constant $\delta>0$,

$$\begin{aligned}
&\mathbb{E}^*\left\|\sqrt{n}\mathbb{M}_n\xi_{q_0}h^f\right\|_{\mathcal{F}}\\
\lesssim&\frac{r}{\epsilon_n}2^{-q_0}\sqrt{\operatorname{Log}N_{q_0}}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}\|F\|_{\mathbb{P},2}\sqrt{\operatorname{Log}N_{q_0}}+\frac{r}{a_{q_0}}\|F\|_{\mathbb{P},2}^2\\
\lesssim&\left[\frac{r}{\epsilon_n}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}+r\right]2^{-q_0}\sqrt{\operatorname{Log}N_{q_0}},
\end{aligned}\tag{2.29}$$

where $N_{q_0}=N_{[]}(\delta\|F\|_{\mathbb{P},2},\mathcal{F},L_2(\mathbb{P}))$.

Finally, we obtain

$$\begin{aligned}
\mathbb{E}^*\left\|\sqrt{n}\mathbb{W}_n(f)\right\|_{\mathcal{F}}\lesssim&\sum_{q=q_0}^{\infty}\left[\frac{r}{\epsilon_n}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}+r\right]2^{-q}\sqrt{\operatorname{Log}N_q}\\
\lesssim&\frac{r}{\epsilon_n}J_{[]}(\|F\|_{\mathbb{P},2},\mathcal{F},L_2(\mathbb{P}))
\end{aligned}$$

by combining the four upper bounds (2.21), (2.27), (2.28) and (2.29). □

2.8.3 Proof of Theorem 2.3.1

Using Lemma 2.8.2 and Lemma 2.3.2, we can give our proof of Theorem 2.3.1. Apart from applying the bound on the expectation of supremum to the concentration inequality of martingale process, we also combine the pilot trial with the main trial which follows the adaptive design.

Proof. First note that $\mathcal{V}(f^*) - \mathcal{V}(f) = \mathbb{E}^{\text{sign}\{f^*\}}(R) - \mathbb{E}^{\text{sign}\{f\}}(R)$, which is the opposite of excess 0-1 risk. For the initial n_0 i.i.d. observations, we know that the excess 0-1 risk is bounded by excess ϕ-risk, that is,

$$\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) \geq \mathbb{E}^{\text{sign}\{f^*\}}(R_i) - \mathbb{E}^{\text{sign}\{f\}}(R_i)$$

for any $i = 1, \ldots, n_0$ and any measurable f by Theorem 3.2 in Zhao et al. (2012). For sequentially generated data $\{\boldsymbol{Z}_i\}_{i=1}^n$, note that conditioning on $\mathcal{G}_{i-1}$,

$$\begin{aligned}
\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) =& \mathbb{E}\left(\frac{R_i\phi(A_i f)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right) - \mathbb{E}\left(\frac{R_i\phi(A_i f^*)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right) \\
\geq& \mathbb{E}\left(\frac{R_i\mathbb{1}\{A_i \neq \text{sign}\{f\}\}}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right) - \mathbb{E}\left(\frac{R_i\mathbb{1}\{A_i \neq \text{sign}\{f^*\}\}}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)}\middle|\mathcal{G}_{i-1}\right) \\
=& \mathbb{E}^{\text{sign}\{f^*\}}(R_i) - \mathbb{E}^{\text{sign}\{f\}}(R_i)
\end{aligned}$$

for any $i = 1, \ldots, n$ and any measurable f. The inequality can be proved similarly as in the i.i.d. case of Theorem 3.2 in Zhao et al. (2012), but with a condition on $\mathcal{G}_{i-1}$. Therefore, the value function difference $\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n)$ is upper bounded by

$$\frac{1}{n_0 + n}\left[\sum_{i=1}^{n_0}\mathbb{E}h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) + \sum_{i=1}^{n}\mathbb{E}_{i-1}h^{\hat{f}_n}(\boldsymbol{Z}_i)\right].$$

In SRAT, $\hat{f}_n$ should be minimizing $\sum_{i=1}^{n_0} g^f(\boldsymbol{Z}_i^{(0)}) + \sum_{i=1}^{n} g^f(\boldsymbol{Z}_i)$, so we have

$$\sum_{i=1}^{n_0} h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) + \sum_{i=1}^{n} h^{\hat{f}_n}(\boldsymbol{Z}_i) \leq 0.$$

It follows that

$$\begin{aligned}
&\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n) \\
\leq &\frac{1}{n_0+n}\left[\sum_{i=1}^{n_0}\mathbb{E}h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) + \sum_{i=1}^{n}\mathbb{E}_{i-1}h^{\hat{f}_n}(\boldsymbol{Z}_i) - \sum_{i=1}^{n_0}h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) - \sum_{i=1}^{n}h^{\hat{f}_n}(\boldsymbol{Z}_i)\right] \\
\leq &\sup_{f\in\mathcal{F}}\frac{1}{n_0+n}\left[\sum_{i=1}^{n_0}\left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] + \sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right].
\end{aligned} \tag{2.30}$$

Now it suffices to bound the right-hand side of (2.30).

We will use Lemma 2.8.2 to bound the martingale part. First we test the conditions of the lemma. Since $\mathcal{R}_n(\mathcal{F})$ and $\sup_{\mathbb{P}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ are on the same scale and the latter one $\sup_{\mathbb{P}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ is in the order of $1/\sqrt{n}$, the first assumption in Lemma 2.8.2 is satisfied. The second one can be satisfied when taking a large class $\mathcal{F}$, for example, a linear class with parameters bounded loosely.

Let $\mathcal{H}(\mathcal{F})$ be the class of functions constructed by h^f as f ranges over $\mathcal{F}$. According to (2.8),

$$\mathcal{R}_n(\mathcal{H}(\mathcal{F})) \leq 2\sup_{\mathbb{P}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{W}_n(f) + \frac{D}{2\sqrt{n}},$$

where $D = \inf_{z\in\mathcal{Z}}\sup_{h^f,h^{f'}\in\mathcal{F}}[h^f(z) - h^{f'}(z)] \geq 0$. Since R and h^f can take value zero, $D = 0$ here. Therefore, $\mathcal{R}_n(\mathcal{H}(\mathcal{F}))$ is bounded by $rJ_{[]}(\|F\|_{\mathbb{P},2}, \mathcal{F}, L_2(\mathbb{P}))/(\sqrt{n}\epsilon_n)$ up to a constant by Lemma 2.3.2. Since $\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)$ is upper bounded by $2rb/\epsilon_n$ for all i and all $f \in \mathcal{F}$, scale (2.9) and we get

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right| > t\right) \leq 8L\exp\left\{-\frac{n\epsilon_n^4}{\log^3 n}\frac{t^2}{Cr^4b^2J^2}\right\}$$

for some constant C and any $t > 0$. In other words,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right| > C\frac{r^2bJ}{\epsilon_n^2}\sqrt{n\log^3 n\delta}\right) \leq e^{-\delta} \tag{2.31}$$

for some constant C and any $\delta > 0$.

To derive a bound for the initial randomized treatments of size n_0, we will take use of a variant of Talagrand's inequality (Talagrand, 1994) in Lemma 2.8.1, which is a common approach in i.i.d.

classification problems. In our setting, $\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)$ has an expectation zero for all i and all $f \in \mathcal{F}$ and $\left\|\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right\|_\infty \leq 2rb/(1/2)$. Note that $\pi_i(A_i) = 1/2$ for all $i \in \{1, \ldots, n_0\}$ in pilot data. By assumption, $\frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathrm{Var}[h^f(\boldsymbol{Z}_i)]$ is bounded by $4b^2r^2$. The key step here is to bound

$$\mu^* = \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \sum_{i=1}^{n_0} [\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})]\right\}$$

in Lemma 2.8.1. By Theorem 2.14.2 in Van der Vaart and Wellner (1996), the expectation of supremum of an empirical process is bounded by the bracketing integral. Following a similar proof of Lemma 2.3.2, with only Freedman's inequality (Freedman, 1975) replaced by Bernstein's inequality, we know $\mu^* \leq rJ\sqrt{n_0}$, since J is the supremum of bracketing integrals over all possible measures. Therefore,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^{n_0} \left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] \geq 3rJ\sqrt{n_0} + \sqrt{4\delta br^2 n_0} + 4rb\delta\right) \leq e^{-\delta}. \tag{2.32}$$

Now by the triangle inequality and the fact that $\mathbb{P}(|X+Y| \geq a+b) \leq \mathbb{P}(|X| \geq a) + \mathbb{P}(|Y| \geq b)$,

$$\begin{aligned}\mathbb{P}\Bigg(\sup_{f\in\mathcal{F}} \frac{1}{n_0+n}\Bigg[\sum_{i=1}^{n_0}\left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] + \sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\Bigg] \\ \geq \frac{C}{n_0+n}\left[(rJ + r\sqrt{\delta b})\sqrt{n_0} + rb\delta + \frac{r^2bJ}{\epsilon_n^2}\sqrt{n\log^3 n\delta}\right]\Bigg) \leq e^{-\delta}\end{aligned} \tag{2.33}$$

for some constant C and any $\delta > 0$. The result on test data follows by combining inequalities (2.30) and (2.33). □

2.8.4 Proof of Theorem 2.3.3

Proof. First note that

$$\left|\frac{1}{n}\sum_{i=1}^n \left[\mathcal{V}(\hat{f}_{i-1}) - R_i\right]\right| \leq \left|\frac{1}{n}\sum_{i=1}^n [R_i - \mathbb{E}_{i-1}R_i]\right| + \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}_{i-1}R_i - \mathcal{V}(\hat{f}_{i-1})\right|. \tag{2.34}$$

Given $\mathcal{G}_{i-1}$ and I_i, $\mathbb{E}(R_i|\mathcal{G}_{i-1}, I_i)$ is actually $\mathcal{V}(\hat{f}_{i-1}I_i)$. Then $\mathbb{E}_{i-1}R_i$ can be written as

$$\mathbb{E}\left[\mathbb{E}(R_i|\mathcal{G}_{i-1}, I_i)|\mathcal{G}_{i-1}\right] = \mathbb{E}[p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)|\mathcal{G}_{i-1}]\mathcal{V}(\hat{f}_{i-1}) + \mathbb{E}[1 - p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)|\mathcal{G}_{i-1}]\mathcal{V}(-\hat{f}_{i-1}),$$

where $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ is the probability of $I_i = 1$. So the second term of the right-hand side of (2.34) is upper bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[1-p_i(\boldsymbol{H}_{i-1},\boldsymbol{X}_i)|\mathcal{G}_{i-1}]|\mathcal{V}(\hat{f}_{i-1})-\mathcal{V}(-\hat{f}_{i-1})|\leq\frac{2r}{n}\sum_{i=1}^{n}\epsilon_i'.$$

For the first term, note that $\{R_i - \mathbb{E}_{i-1}R_i\}_{i=1}^n$ is a martingale difference sequence. We will use the Freedman's inequality in Lemma 2.8.3. The two parameters can be bounded as $\|R_i - \mathbb{E}_{i-1}R_i\|_\infty \leq 2r$ and $\left\|\sum_{i=1}^n \mathbb{E}_{i-1}(R_i - \mathbb{E}_{i-1})^2/n\right\|_\infty \leq nr^2$. Therefore,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}[R_i-\mathbb{E}_{i-1}R_i]\right|\geq t\right)\leq 2\exp\left\{-\frac{1}{2}\frac{nt^2}{r^2+2rt/3}\right\}. \tag{2.35}$$

Let the right-hand side be $e^{-\delta}$ and the result follows. □

2.8.5 Proof of Corollary 2.3.4

Proof. First note that

$$\begin{aligned}&\left|\mathcal{V}(f^*)-\bar{R}_n\right|\\ \leq&\left|\frac{1}{n}\sum_{i=1}^{n}[R_i-\mathbb{E}_{i-1}R_i]\right|+\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{i-1}R_i-\mathcal{V}(\hat{f}_{i-1})\right|+\frac{1}{n}\sum_{i=1}^{n}\left|\mathcal{V}(\hat{f}_{i-1})-\mathcal{V}(f^*)\right|,\end{aligned}$$

where the first two terms are the same as in the decomposition (2.34). Now let the right-hand side of (2.35) be $e^{-\delta}/3$ and the right-hand side of (2.31) and (2.32) be $e^{-\delta}/3n$ by inverting the two bounds in the proof of Theorem 2.3.1. Note that in the third term we are comparing $\mathcal{V}(\hat{f}_i)$ with $\mathcal{V}(f^*)$ for $i = 0, \ldots, n-1$. When $i = 0$, the term in (2.31) does not actually exist. Hence letting $i\log^3 i = 0$ will work. □

2.8.6 Proof of Theorem 2.3.5

Proof. For the test regret bound (2.5), note that the last term

$$\frac{1}{n_0+n}\frac{r^2bJ}{\epsilon_n^2}\sqrt{\delta n\log^3 n}$$

is in the order of $O(n^{-1/2}(\log n)^{3/2}\epsilon_n^{-2})$ and the sum of the first two terms

$$\frac{1}{n_0+n}\left[(J+\sqrt{\delta b})r\sqrt{n_0}+rb\delta\right]$$

is in the order of $O_p(n^{-1/2})$. So the last term dominates. In addition, we also want ϵ to be non-increasing and the last term to converge to 0. Therefore, θ should be no greater than 1 and larger than 0. Similarly, for the training regret bound (2.7), the second term that contains ϵ_n dominates. The result follows by substituting ϵ_n into the convergence rate and letting the two rates be equal. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i = O(\left(\frac{1}{n}\int_0^n x^{-(1-\theta)/4}\mathrm{d}x\right) = O(n^{-(1-\theta)/4}).$$

□

2.8.7 Additional Simulation Results

Figure 2.7 of scenario 2 demonstrates similar results as Figure 2.2. However, since the dimension of predictors increases, the regret and false decision ratio of the test set are larger than that of scenario 1, especially for small sample sizes. LinUCB is not largely affected by the dimension compared to other methods. Therefore, SRATs with $\epsilon_0 = 0.1, \theta = 1$ exceed LinUCB on the test set only when n is larger than 600. RCT is better than AL-GP on the test set in this scenario, possibly because the nonparametric method is not efficient in a linear setting with a high dimension.

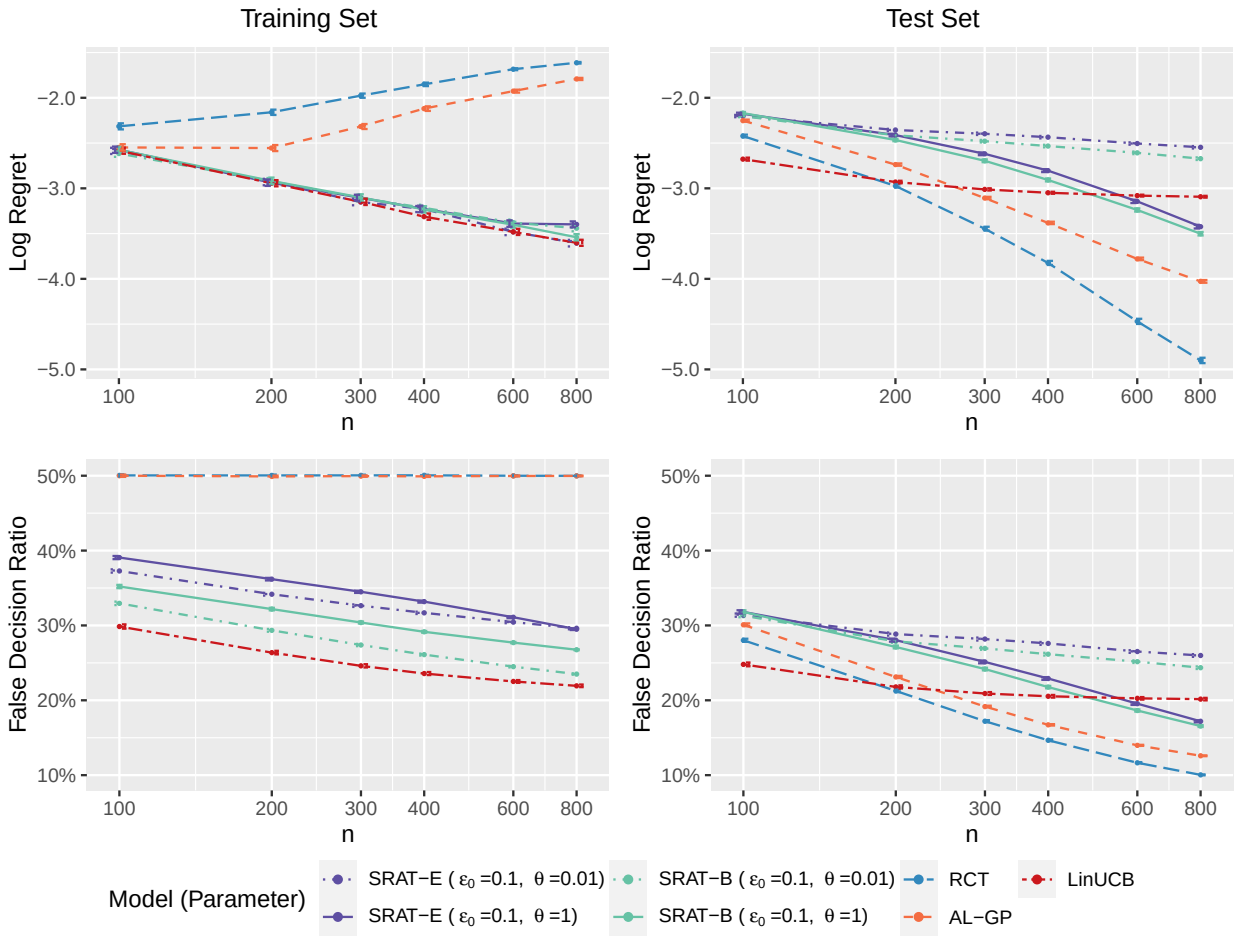

Figure 2.7: Scenario 2. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size n.

CHAPTER 3

Asymptotic Inference for Multi-Stage Stationary Treatment Policy with High Dimensional Features

3.1 Introduction

Dynamic treatment rules (DTRs) or policies have recently attracted great attentions in precision medicine, and they are a sequence of decision functions to prescribe treatments over stages based on a patient's features which can evolve over stages. When the patient's features consist of both baseline variables and clinical variables that are routinely collected over stages, one important class of policies, which we name as multistage stationary treatment policies (MSTPs), are to prescribe from the same set of treatments using the same decision function over all stages. For example in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Rush et al., 2004), demographics and the initial score of Hamilton Rating Scale for Depression (primary outcome) are baseline variables, and the side effect ratings, clinical global impression of improvement, and other depression related symptoms are time-evolving variables. Moreover, the policies are stochastic, meaning that individuals have a large probability to receive recommended treatments. We focus on MSTPs in this chapter because they are particularly useful for treating chronic diseases in clinical practices. First, the policies use the same decision function and the same set of variables so they are convenient for both implementation and interpretation. Second, the policies are dynamic by incorporating individual's evolving features in decisions, which are often known to be important for disease prognostics and thus are routinely collected during clinical visits or lab tests. Third, using stochastic decisions or soft policies is more flexible than relying on a hard policy whose estimation is known to be sensitive to evidence bias and noise. Moreover, the soft policies are useful for exploration when applied to future unknown individuals.

Many approaches have been developed to estimate optimal DTRs using data from a multi-stage study. For example, Q-learning derives the optimal DTRs by maximizing the so-called Q-functions,

which are the conditional means of a total optimal reward outcome given the features at each stage and are often estimated using regression models in a backward fashion (Murphy, 2005b; Zhao et al., 2011; Moodie et al., 2012; Zhu et al., 2019). On the other hand, A-learning poses assumptions only on the interaction effect between the treatments and actions, and thus avoids potential problems of misspecified treatment-free effect (Murphy, 2003; Shi et al., 2018; Jeng et al., 2018). Some other methods find the DTR within a function class that directly maximizes the value function, which is estimated using inverse probability weighting (IPW) with a convex surrogate loss (Zhao et al., 2015) or its augmentation (AIPWE) (Zhang et al., 2013; Liu et al., 2018b).

In addition to the DTR estimation, a number of literature have considered obtaining valid inference for the value function associated with the estimated DTR. For example, Luedtke and Van Der Laan (2016); Zhu et al. (2019) and Shi et al. (2020a) considered the value inference allowing the situation when the treatment is neither beneficial nor harmful for a subpopulation. Similar inference has been studied in the reinforcement learning framework under Markov decision process assumptions (Kallus and Uehara, 2020; Luckett et al., 2020; Liao et al., 2021; Shi et al., 2021b,a). There are very few methods to make inference for the treatment policies themselves. The latter is especially important for studying MSTPs with many feature variables: clinicians usually favor a simple and parsimonious decision because of concern about implementation and cost in collecting disease biomarkers in routine visits or lab tests. Zhu et al. (2019) and Jeng et al. (2018) obtained the asymptotic distribution of the parameters in Q-learning or A-learning by assuming the regression models to be correct, which may not be plausible when there are many features. More recently, Liang et al. (2022) studied inference for a hard treatment decision in a high-dimensional setting, but only restricted to one stage.

This work aims to fill in the above gap through obtaining a valid inference for MSTP with high-dimensional features, assuming data from a sequentially multiple-assignment randomized trial (SMART). Our stationary treatment policy is a probability function of a linear combination of all the feature variables, and a tuning parameter is used in this function to approximate a hard decision. For inference, we first estimate the average value function using the AIPWE method (Liu et al., 2018b), where the augmentation term is constructed by fitting a working model for the single-stage Q-function. We show that any specification of the outcome regression model does not affect the consistency or convergence rate of the resulting estimator due to the construction

method of the augmentation term. Furthermore, to find a sparse estimator of MSTP, we impose a L_1 penalty for the purpose of variable selection. The classical inferential theory fails in this case due to the presence of high-dimensional parameters. The asymptotic distribution of the estimated parameter becomes intractable due to the non-ignorable estimation bias and the sparsity effect of the nuisance parameters. To validate the inference for the parameter estimators, we adopt the idea of one-step estimation (Zhang and Zhang, 2014; Ning and Liu, 2017) to remove bias in the regularized estimators. The augmentation in our first step shares a similar spirit with debiased machine learning methods in Chernozhukov et al. (2018a) to decorrelate the nuisance parameters (Q-functions) with the policy parameters, except that we do not rely on data-splitting to ensure more reliable estimation with limited data. The one-step improvement in our second step follows (Ning and Liu, 2017) to further decorrelate the parameter of interest from the remaining high-dimensional parameters in MSTP. In other words, our proposed method incorporates two decorrelation procedures for inference. The theoretical analysis is more involved since we need to take care of the high-dimensional plug-in estimator of the nuisance parameters. In addition, the objective function to be minimized in our problem is not convex, which invalidates some of the arguments in Ning and Liu (2017). Theoretically, we show that the final estimators for the policy parameters are asymptotically normal even if (1) the dimension of the feature variables increases exponentially with n; or (2) the models for Q-functions are misspecified and their parameters are estimated at a rate arbitrarily slow.

The rest of this chapter is organized as follows. In Section 3.2, we define the objective of our problem and describe our method for learning the MSTP. Then we introduce the procedure of constructing the one-step estimator in theory and specify some implementation details. In Section 3.3, we show that the one-step estimator is asymptotically normal and provide its confidence intervals. In Sections 3.4, we demonstrate our method in simulation studies. Finally, we conclude this chapter with some discussion in Section 3.5.

3.2 Methodology

Consider treatment decisions for T stages. Let $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{X}$ be a d-dimensional feature vector and $\mathcal{A} = \{-1, 1\}$ be the set of treatment options. We assume that at each stage, the

feature variables are from $\mathcal{X}$ and the treatments are from the same set $\mathcal{A}$, but their distributions may vary with stages. An MSTP is defined as a sequence of the same probability function, denoted by $\pi(a|\boldsymbol{X})$, to assign treatment at all stages. We assume that $\pi(a|\boldsymbol{X})$ is from the following class

$$\mathcal{B}(\mathcal{A}) = \left\{ \pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}) = \frac{e^{ag(\boldsymbol{X},\boldsymbol{\theta})/\tau}}{1+e^{ag(\boldsymbol{X},\boldsymbol{\theta})/\tau}} : \|\boldsymbol{\theta}\|_2 = 1, a \in \mathcal{A} \right\},$$

where τ is a constant scaling parameter, and the restriction $\|\boldsymbol{\theta}\|_2 = 1$ is for avoiding non-identifiability. We take $g(\boldsymbol{X},\boldsymbol{\theta}) = \sum_{j=1}^{d} X_j\theta_j$ to be the linear function for easy interpretability, which is important in clinical practice. Note that the scaling parameter τ can be used to adjust the influence of parameters on the action probability. A smaller τ leads to a decision closer to a hard policy. When $\pi(a|\boldsymbol{X})$ is from $\mathcal{B}(\mathcal{A})$, the corresponding MSTP is indexed by $\boldsymbol{\theta}$. Without confusion, we use $\mathbb{E}_{\boldsymbol{\theta}}$ to denote the expectation under the evaluation policy indexed by $\boldsymbol{\theta}$, which means that the treatment in each stage is taken with probability indicated by $\pi^{\boldsymbol{\theta}}$. Our goal is to estimate the optimal MSTP and obtain a proper inference for $\boldsymbol{\theta}$.

We assume that data are obtained from a T-stage study, so the observations of the i-th subject can be represented as

$$\boldsymbol{D}_i = \{\boldsymbol{X}_{i,1}, A_{i,1}, R_{i,1}, \boldsymbol{X}_{i,2}, A_{i,2}, R_{i,2}, \ldots, \boldsymbol{X}_{i,T}, A_{i,T}, R_{i,T}\},$$

where the reward $R_{i,t} \in \mathbb{R}$ is an unknown function of the data $\{\boldsymbol{X}_{i,1}, A_{i,1}, R_{i,1}, \ldots, \boldsymbol{X}_{i,t}, A_{i,t}\}$ observed prior to or at time t. Let the domain of $\boldsymbol{D}_i$ be defined as $\mathcal{D} := (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^T$. Assume there are n subjects and their trajectories $\{\boldsymbol{D}_i\}_{i=1}^n$ are independent and identically distributed. We assume that each action $A_{i,t}$ is taken randomly with probability depending on the history $\boldsymbol{H}_{i,t} = \{\boldsymbol{X}_{i,1}, A_{i,1}, R_{i,1}, \ldots, \boldsymbol{X}_{i,t}\}$, and denote $\mu_t(a|\boldsymbol{H}_{i,t})$ as the conditional probability of $A_{i,t} = a$ given $\boldsymbol{H}_{i,t}$. The collection $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_T)$ is sometimes called the behavior policy in the reinforcement learning literature. We use $\mathbb{E}$ without subscript to denote the expectation under this policy.

We need some basic assumptions to infer the optimal MSTP using the observed data. Specifically, denote $\boldsymbol{X}_t(a_{1:(t-1)})$ the potential state at time t if an action sequence $a_{1:(t-1)} \in \mathcal{A}^{t-1}$ were taken. We make the following three standard assumptions (Murphy, 2003; Nie et al., 2021).

Assumption 8 (Sequential ignorability)**.** The sequence of potential outcomes (states) $\{\boldsymbol{X}_{t'}(A_{1:(t-1)}, a_{t:(t'-1)})\}_{t'=t+1}^{T+1}$ is independent of the treatment A_t given the observed information $\{\boldsymbol{X}_1, A_1, \ldots, \boldsymbol{X}_{t-1}, A_{t-1}\}$ for all $a_t \in \mathcal{A}$, $t = 2, \ldots, T$.

Assumption 9 (Consistency)**.** The observed states are consistent with the potential states, $\boldsymbol{X}_t = \boldsymbol{X}_t(A_{1:(t-1)})$ for all $t = 2, \ldots, T+1$.

Assumption 10 (Positivity)**.** There exists a constant $p_0 > 0$ such that $\mu_t(a|\boldsymbol{H}_t) \geq p_0$ for all $a \in \mathcal{A}$ and $t = 1, \ldots, T$.

As a note, the first assumption holds if the data are from a SMART, which we assume for the subsequent development.

3.2.1 Estimate Policy Parameter with Variable Selection

An important metric to evaluate an MSTP with parameter $\boldsymbol{\theta}$ is called the value function, which is defined as the sum of rewards $V(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(\sum_{t=1}^{T} R_t)$. The optimal MSTP, whose parameter value is denoted as $\boldsymbol{\theta}^*$, is the one maximizing $V(\boldsymbol{\theta})$.

To estimate $V(\boldsymbol{\theta})$ using the SMART data with known μ_t, we first denote the expectation of R_t given $\boldsymbol{X}_t$ and A_t under a given policy with parameter β as

$$Q_t^{\boldsymbol{\theta}}(\boldsymbol{x}_t, a_t) := \mathbb{E}_{\boldsymbol{\theta}}[R_t|\boldsymbol{X}_t = \boldsymbol{x}_t, A_t = a_t]$$

and denote the expectation of Q_t given $\boldsymbol{X}_t$ as

$$U_t^{\boldsymbol{\theta}}(\boldsymbol{x}_t) := \mathbb{E}_{\boldsymbol{\theta}}[Q_t^{\boldsymbol{\theta}}(\boldsymbol{X}_t, A_t)|\boldsymbol{X}_t = \boldsymbol{x}_t].$$

Let $\boldsymbol{Q} := (Q_1, \ldots, Q_T)$ and $\boldsymbol{U} := (U_1, \ldots, U_T)$. Then according to AIPWE given in Liu et al. (2018b), we propose the following estimator for $V(\boldsymbol{\theta})$:

$$\hat{V}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\frac{\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}[R_{i,t} - Q_t^{\boldsymbol{\theta}}(\boldsymbol{X}_{i,t}, A_{i,t})] + \frac{\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}U_t^{\boldsymbol{\theta}}(\boldsymbol{X}_{i,t})\right\}.$$

It uses the step-wise weight instead of the trajectory-wise weight to reduce the variability. One major advantage of using the AIPWE for estimating $V(\boldsymbol{\theta})$ is that the asymptotic limit of $\hat{V}(\boldsymbol{\theta})$ remains to be unbiased even if we use a misspecified model for $\boldsymbol{Q}$, as long as U_t is correctly evaluated as the conditional expectation of Q_t. To see this, notice that the expectation of the estimator $\hat{V}(\boldsymbol{\theta})$ is

$$\begin{aligned}\mathbb{E}\hat{V}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} &\left\{\sum_{t=1}^{T} R_{i,t}\right\} \\ - \mathbb{E} &\left\{\sum_{t=1}^{T} \frac{\prod_{k=1}^{t-1} \pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t-1} \mu_k(A_{i,k}|\boldsymbol{H}_{i,k})} \mathbb{E}\left[\frac{\pi^{\boldsymbol{\theta}}(A_{i,t}|\boldsymbol{X}_{i,t})}{\mu_t(A_{i,t}|\boldsymbol{H}_{i,t})} Q_t(\boldsymbol{X}_{i,t}, A_{i,t}) - U_t(\boldsymbol{X}_{i,t}) \middle| \boldsymbol{H}_{i,k}\right]\right\}\end{aligned}$$

given any Q_t and U_t. Now by the definition of U_t,

$$\begin{aligned}&\mathbb{E}\left[\frac{\pi^{\boldsymbol{\theta}}(A_{i,t}|\boldsymbol{X}_{i,t})}{\mu_t(A_{i,t}|\boldsymbol{H}_{i,t})} Q_t(\boldsymbol{X}_{i,t}, A_{i,t}) - \hat{U}_t(\boldsymbol{X}_{i,t}) \middle| \boldsymbol{H}_{i,k}\right] \\ =&\sum_{a\in\mathcal{A}} \mu_t(a|\boldsymbol{H}_{i,t}) \frac{\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})}{\mu_t(a|\boldsymbol{H}_{i,t})} Q_t(\boldsymbol{X}_{i,t}, a) - U_t(\boldsymbol{X}_{i,t}) = 0\end{aligned}$$

once $U_t(\boldsymbol{X}_{i,t}) = \mathbb{E}_{\boldsymbol{\theta}}[Q_t^{\boldsymbol{\theta}}(\boldsymbol{X}_{i,t}, A_{i,t})|\boldsymbol{X}_t]$. Therefore, $\mathbb{E}\hat{V}(\boldsymbol{\theta}) = V(\boldsymbol{\theta})$ for any Q_t. As a remark, Thomas and Brunskill (2016) and Jiang and Li (2016) proposed to construct the augmentation term in an AIPWE based on the long-term Q-function in reinforcement learning settings, and Kallus and Uehara (2020) proved the double robustness and semiparametric efficiency of this value estimator.

The above property implies that we can always obtain a consistent estimator for $\boldsymbol{\theta}^*$ with any working model for Q_t; however, a good choice of Q_t is likely to improve the efficiency of the estimator. For example, a simple estimate, denoted by $\hat{Q}_t$, for Q_t can be obtained by fitting the single-stage outcomes $\{R_{i,t}\}_{i=1}^n$ against the covariates $\{\boldsymbol{X}_{i,t}\}_{i=1}^n$, using models like parametric and nonparametric regression or other machine learning models for all t. To ensure the relationship between Q_t and U_t to be correct, we estimate U_t using

$$\hat{U}_t(\boldsymbol{x}) = \sum_{a\in\mathcal{A}} \pi^{\boldsymbol{\theta}}(a|\boldsymbol{x}) \hat{Q}_t(\boldsymbol{x}, a). \tag{3.1}$$

After plugging $\hat{Q}_t$ and $\hat{U}_t$ into the expression of $\hat{V}(\boldsymbol{\theta})$, we aim to maximize $\hat{V}(\boldsymbol{\theta})$, or equivalently, minimize the function

$$\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) := \frac{1}{n} \sum_{i=1}^{n} l_i(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}), \tag{3.2}$$

where $\boldsymbol{\eta} := \boldsymbol{Q}$ is the nuisance parameter, and

$$\begin{aligned} l_i(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) := -\sum_{t=1}^{T} \Bigg\{ & \frac{\prod_{k=1}^{t} \pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t} \mu_k(A_{i,k}|\boldsymbol{H}_{i,k})} [R_{i,t} - \hat{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})] \\ & + \frac{\prod_{k=1}^{t-1} \pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t-1} \mu_k(A_{i,k}|\boldsymbol{H}_{i,k})} \hat{U}_t(\boldsymbol{X}_{i,t}) \Bigg\}. \end{aligned} \tag{3.3}$$

Assume $\boldsymbol{\eta} \in \mathcal{H}$, where $\mathcal{H}$ is a convex subset of some normed vector space. Suppose the estimator $\hat{\boldsymbol{\eta}}$ converges in probability uniformly to some deterministic limits $\bar{\boldsymbol{\eta}}$. Only the limit $\bar{\boldsymbol{\eta}}$ but not the true value $\boldsymbol{\eta}^*$ will appear in the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Finally, since $\boldsymbol{\theta}$ is high dimensional, we include L_1-penalty to obtain a sparse MSTP. That is, we estimate $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} [\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) + \lambda_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1] \qquad \text{s.t. } \|\boldsymbol{\theta}\|_2 = 1, \tag{3.4}$$

where $\lambda_{\boldsymbol{\theta}}$ is a tuning parameter for $\boldsymbol{\theta}$.

Since $\boldsymbol{\eta}$ depends on $\boldsymbol{\theta}$, it is computationally challenging to directly minimize the above loss function for estimation. Instead, we propose to estimate $\boldsymbol{\eta}$ using an initial estimator $\check{\boldsymbol{\theta}}$ that does not depend on the $\boldsymbol{\eta}$ and is a consistent estimator of $\boldsymbol{\theta}^*$. For example, $\check{\boldsymbol{\theta}}$ could be the one minimizing the same loss by setting $\hat{Q}_t = 0$, i.e., without augmentation. We will give some additional examples on how to estimate $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{Q}}$ in Section 3.2.3.

3.2.2 Statistical Inference for Sparse High Dimensional Parameters

To obtain valid inference, we follow Zhang and Zhang (2014) and Ning and Liu (2017) to construct one-step improvement of the estimator obtained in the previous section. More specifically, we denote $\boldsymbol{\theta}_{-j}$ the parameters in $\boldsymbol{\theta}$ except θ_j. Without loss of generality, we can put θ_j at the first position in the parameter vector, so that $\boldsymbol{\theta}$ can be written as $(\theta_j, \boldsymbol{\theta}_{-j})$. Similarly, the true parameter $\boldsymbol{\theta}^*$ can be partitioned as $(\theta_j^*, \boldsymbol{\theta}_{-j}^*)$. Given a vector $\boldsymbol{\theta}$, we can define the matrix $\mathbf{I} := \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})]^2$ under regularity conditions since $\mathbb{E}[\ell(\boldsymbol{\theta}, \boldsymbol{\eta})] = -V(\boldsymbol{\theta})$ for any $\boldsymbol{\eta}$. With $I_{\theta_j\theta_j}, \mathbf{I}_{\theta_j\boldsymbol{\theta}_{-j}}, \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}, \mathbf{I}_{\boldsymbol{\theta}_{-j}\theta_j}$

being the corresponding submatrix of $\mathbf{I}$ to the parameters, let $I_{\theta_j|\boldsymbol{\theta}_{-j}} = I_{\theta_j\theta_j} - \mathbf{I}_{\theta_j\boldsymbol{\theta}_{-j}}\mathbf{I}^{-1}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\mathbf{I}_{\boldsymbol{\theta}_{-j}\theta_j}$. Next, a decorrelated score function is defined as

$$S(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta}) = \nabla_{\theta_j}\ell(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta}) - \boldsymbol{w}^T\nabla_{\boldsymbol{\theta}_{-j}}\ell(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta}),$$

where $\boldsymbol{w}^T = \mathbf{I}_{\theta_j\boldsymbol{\theta}_{-j}}\mathbf{I}^{-1}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}$. It is uncorrelated with the nuisance score function in the sense that

$$\mathbb{E}[S(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta})\nabla_{\boldsymbol{\theta}_{-j}}\ell(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta})] = \mathbf{0}.$$

Given the estimated parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}$, the sparse estimator $\hat{\boldsymbol{w}}$ of $\boldsymbol{w}$ can be obtained using the Dantzig estimator

$$\hat{\boldsymbol{w}} = \arg\min \|\boldsymbol{w}\|_1, \qquad \text{s.t.} \quad \left\|\nabla^2_{\theta_j\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\right\|_\infty \le \lambda_{\boldsymbol{w}}, \tag{3.5}$$

where $\lambda_{\boldsymbol{w}}$ is tuned by cross-validation. This Dantzig estimator (3.5) is essentially the best sparse linear combination of the nuisance score function $\nabla_{\boldsymbol{\theta}_{-j}}\ell(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta})$ that approximates the score function $\nabla_{\theta_j}\ell(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{\eta})$ of the parameter of interest with error $\lambda_{\boldsymbol{w}}$. It can also be estimated by penalized least squares using the gradient or Hessian matrix of ℓ (see Ning and Liu, 2017, Remark 3). Let $\hat{\boldsymbol{v}} := (1, -\hat{\boldsymbol{w}}^T)^T$. Then $S(\theta_j, \boldsymbol{\theta}_{-j}, \hat{\boldsymbol{\eta}})$ can be estimated by plugging in $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$ as

$$\hat{S}(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}}) = \nabla_{\theta_j}\ell(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{w}}^T\nabla_{\boldsymbol{\theta}_{-j}}\ell(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}}). \tag{3.6}$$

This score function can be used for hypothesis testing (see Ning and Liu, 2017).

Finally, given the sparse estimator $\hat{\boldsymbol{\theta}}$ from the previous section, a one-step estimator is defined as

$$\tilde{\theta}_j := \hat{\theta}_j - \hat{S}(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}})/\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}, \qquad \text{where } \hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}} = \nabla^2_{\theta_j\theta_j}\ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{w}}^T\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}). \tag{3.7}$$

It is the solution to

$$\hat{S}(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}}) + \hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}(\theta_j - \hat{\theta}_j) = 0,$$

which is the estimating equation of the one-step method for solving

$$\hat{S}(\theta_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}}) = 0.$$

For the true parameter $\boldsymbol{\theta}^*$, define the matrix $\mathbf{I}^* := \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]^2$ and the vectors $\boldsymbol{w}^* = \mathbf{I}^{*-1}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j}$, $\boldsymbol{v}^* = (1, -\boldsymbol{w}^{*T})^T$. Let

$$\boldsymbol{\Sigma}^* := \mathrm{Var}[\nabla_{\boldsymbol{\theta}} l_0(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})] \quad \text{and} \quad \sigma_S^* := \boldsymbol{v}^{*T} \boldsymbol{\Sigma}^* \boldsymbol{v}^*, \tag{3.8}$$

where l_0 is an independent copy of l_i for any i. As will be shown in Section 3.3, $\tilde{\theta}_j$ is asymptotically normal with mean $\boldsymbol{\theta}^*$ and asymptotic variance $\sigma_S^*/[nI^{*2}_{\theta_j|\boldsymbol{\theta}_{-j}}]$, which can be used for constructing the confidence intervals.

We summarize the theoretical steps of estimating the high dimensional parameters and making inference with one-step estimators in Algorithm 2. Note that the confidence level is for every single confidence interval rather than for multiple confidence intervals simultaneously. Some implementation details will be specified in Section 3.2.3.

3.2.3 Implementation

To get an initial estimate of the policy parameter, we propose to use the sum of rewards without the augmentation term as an estimate of the value function. This estimate is unbiased and does not require estimation of the nuisance parameters. Besides, the probability ratio $\frac{\prod_{k=1}^t \pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^t \mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}$ in (3.2), which we denote as $\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}$, may be highly unstable in numerical computation since it is a product of t probability ratios. The variance of the value estimator can thus grow exponentially with the horizon T (Liu et al., 2018a) and leads to performance worse than that of a biased estimator (Sugiyama, 2015). A practical solution is to use the weighted importance sampling, which weights the probability ratio by the average ratio across all episodes at this stage. Therefore, the initial estimator $\check{\boldsymbol{\theta}}$ is found by

$$\check{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[-\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \frac{\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}}{w_{1:t}} R_{i,t} + \lambda_{\check{\boldsymbol{\theta}}} \|\boldsymbol{\theta}\|_1 \right] \qquad \text{s.t. } \|\boldsymbol{\theta}\|_2 = 1, \tag{3.10}$$

Algorithm 2: Procedure for the estimation and inference of high-dimensional policy parameters

Input : n samples

Output: $\tilde{\boldsymbol{\theta}}$ and the confidence intervals for each coefficient

Obtain an initial estimator $\check{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$;

Estimate $\hat{\boldsymbol{Q}}$ using the initial estimator $\check{\boldsymbol{\theta}}$;

Estimate the policy parameter $\hat{\boldsymbol{\theta}}$ by (3.4) with $\hat{\boldsymbol{Q}}$, where $\lambda_{\boldsymbol{\theta}}$ is tuned by cross-validation;

for $j = 1, \ldots, d$ **do**

Partition the sparse estimate $\hat{\boldsymbol{\theta}}$ as $(\hat{\theta}_j, \hat{\boldsymbol{\theta}}_{-j})$;

Obtain the Dantzig type estimator $\hat{\boldsymbol{w}}$ by (3.5);

Obtain the decorrelated score function $\hat{S}(\theta_j, \hat{\boldsymbol{\theta}}_{-j}, \hat{\boldsymbol{\eta}})$ by (3.6);

Calculate the one-step estimator $\tilde{\theta}_j$ by (3.7);

Construct the $(1-\alpha)\times 100\%$ confidence interval of θ_j by

$$\left(\tilde{\theta}_j - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\hat{\sigma}_S}}{\sqrt{n}\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}}, \tilde{\theta}_j + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\hat{\sigma}_S}}{\sqrt{n}\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}}\right), \tag{3.9}$$

where Φ is the cumulative distribution function of the standard normal distribution, and

$$\hat{\sigma}_S = (1, -\hat{\boldsymbol{w}}^T)\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}} l_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\nabla_{\boldsymbol{\theta}} l_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})^T\right](1, -\hat{\boldsymbol{w}}^T)^T.$$

end

and the final sparse estimator $\hat{\boldsymbol{\theta}}$ is found by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \Bigg[-\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}}{w_{1:t}}[R_{i,t} - \hat{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})] + \frac{\rho_{i,1:(t-1)}^{\boldsymbol{\theta},\boldsymbol{\mu}}}{w_{1:t-1}}\hat{U}_t(\boldsymbol{X}_{i,t}) + \lambda_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 \Bigg] \quad \text{s.t. } \|\boldsymbol{\theta}\|_2 = 1, \tag{3.11}$$

where $w_{1:t} = \frac{1}{n}\sum_{j=1}^{n} \rho_{j,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}$ and $\lambda_{\check{\boldsymbol{\theta}}}, \lambda_{\boldsymbol{\theta}}$ are tuning parameters for the initial and final sparse estimators respectively. Similar weighted ratios have been used and discussed in Precup (2000b); Thomas (2015); Thomas and Brunskill (2016).

To solve for $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$, note that (3.10) and (3.11) are both constrained nonconvex nondifferentiable optimization problems. To deal with the L_1 penalty, we use the proximal coordinate descent algorithm. To ensure that the L_2 norm of the estimated parameter is bounded by one, we normalize the parameter by its L_2 norm in each iteration of the coordinate descent. Since nonconvex problem may converge to a local minimizer, we try to start the optimization from different starting points for better numerical results. In addition, it has been shown that a refitted Lasso estimator usually leads to a better finite sample performance than the original Lasso estimator (Zhang and Zhang, 2014; Ning and Liu, 2017). A refitted Lasso estimator means that we re-estimate the parameter on the support of the original Lasso estimator using the original loss function without the L_1 penalty. This refitted estimator may be less biased and less sensitive to the choice of tuning parameters of the penalty. Therefore, we refit $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ on their support using the trust-region constrained algorithm, which is suitable for minimization problems with constraints. The full optimization procedure is summarized in Algorithm 4.

Based on this initial estimate $\check{\boldsymbol{\theta}}$, we model the nuisance parameter $\boldsymbol{Q}$ using a linear model of basis functions for each stage separately. Assume that $\phi(\boldsymbol{x}_t)$ is a basis function of the covariates $\boldsymbol{x}_t$ at some stage t of dimension d' and it includes an intercept. In practice, $\phi(\boldsymbol{x}_t)$ can be taken to be the linear function, polynomial function, Gaussian radial basis functions, splines, wavelet basis, etc. of $\boldsymbol{x}_t$ (Luckett et al., 2020; Shi et al., 2021b). Let $\Phi(\boldsymbol{X}_t, A_t) := [\phi^T(\boldsymbol{X}_t), A_t \cdot \phi^T(\boldsymbol{X}_t)]^T$ and we can fit the model $Q_t(\boldsymbol{X}_t, A_t) = \Phi(\boldsymbol{X}_t, A_t)^T\boldsymbol{\beta}_t$ at stage t, where $\boldsymbol{\beta}_t$ is in the dimension $2d'$. We discuss two different ways for constructing the loss function of $\boldsymbol{\beta}_t$ here. In the first method, we minimize the square loss function with the L_1 penalty to deal with the high dimensional covariates,

so that

$$\hat{\boldsymbol{\beta}}_t^{(1)} = \underset{\boldsymbol{\beta}_{i,t}\in\mathbb{R}^{2d'}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left[R_t - \Phi(\boldsymbol{X}_{i,t}, A_{i,t})^T\boldsymbol{\beta}_t\right]^2 + \lambda_{\boldsymbol{\beta}_t^{(1)}}\|\boldsymbol{\beta}_t\|_1, \tag{3.12}$$

where $\lambda_{\boldsymbol{\beta}_t^{(1)}}$ is a tuning parameter. In the second method, we propose to minimize the variance of the estimator $\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ of the value function at stage t, where

$$\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}[R_{i,t} - \hat{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})] + \rho_{i,1:(t-1)}^{\boldsymbol{\theta},\boldsymbol{\mu}}\hat{U}_t(\boldsymbol{X}_{i,t})\right].$$

Note that $Var(\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})) = \mathbb{E}(\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}))^2 - (\mathbb{E}\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}))^2$. Since $\mathbb{E}\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \mathbb{E}_{\boldsymbol{\theta}}R_t$ for any $\boldsymbol{\theta}$, $\mathbb{E}\hat{V}_t(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$ is approximately the same for all $\hat{\boldsymbol{\eta}}$ when $\hat{\boldsymbol{\theta}}$ is close enough to the true optimal $\boldsymbol{\theta}^*$. Therefore, we only need to minimize the sample average of $\mathbb{E}(\hat{V}_t(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}))^2$. With the linear model of Q_t and the weighted probability ratio, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t^{(2)} = &\underset{\boldsymbol{\beta}_{i,t}\in\mathbb{R}^{2d'}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\frac{\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}}{w_{1:t}}R_{i,t} - \\ &\left[\frac{\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}}{w_{1:t}}\Phi(\boldsymbol{X}_{i,t}, A_{i,t}) - \sum_{a\in\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\Phi(\boldsymbol{X}_{i,t}, a)\right]^T\boldsymbol{\beta}_t\Bigg\}^2 + \lambda_{\boldsymbol{\beta}_t^{(2)}}\|\boldsymbol{\beta}_t\|_1,\end{aligned} \tag{3.13}$$

where $\lambda_{\boldsymbol{\beta}_t^{(2)}}$ is a tuning parameter. Now (3.12) and (3.13) can be easily solved by existing Lasso packages. To improve the finite sample performance, we also refit $\hat{\boldsymbol{\beta}}_t$ on its nonzero components for all t. Then the estimated function Q at stage t is

$$\hat{Q}_t^{(m)}(\boldsymbol{X}_t, A_t) = \Phi(\boldsymbol{X}_{1:t}, A_{1:t})^T\hat{\boldsymbol{\beta}}_t^{(m)} \tag{3.14}$$

for $m = 1, 2$.

The confidence interval of $\tilde{\theta}_j$ in Algorithm 2 relies on the gradient and the Hessian matrix of the loss function. However, our simulation experience shows that the gradient and the Hessian matrix can be highly unstable due to the product of a series of sampling probability $\pi^{\boldsymbol{\theta}}$ and μ_t. Therefore, we propose two amendments to solve this problem in practice. First, for the gradient and the Hessian matrix in (3.5), (3.6), and (3.7), we calculate them numerically using the symmetric difference quotient. Since the ℓ_2-norm of $\hat{\boldsymbol{\theta}}$ is restricted to be 1, we need to take care of the special

case when $\hat{\theta}_0 = 0$ and only the Newton's difference quotient is available. Second, we use bootstrap to find the confidence interval of $\tilde{\theta}_j$ for step (3.9).

The full procedure for finding the estimate and the confidence interval of high-dimensional policy parameters is summarized in Algorithm 3. Note that $\lambda_{\boldsymbol{w}}$ is tuned by cross-validation, where we minimize the average projection error $\|\nabla^2_{\theta_j \boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \boldsymbol{w}^T \nabla^2_{\boldsymbol{\theta}_{-j} \boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\|_\infty$ across all $j = 1, \ldots, j$. In addition, since $\tilde{\boldsymbol{\theta}}$ may be unstable with a small $\lambda_{\boldsymbol{w}}$ when the Hessian matrix is approximately singular, the one standard error rule is used to select the best $\lambda_{\boldsymbol{w}}$. In particular, we choose the largest $\lambda_{\boldsymbol{w}}$ whose average projection error on the validation set is smaller than the minimum average projection error plus its standard error. Then we use the same $\lambda_{\boldsymbol{w}}$ for all j and for the bootstrap estimators.

3.3 Theoretical Results

Denote $s_{\boldsymbol{\theta}} := \|\boldsymbol{\theta}^*\|_0$ and $s_{\boldsymbol{w}} := \|\boldsymbol{w}^*\|_0$ to be the number of nonzero elements in the corresponding vectors. We assume the following assumptions hold for the variables and the convergence rate of the nuisance parameters.

Assumption 11. Assume the rewards R_t are bounded in the sense that $\|R_t\|_\infty \leq r$ for some $r > 0$ and for all $t = 1, \ldots, T$. Assume the covariates are bounded such that $\|\boldsymbol{X}_t\|_\infty \leq z$ and $\left|\boldsymbol{v}^{*T} \boldsymbol{X}_t\right| \leq z$ for some $z > 0$ and for all $t = 1, \ldots, T$.

Assumption 12. Suppose that $\hat{\boldsymbol{\eta}} \in \mathcal{H}_n$ with probability no less than $1 - \Delta_n$, where

$$\mathcal{H}_n := \left\{ \boldsymbol{Q} \in \mathcal{H} : \|Q_t\|_{\mathbb{P},\infty} \leq r, \|Q_t - \bar{Q}_t\|_{\mathbb{P},2} \leq \delta_n \text{ for all } t \right\}$$

and $\delta_n = o(1), \Delta_n = o(1)$ are positive constants. In addition, we assume that $\log d = o(\sqrt{n})$ and $s_{\boldsymbol{\theta}} + s_{\boldsymbol{w}} = O(1)$.

Assumption 13. Suppose that the covariance matrix $\boldsymbol{\Sigma}^* := \text{Var}[\nabla_{\boldsymbol{\theta}} l_0(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]$ is positive definite and finite.

Assumption 14. Assume $V(\boldsymbol{\theta})$ is κ-strongly concave at $\boldsymbol{\theta}^*$. That is, there exists $\kappa > 0$ such that $\langle \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}^*), \Delta_{\boldsymbol{\theta}} \rangle - [V(\boldsymbol{\theta}^* + \Delta_{\boldsymbol{\theta}}) - V(\boldsymbol{\theta}^*)] \geq \frac{\kappa}{2} \|\Delta_{\boldsymbol{\theta}}\|_2^2$ for all $\Delta_{\boldsymbol{\theta}} \in \mathbb{B}(R)$ for some radius R, where $\mathbb{B}(R)$ is the ball with radius R defined by L_2 norm.

Algorithm 3: Implementation procedure for the estimation and inference of high-dimensional policy parameters

Input : n samples, the number of bootstrap iterations B, the confidence level $1-\alpha$

Output: $\tilde{\boldsymbol{\theta}}$ and the confidence intervals for each coefficient

Solve (3.10) using Algorithm 4 to obtain an initial estimator $\check{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, where $\lambda_{\check{\boldsymbol{\theta}}}$ is tuned by cross-validation;

Estimate $\hat{Q}_t$ for $t = T, \ldots, 1$ with the initial estimator $\check{\boldsymbol{\theta}}$ by (3.14) using solvers for Lasso and re-estimate it on the support, where $\lambda_{\boldsymbol{\beta}_1}, \ldots, \lambda_{\boldsymbol{\beta}_T}$ are tuned by cross-validation at each stage t;

Solve (3.11) using Algorithm 4 to obtain the policy parameter $\hat{\boldsymbol{\theta}}$ with $\hat{Q}_t$, where $\lambda_{\boldsymbol{\theta}}$ is tuned by cross-validation;

Tune $\lambda_{\boldsymbol{w}}$ by cross-validation with the one standard error rule;

for $j = 1, \ldots, d$ **do**

  Calculate the one-step estimator $\tilde{\theta}_j$ following the steps in (3.5), (3.6), and (3.7) with $\lambda_{\boldsymbol{w}}$ and the gradient and Hessian matrix estimated numerically by (3.16) and (3.20);

end

for $b = 1, \ldots, B$ **do**

  Obtain a bootstrap sample of size n;

  Obtain the bootstrap estimator $\hat{\boldsymbol{\theta}}_{bs}$ by refitting (3.11) with the bootstrap sample, $\hat{Q}_t$ and $\lambda_{\boldsymbol{\theta}}$ using Algorithm 4;

  for $j = 1, \ldots, d$ **do**

    Calculate the one-step estimator $\tilde{\theta}_{b,j}$ following the steps in (3.5), (3.6), and (3.7) with $\lambda_{\boldsymbol{w}}$ and the gradient and Hessian matrix estimated numerically by (3.16) and (3.20);

  end

end

for $j = 1, \ldots, d$ **do**

  Obtain the $(1-\alpha) \times 100\%$ confidence interval of $\tilde{\theta}_j$ by finding the $\alpha/2$ and $1-\alpha/2$ quantiles of $\{\tilde{\theta}_{b,j}\}_{b=1}^{B}$.

end

Assumption 11 requires the boundedness of the variables. The conditions about the covariates $\boldsymbol{X}_t$ follow the example of generalized linear models in Ning and Liu (2017). Assumption 12 deals with the convergence rate of the nuisance parameters. Since we assume that we are using the data from a randomized controlled trial and $\boldsymbol{\mu}$ is known, we do not need the model of Q_t to be correctly specified. Besides, $\hat{Q}_t$ can converge to its limit at any rate. This assumption can be easily satisfied by almost all learning methods, including regularized methods like Lasso, ridge regression or elastic net. Different from the case of Liang et al. (2022) in the single stage, our theorem relies on the assumption that the number of nonzero elements $s_{\boldsymbol{\theta}}$ and $s_{\boldsymbol{w}}$ should be bounded. This assumption is needed to take care of the non-convexity of the loss function. Weaker assumptions in Ning and Liu (2017); Liang et al. (2022) rely on the special structure of their convex loss functions. Assumption 13 is used in the multivariate central limit theorem to prove the asymptotic normality of the score function. Assumption 14 is used to verify the restricted strong convexity, which is one of the sufficient conditions for proving the convergence of parameters regularized by L_1 penalty.

We follow the proof of Ning and Liu (2017) to decorrelate $\boldsymbol{\theta}_{-j}$ from the parameter θ_j that we are interested in, and the assumptions will be used to verify the conditions in Theorem 3.2 (Ning and Liu, 2017). The main challenge lies in the nuisance parameters $\boldsymbol{\eta}$, which need to be estimated and are also high-dimensional. We will use the technique in Chernozhukov et al. (2018a) to decorrelate the nuisance parameters.

Theorem 3.3.1. *Under Assumptions 8-14, the one-step estimator $\tilde{\theta}_j$ satisfies*

$$\sqrt{n}(\tilde{\theta}_j - \theta_j^*) I^*_{\theta_j|\boldsymbol{\theta}_{-j}} \Rightarrow N(0, \sigma_S^*), \tag{3.15}$$

where σ_S^* is defined in (3.8).

Theorem 3.3.1 shows that the plug-in one-step estimator is asymptotically normal. Since $\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}$ is consistent for $I^*_{\theta_j|\boldsymbol{\theta}_{-j}}$ and $\hat{\sigma}_S$ is consistent for σ_S^*, we can construct confidence intervals as in (3.9) based on the theorem. Note that $\boldsymbol{\Sigma}^*$ and thus σ_S^* depend on the nuisance parameter $\hat{\boldsymbol{\eta}}$. Therefore, although the limit and the convergence rate of $\hat{\boldsymbol{\eta}}$ does not affect the convergence rate of $\tilde{\theta}_j$, the limit $\bar{\boldsymbol{\eta}}$ does influence the asymptotic variance of $\tilde{\theta}_j$.

3.4 Simulation Study

In this section, we test our proposed procedure for estimating the confidence intervals of low dimensional parameters in high dimensional settings in two simulated scenarios. Assume that the data are from a SMART, and the action $A_{i,t}$ takes value from $\{-1, 1\}$ with equal probability at each stage t for all patients i. The horizon T is taken to be 1 or 3. The initial states are generated independently and identically such that $\boldsymbol{X}_{i,1} \sim N(\mathbf{0}, \mathbf{I}_d)$ for all $i = 1, \ldots, n$. For stages $t \geq 2$, let $\epsilon_{i,t,j} \overset{i.i.d.}{\sim} N(0, 0.2)$ for all $i = 1, \ldots, n, j = 1, \ldots, d$, where j represents the coordinate of the state vector. The useful variables and the rewards are generated as follows. For $i = 1, \ldots, n$,

$$X_{i,t,1} = 0.6A_{i,t-1}\tilde{X}_{i,t-1,1} + 0.2\tilde{X}_{i,t-1,1} + 0.1\tilde{X}_{i,t-1,2} + \epsilon_{i,t,1}, t = 2, \ldots, T$$
$$X_{i,t,2} = -0.6A_{i,t-1}\tilde{X}_{i,t-1,2} + 0.3\tilde{X}_{i,t-1,1}\tilde{X}_{i,t-1,2} + \epsilon_{i,t,2}, t = 2, \ldots, T$$
$$R_{i,t} = \exp\left\{\frac{1}{2}(X_{i,t+1,1} + X_{i,t+1,2}) - 0.2A_{i,t} - 1\right\}, t = 1, \ldots, T,$$

in Scenario 1, and

$$X_{i,t,1} = 0.5A_{i,t-1}\tilde{X}_{i,t-1,1} + 0.3\tilde{X}_{i,t-1,1} + 0.1\tilde{X}_{i,t-1,2} + \epsilon_{i,t,1}, t = 2, \ldots, T$$
$$X_{i,t,2} = 0.5A_{i,t-1}\tilde{X}_{i,t-1,2} + 0.1\tilde{X}_{i,t-1,1} + 0.3\tilde{X}_{i,t-1,2} + \epsilon_{i,t,2}, t = 2, \ldots, T$$
$$R_{i,t} = X_{i,t+1,1} + X_{i,t+1,2} - 0.5A_{i,t}, t = 1, \ldots, T$$

in Scenario 2. Here $\tilde{X}_{i,t,j}$ is the sequence of exponentially weighted moving average of $X_{i,t,j}$ such that $\tilde{X}_{i,1,j} = X_{i,1,j}$ and $\tilde{X}_{i,t,j} = 0.2\tilde{X}_{i,t-1,j} + 0.8X_{i,t,j}$ for $j = 1, 2$ and $t \geq 2$. The other variables $X_{i,t,j} = \epsilon_{i,t,j}$ for $j = 3, \ldots, d$ are noise variables. Under this scenario, the Markov assumption is violated since the state at each stage depends on the states in all previous stages.

We find the true minimizer $\boldsymbol{\theta}^*$ of the loss function within the class $\mathcal{B}(\mathcal{A})$ by grid-search. In particular, we estimate the value function on an independent test set of size $200,000$ for θ_1, θ_2 on the grids inside the unit ball and let $\theta_0 = \sqrt{1 - \theta_1^2 - \theta_2^2}$. Since the value function on close grid points may be quite similar, we repeat the grid-search process 4 times and average the $\boldsymbol{\theta}^*$'s. Finally, we find $\boldsymbol{\theta}^* = (-0.39, 0.68, -0.62)$ when $T = 1$, and $\boldsymbol{\theta}^* = (-0.45, 0.53, -0.72)$ when $T = 3$ for Scenario 1 and $\boldsymbol{\theta}^* = (-0.57, 0.58, 0.58)$ when $T = 1$, and $\boldsymbol{\theta}^* = (-0.57, 0.58, 0.58)$ when $T = 3$ for Scenario 2.

In this simulation, we experiment with 3 different constructions of the nuisance parameter $\boldsymbol{Q}$. For the first method, we take $\hat{Q}_t^{(0)}(\boldsymbol{x}, a) = 0$ for all $\boldsymbol{x}$ and a, so that $\hat{\boldsymbol{\theta}}$ is actually equal to $\check{\boldsymbol{\theta}}$. The second and third estimators $\hat{Q}_t^{(1)}, \hat{Q}_t^{(2)}$ are found by (3.12) and (3.13) respectively. We use the function `scipy.optimize.minimize` with `method='trust-constr'` in Python to solve the constrained minimization problem when refitting $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ on their support. To estimate the Dantzig type estimator in (3.5), we use the package `cvxpy` in Python to solve the constrained convex minimization problem.

We test our procedure for three different settings, where $n = 500, d = 30$, $d = 800, d = 30$ and $n = 800, d = 50$ respectively. The parameter τ is fixed at 0.1, since our simulation experiments show that a larger τ leads to a soft policy far from the true optimal hard policy, and a smaller τ may cause unstable computation. We take the number of bootstraps to be $B = 100$. The value function of each estimated policy is calculated based on an independent test set of size $10,000$ generated by this policy. We repeat the whole procedure for 100 times for each scenario.

We compare our proposed method with penalized efficient augmentation and relaxation learning (PEARL) (Liang et al., 2022) when $T = 1$. PEARL is a method for estimating the optimal ITR and conducting statistical inference from high-dimensional data in single-stage decision problems. It utilizes the data-splitting method to allow for slow convergence rate of the nuisance parameter estimations. In addition, it also follows the inference procedure in Ning and Liu (2017) to first find a sparse estimator and then obtain the one-step estimator. We denote the sparse estimator by $\hat{\boldsymbol{\theta}}_{\text{PEARL}}$ and the one-step estimator by $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$. We use the package `ITRInference` for implementation (Liang et al., 2022). The package does not provide an inference result for the intercept θ_0 and does not have requirements on the scale of the parameters. Therefore, we keep $\tilde{\theta}_{\text{PEARL},0} = \hat{\theta}_{\text{PEARL},0}$ and then normalize $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$ by its L_2 norm. Since the estimated ITR takes the selected treatment with probability one, the scale of the parameters does not affect the ITR. In this way, we can estimate the coverage probability of $\tilde{\boldsymbol{\theta}}_{PEARL}$ on the same scale of $\boldsymbol{\theta}^*$.

The simulation results for Scenario 1 when $T = 1$ are shown in Tables 3.1. We report the estimated value function, the mean absolute deviations (MADs) and the coverage probabilities (CPs). For the value functions, we report their means and standard deviations of the estimated policies corresponding to the sparse estimator of PEARL $\hat{\boldsymbol{\theta}}_{\text{PEARL}}$, the one-step estimator of PEARL $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$, the proposed sparse estimator $\hat{\boldsymbol{\theta}}$, and the proposed one-step estimator $\tilde{\boldsymbol{\theta}}$. Besides, we report

Table 3.1: Value functions, MADs and CPs of the learnt ITR for $T = 1$ in Scenario 1.

			$n = 500, d = 30$			$n = 800, d = 30$			$n = 800, d = 50$			
$V(\boldsymbol{\theta})$		Train	0.1159 (0.0218)			0.1173 (0.0208)			0.1138 (0.0205)			
		$\hat{\boldsymbol{\theta}}_{\text{PEARL}}$	0.4854 (0.0040)			0.4829 (0.0052)			0.4871 (0.0041)			
		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	0.4822 (0.0048)			0.4723 (0.0059)			0.4839 (0.0044)			
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.4731 (0.0071)			0.4767 (0.0060)			0.4747 (0.0066)			
		$\hat{\boldsymbol{Q}}^{(1)}$	0.4761 (0.0075)			0.4794 (0.0061)			0.4789 (0.0056)			
		$\hat{\boldsymbol{Q}}^{(2)}$	0.4753 (0.0074)			0.4791 (0.0056)			0.4785 (0.0055)			
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.4519 (0.0338)			0.4600 (0.0346)			0.4541 (0.0288)			
		$\hat{\boldsymbol{Q}}^{(1)}$	0.4661 (0.0213)			0.4724 (0.0084)			0.4610 (0.0361)			
		$\hat{\boldsymbol{Q}}^{(2)}$	0.4630 (0.0213)			0.4677 (0.0340)			0.4619 (0.0228)			
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
MAD		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	0.0225	0.0228	0.0174	0.0321	0.0310	0.0294	0.0175	0.0180	0.0135	Estimated
			0.0231	0.0240	0.0209	0.0221	0.0246	0.0294	0.0139	0.0173	0.0145	Empirical
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0338	0.0447	0.0323	0.0265	0.0354	0.0298	0.0280	0.0364	0.0267	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0292	0.0352	0.0273	0.0234	0.0285	0.0234	0.0239	0.0281	0.0225	Estimated
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0284	0.0353	0.0295	0.0218	0.0272	0.0253	0.0237	0.0277	0.0241	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0301	0.0370	0.0349	0.0239	0.0395	0.0303	0.0234	0.0316	0.0290	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0291	0.0317	0.0278	0.0267	0.0277	0.0240	0.0263	0.0245	0.0246	Empirical
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0257	0.0294	0.0315	0.0267	0.0272	0.0265	0.0260	0.0239	0.0266	
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
CP		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	96%	93%	89%	99%	99%	94%	98%	94%	91%	
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	97%	94%	100%	95%	93%	100%	96%	89%	100%	
		$\hat{\boldsymbol{Q}}^{(1)}$	97%	93%	99%	98%	96%	99%	93%	91%	99%	
		$\hat{\boldsymbol{Q}}^{(2)}$	97%	95%	100%	97%	97%	100%	98%	91%	100%	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	98%	97%	99%	98%	100%	99%	99%	95%	99%	
		$\hat{\boldsymbol{Q}}^{(1)}$	94%	98%	97%	95%	96%	96%	98%	97%	97%	
		$\hat{\boldsymbol{Q}}^{(2)}$	93%	99%	98%	95%	97%	97%	97%	99%	98%	

* The notation $\hat{\boldsymbol{\theta}}_{\text{PEARL}}$ represents the sparse estimator of PEARL, $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$ is the one-step estimator of PEARL, $\hat{\boldsymbol{\theta}}$ is the proposed sparse estimator, and $\tilde{\boldsymbol{\theta}}$ is the proposed one-step estimator. The nuisance parameters in $\hat{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}$ can be constructed in three different ways, where $\hat{\boldsymbol{Q}}^{(0)} = 0$, and $\hat{\boldsymbol{Q}}^{(1)}, \hat{\boldsymbol{Q}}^{(2)}$ are obtained by (3.12) and (3.13) respectively.

Table 3.2: Value functions, MADs and CPs of the learnt ITR for $T = 3$ in Scenario 1.

			$n = 500, d = 30$			$n = 800, d = 30$			$n = 800, d = 50$			
$V(\boldsymbol{\theta})$		Train	0.1903 (0.1154)			0.1890 (0.0910)			0.1876 (0.0832)			
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	1.0751 (0.0223)			1.0918 (0.0145)			1.0862 (0.0153)			
		$\hat{\boldsymbol{Q}}^{(1)}$	1.0843 (0.0178)			1.0943 (0.0124)			1.0906 (0.0145)			
		$\hat{\boldsymbol{Q}}^{(2)}$	1.0828 (0.0175)			1.0961 (0.0123)			1.0899 (0.0143)			
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	1.0521 (0.0387)			1.0398 (0.1027)			1.0427 (0.0624)			
		$\hat{\boldsymbol{Q}}^{(1)}$	1.0467 (0.1015)			1.0651 (0.0786)			1.0615 (0.0353)			
		$\hat{\boldsymbol{Q}}^{(2)}$	1.0511 (0.0737)			1.0586 (0.0719)			1.0615 (0.0291)			
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
MAD	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0564	0.0502	0.0359	0.0414	0.0404	0.0347	0.0437	0.0440	0.0298	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0508	0.0452	0.0343	0.0371	0.0366	0.0316	0.0396	0.0395	0.0281	Estimated
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0489	0.0435	0.0346	0.0379	0.0359	0.0323	0.0398	0.0379	0.0280	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0607	0.0421	0.0343	0.0382	0.0416	0.0358	0.0382	0.0400	0.0317	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0541	0.0392	0.0348	0.0319	0.0351	0.0324	0.0379	0.0450	0.0305	Empirical
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0494	0.0350	0.0358	0.0324	0.0310	0.0344	0.0360	0.0384	0.0304	
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
CP	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	94%	91%	100%	94%	92%	99%	91%	94%	100%	
		$\hat{\boldsymbol{Q}}^{(1)}$	94%	95%	99%	91%	95%	99%	88%	95%	99%	
		$\hat{\boldsymbol{Q}}^{(2)}$	94%	94%	99%	91%	94%	99%	89%	95%	100%	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	98%	94%	99%	94%	94%	98%	92%	94%	99%	
		$\hat{\boldsymbol{Q}}^{(1)}$	94%	96%	98%	92%	96%	98%	92%	95%	99%	
		$\hat{\boldsymbol{Q}}^{(2)}$	95%	95%	98%	93%	93%	98%	93%	94%	99%	

Table 3.3: Value functions, MADs and CPs of the learnt ITR for $T = 1$ in Scenario 2.

			$n = 500, d = 30$			$n = 800, d = 30$			$n = 800, d = 50$			
$V(\boldsymbol{\theta})$		Train	0.0053 (0.0427)			0.0034 (0.0337)			0.0049 (0.0349)			
		$\hat{\boldsymbol{\theta}}_{\text{PEARL}}$	0.6883 (0.0082)			0.6873 (0.0080)			0.6882 (0.0080)			
		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	0.6870 (0.0063)			0.6646 (0.0108)			0.6873 (0.0054)			
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.6834 (0.0074)			0.6868 (0.0061)			0.6868 (0.0061)			
		$\hat{\boldsymbol{Q}}^{(1)}$	0.6861 (0.0057)			0.6886 (0.0048)			0.6873 (0.0060)			
		$\hat{\boldsymbol{Q}}^{(2)}$	0.6844 (0.0065)			0.6883 (0.0047)			0.6870 (0.0055)			
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.6654 (0.0191)			0.6761 (0.0074)			0.6684 (0.0098)			
		$\hat{\boldsymbol{Q}}^{(1)}$	0.6768 (0.0093)			0.6812 (0.0068)			0.6761 (0.0080)			
		$\hat{\boldsymbol{Q}}^{(2)}$	0.6726 (0.0090)			0.6790 (0.0060)			0.6724 (0.0077)			
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
MAD		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	0.0212	0.0210	0.0188	0.0332	0.0335	0.0326	0.0165	0.0166	0.0143	Estimated
			0.0200	0.0197	0.0200	0.0329	0.0304	0.0333	0.0165	0.0195	0.0151	Empirical
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0309	0.0305	0.0265	0.0229	0.0226	0.0227	0.0229	0.0228	0.0214	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0229	0.0223	0.0216	0.0171	0.0171	0.0183	0.0176	0.0174	0.0178	Estimated
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0267	0.0269	0.0243	0.0202	0.0203	0.0208	0.0200	0.0197	0.0199	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0215	0.0260	0.0264	0.0230	0.0205	0.0232	0.0205	0.0183	0.0222	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0176	0.0198	0.0223	0.0170	0.0161	0.0183	0.0171	0.0148	0.0177	Empirical
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0195	0.0255	0.0245	0.0213	0.0194	0.0215	0.0191	0.0174	0.0204	
			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
CP		$\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$	72%	77%	92%	92%	98%	93%	65%	56%	92%	
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	95%	95%	100%	93%	91%	100%	95%	91%	100%	
		$\hat{\boldsymbol{Q}}^{(1)}$	94%	92%	99%	85%	92%	99%	93%	95%	99%	
		$\hat{\boldsymbol{Q}}^{(2)}$	95%	93%	100%	92%	91%	100%	97%	91%	100%	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	95%	97%	98%	91%	92%	97%	96%	94%	97%	
		$\hat{\boldsymbol{Q}}^{(1)}$	98%	98%	96%	92%	90%	94%	91%	93%	95%	
		$\hat{\boldsymbol{Q}}^{(2)}$	92%	96%	97%	93%	94%	96%	97%	96%	97%	

Table 3.4: Value functions, MADs and CPs of the learnt ITR for $T = 3$ in Scenario 2.

			$n = 500, d = 30$			$n = 800, d = 30$			$n = 800, d = 50$			
$V(\boldsymbol{\theta})$	Train		0.0111 (0.2468)			0.0041 (0.1803)			-0.0006 (0.1984)			
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	1.8715 (0.0226)			1.8892 (0.0167)			1.8849 (0.0180)			
		$\hat{\boldsymbol{Q}}^{(1)}$	1.8775 (0.0174)			1.8911 (0.0131)			1.8865 (0.0143)			
		$\hat{\boldsymbol{Q}}^{(2)}$	1.8771 (0.0160)			1.8904 (0.0132)			1.8841 (0.0160)			
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	1.8344 (0.1288)			1.8644 (0.0178)			1.8266 (0.1584)			
		$\hat{\boldsymbol{Q}}^{(1)}$	1.8542 (0.0239)			1.8716 (0.0165)			1.8549 (0.0498)			
		$\hat{\boldsymbol{Q}}^{(2)}$	1.8209 (0.1957)			1.8707 (0.0177)			1.8531 (0.0338)			
MAD			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0319	0.0316	0.0305	0.0246	0.0247	0.0263	0.0263	0.0267	0.0260	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0295	0.0289	0.0282	0.0239	0.0236	0.0251	0.0232	0.0239	0.0231	Estimated
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0301	0.0302	0.0293	0.0240	0.0229	0.0257	0.0240	0.0240	0.0235	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	0.0341	0.0300	0.0308	0.0239	0.0207	0.0281	0.0275	0.0238	0.0271	
		$\hat{\boldsymbol{Q}}^{(1)}$	0.0311	0.0318	0.0297	0.0215	0.0207	0.0256	0.0250	0.0223	0.0243	Empirical
		$\hat{\boldsymbol{Q}}^{(2)}$	0.0299	0.0268	0.0311	0.0203	0.0254	0.0269	0.0252	0.0266	0.0254	
CP			θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	θ_1	θ_2	$\theta_{3:d}$	
	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	90%	90%	99%	85%	89%	100%	91%	79%	100%	
		$\hat{\boldsymbol{Q}}^{(1)}$	93%	88%	99%	92%	93%	99%	92%	83%	99%	
		$\hat{\boldsymbol{Q}}^{(2)}$	90%	89%	99%	87%	89%	99%	89%	85%	100%	
	$\tilde{\boldsymbol{\theta}}$	$\hat{\boldsymbol{Q}}^{(0)}$	93%	93%	98%	92%	93%	98%	96%	93%	98%	
		$\hat{\boldsymbol{Q}}^{(1)}$	96%	95%	97%	95%	95%	97%	94%	91%	98%	
		$\hat{\boldsymbol{Q}}^{(2)}$	94%	93%	98%	93%	94%	97%	98%	91%	98%	

the value function on the training set, which is calculated by taking the average total reward of all observations. For the MADs, we report that of the two important variables θ_1, θ_2 and the average MAD of $\theta_3, \ldots, \theta_d$ for $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$ and $\tilde{\boldsymbol{\theta}}$. For the one-step estimator of PEARL $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$, the estimated value is the average estimated MAD based on the asymptotic normal distribution across all replications, calculated by $\tilde{\sigma}_{\text{PEARL}}/1.4826$ where $\tilde{\sigma}_{\text{PEARL}}$ is the estimated standard deviation of $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$. The empirical value is the MAD of one-step estimators $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$ in all replications. For the proposed one-step estimator $\tilde{\boldsymbol{\theta}}$, the estimated value is the average estimated MAD by bootstrap across all replications, and the empirical value is the MAD of one-step estimators $\tilde{\boldsymbol{\theta}}$ in all replications. For the CPs, we report that of the two important variables θ_1, θ_2 and the average CPs of $\theta_3, \ldots, \theta_d$ for $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$, $\check{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. The confidence intervals of the sparse estimator $\hat{\boldsymbol{\theta}}$ are estimated by bootstrap as well. Note that each proposed sparse estimator $\hat{\boldsymbol{\theta}}$ and one-step estimator $\tilde{\boldsymbol{\theta}}$ can be constructed by three different nuisance parameters $\hat{\boldsymbol{Q}}^{(0)}, \hat{\boldsymbol{Q}}^{(1)}, \hat{\boldsymbol{Q}}^{(2)}$ The results for $T = 1$ in Scenario 1 and $T = 1, 3$ in Scenario 2 are included in Tables 3.2, 3.3, 3.4.

In the single-stage settings, the value function of the proposed method is slightly smaller than that of PEARL, since PEARL estimates a hard policy while the proposed method estimates a soft policy. However, the difference is not significant. The one-step estimator $\tilde{\boldsymbol{\theta}}$ may affect the value to a small extent, but it is suitable for estimating the confidence intervals. The estimated MAD of $\tilde{\boldsymbol{\theta}}$ is close to its empirical value, which suggests that the bootstrapped confidence interval is an approximation for the confidence interval of $\tilde{\boldsymbol{\theta}}$. The coverage probabilities of $\tilde{\boldsymbol{\theta}}$ are concentrated near the nominal coverage 95%, while that of the sparse estimator $\hat{\boldsymbol{\theta}}$ can be significantly lower than 95%. This demonstrates the necessity of using the one-step estimator for statistical inference. The MAD of $\tilde{\boldsymbol{\theta}}_{\text{PEARL}}$ is smaller than that of $\tilde{\boldsymbol{\theta}}$, but its coverage probabilities cannot reach 95%.

Compared to the zero estimator $\hat{\boldsymbol{Q}}^{(0)}$ of the nuisance parameter $\boldsymbol{Q}$, we can see that the estimated value functions of the sparse estimators $\hat{\boldsymbol{\theta}}$ of $\hat{\boldsymbol{Q}}^{(1)}$ and $\hat{\boldsymbol{Q}}^{(2)}$ always have larger means and smaller standard deviations, with an exception when $n = 500$ or 800 and $d = 30$ in Scenario 1. In addition, the true MADs of the one step estimators $\tilde{\boldsymbol{\theta}}$ of $\hat{\boldsymbol{Q}}^{(1)}$ and $\hat{\boldsymbol{Q}}^{(2)}$ are usually smaller than that of $\hat{\boldsymbol{Q}}^{(0)}$. This leads to the fact that its coverage probability is close to 1 in Scenario 1 when $T = 1$, indicating the over-estimate of the its confidence intervals. Between $\hat{\boldsymbol{Q}}^{(1)}$ and $\hat{\boldsymbol{Q}}^{(2)}$, we can conclude that they generally have similar performance in Scenario 1, and $\hat{\boldsymbol{Q}}^{(1)}$ is always better than $\hat{\boldsymbol{Q}}^{(2)}$ in Scenario 2. Therefore, the above experiments demonstrate the advantage of the proposed AIPWE of the value

function over the IPW estimator, and the influence of the nuisance parameters on the asymptotic efficiency.

3.5 Discussion

In this work, we focus on the multi-stage decision problem and propose a method for estimating the high-dimensional MSTP and the confidence intervals of its parameters. We first estimate the MSTP based on the AIPWE of the value function with an L_1 penalty to encourage sparsity and an L_2 constraint to avoid non-identifiability. Then we find the one-step estimators which is asymptotically normal and suitable for statistical inference. We show that there is a tradeoff between the value function and the coverage probability. While the sparse MSTP is better for generating treatment suggestions with higher value functions, the one-step estimator is suitable for constructing confidence intervals. The proposed one-step estimator is shown to achieve nominal coverage probabilities in simulation studies. While the choice of the nuisance parameter estimation does not affect the convergence rate of the low-dimensional policy parameter, it affects the asymptotic efficiency. We compare different estimators $\hat{\boldsymbol{Q}}$ in the simulation study, and show that the AIPWE generates higher value function and smaller MAD of the estimated policy than the IPW estimator.

We assume that the behavior policy $\boldsymbol{\mu}$ is known as in randomized controlled trials. When the data from an observational study are used for learning, $\boldsymbol{\mu}$ is unknown and needs to be estimated. For binary treatments with continuous covariates, a common approach is to fit a logistic regression for the behavior policy, possibly with various penalties to handle high-dimensional covariates (Hernán and Robins, 2020). Other nonparametric or machine learning methods can also be used for the estimation.

For problems with a long horizon, similar doubly robust estimators of the value function can be applied as well (Jiang and Li, 2016; Thomas and Brunskill, 2016; Kallus and Uehara, 2020). The results for the inference of MSTP parameters can be extended to these settings. However, our simulation experiments show that the ratio $\rho_{1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}$ can become extremely unstable numerically when t is large and it will affect the construction of a valid confidence interval. How to stabilize the weight when the horizon is long remains an open question. An alternative method based on

the marginalized distribution of the current state and action has been proposed in (Chernozhukov et al., 2018a). However, it requires the Markov assumption which may not be satisfied in reality.

3.6 Supplementary Materials

3.6.1 Implementation Details

3.6.1.1 Optimization with L_1 Penalty and L_2 Constraint

We present the full algorithm of the optimization method for a nonconvex loss function with L_1 penalty and L_2 constraint as described in Section 3.2.3. Notice that to find the global minimizer of a nonconvex function, we need to try different starting point. Since running the full algorithm from multiple starting points is computationally heavy, we compare the function value of multiple points for each coordinate separately at the beginning. In practice, ξ in line 2 can be taken as several discrete values, for example $-0.8, -0.4, 0, 0.4, 0.8$.

3.6.1.2 Numerical Computation of the Gradient and Hessian matrix of Loss Function

The direct estimation of the gradient and Hessian matrix of $\rho_{i,1:t}^{\boldsymbol{\theta},\boldsymbol{\mu}}$ is unstable due to the probability product, which may cause even larger variability when constructing confidence intervals for the one-step estimator and lead to the undercoverage of the true parameter. Therefore, we propose to calculate them numerically. An estimate of the partial derivative at the coordinate $j = 1, \ldots, d$ is calculated using the symmetric difference quotient as

$$\widehat{\nabla}_{\theta_j} \ell(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{\ell(\boldsymbol{\theta}^+, \boldsymbol{\eta}) - \ell(\boldsymbol{\theta}^-, \boldsymbol{\eta})}{2h_j}, \tag{3.16}$$

Algorithm 4: Optimization of a Loss Function with L_1 penalty and L_2 constraint

Input : a loss function ℓ to be minimized with L_1 penalty and L_2 constraint, the tuning parameter λ for the L_1 penalty, a starting point $\boldsymbol{\theta}^{(0)}$, the error threshold e, the maximum number of iterations M

Output: The estimated parameter $\check{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}$

for *each coordinate* $j = 1, \ldots, d$ **do**

Find the best starting point of the coordinate j

$$\theta_j^{(1)} = \underset{\boldsymbol{\theta}_j \in \left\{\boldsymbol{\theta}_j^{(0)} + \xi : \xi \in [-1,1]\right\}}{\arg\min} \ell((\theta_1^{(0)}, \ldots, \theta_j, \ldots, \theta_d^{(0)}));$$

end

Normalize $\boldsymbol{\theta}^{(1)}$ by its ℓ_2-norm;

Initialize the number of iterations $m = 1$;

while $m \leq M$ *and* $\|\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}\|_2 \geq e$ **do**

$m \leftarrow m + 1$;

for *each coordinate* $j = 1, \ldots, d$ **do**

Find the minimizer of ℓ with respect to the current coordinate using BFGS algorithm

$$\tilde{\theta}_j^{(m)} = \underset{\boldsymbol{\theta}_j}{\arg\min}\, \ell((\theta_1^{(m)}, \ldots, \theta_{j-1}^{(m)}, \theta_j, \theta_{j+1}^{(m-1)}, \ldots, \theta_d^{(m-1)}));$$

Shrink $\tilde{\theta}_j^{(m)}$ using soft-thresholding

$$\theta_j^{(m)} = \text{sign}\{\tilde{\theta}_j^{(m)}\} \max(|\tilde{\theta}_j^{(m)}| - \lambda, 0);$$

Normalize $(\theta_1^{(m)}, \ldots, \theta_j^{(m)}, \theta_{j+1}^{(m-1)}, \ldots, \theta_d^{(m)})$ by its ℓ_2-norm;

end

end

Re-estimate $\check{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}$ by minimizing the loss function ℓ with the L_2 constraint on the support of $\boldsymbol{\theta}^{(m)}$ using the trust-region constrained algorithm.

where

$$h_j = \min\left\{\frac{1}{\sqrt{nT}}, \sqrt{1-\sum_{l\neq 0,j}\theta_l^2} - |\theta_j|\right\},$$

$$\theta_j^+ = \theta_j + h_j,$$

$$\theta_0^+ = \text{sign}\{\theta_0\}\sqrt{1-\sum_{l\neq 0,j}\theta_l^2 - (\theta_j^+)^2}, \tag{3.17}$$

$$\theta_l^+ = \theta_l \quad \text{for } l \neq 0, j, \tag{3.18}$$

and $\boldsymbol{\theta}^-$ can be defined similarly with $\theta_j^- = \theta_j - h_j$. Here $1/\sqrt{nT}$ is the common value for h_j taken in numerically computations. However, since the ℓ_2-norm is restricted to be 1 in our case, h_j sometimes needs to be even smaller depending on the initial estimate θ_j. If $\theta_0 = 0$, then we have $\sqrt{1-\sum_{l\neq 0,j}\theta_l^2} - |\theta_j| = 0$. In this case, when $\theta_j \neq 0$, we estimate the gradient using Newton's difference quotient as

$$\widehat{\nabla}_{\theta_j}\ell(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\ell(\boldsymbol{\theta}^+,\boldsymbol{\eta}) - \ell(\boldsymbol{\theta},\boldsymbol{\eta})}{-h_j\,\text{sign}\{\theta_j\}},$$

where

$$h_j = \min\left\{\frac{1}{\sqrt{nT}}, 2\,|\theta_j|\right\},$$

$$\theta_j^+ = \theta_j - h_j\,\text{sign}\{\theta_j\}$$

and the other coordinates θ_l^+ for $l \neq j$ are calculated as (3.17) and (3.18). When $\theta_0 = \theta_j = 0$, we use the gradient of $\boldsymbol{\theta}'$ for approximation, where

$$\theta_0' = 0, \qquad \theta_j' = \frac{1}{\sqrt{nT}}, \qquad \theta_l' = \theta_l\sqrt{1-(\theta_j')^2} \text{ for } l \neq 0, j. \tag{3.19}$$

Similarly, we use the symmetric difference quotient to find an estimate of the Hessian matrix

$$\widehat{\nabla}_{\theta_j\theta_k}\ell(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\ell(\boldsymbol{\theta}^{++},\boldsymbol{\eta}) - \ell(\boldsymbol{\theta}^{+-},\boldsymbol{\eta}) - \ell(\boldsymbol{\theta}^{-+},\boldsymbol{\eta}) + \ell(\boldsymbol{\theta}^{--},\boldsymbol{\eta})}{(2h)^2}. \tag{3.20}$$

When $j \neq k$,

$$h_{jk} = \min\left\{\frac{1}{\sqrt{nT}}, \frac{1-\sqrt{\sum_{l\neq 0}\theta_l^2}}{\sqrt{2}}\right\},$$

$$\theta_j^{++} = \theta_j + h_{jk}, \qquad \theta_k^{++} = \theta_k + h_{jk},$$

$$\theta_0^{++} = \operatorname{sign}\{\theta_0\}\sqrt{1-\sum_{l\neq 0,j,k}\theta_l^2 - (\theta_j^{++})^2 - (\theta_k^{++})^2}, \tag{3.21}$$

$$\theta_l^{++} = \theta_l \quad \text{for } l \neq 0, j, k, \tag{3.22}$$

and $\boldsymbol{\theta}^{+-}, \boldsymbol{\theta}^{-+}, \boldsymbol{\theta}^{--}$ can be defined similarly with

$$\theta_j^{+-} = \theta_j + h_{jk}, \qquad \theta_k^{+-} = \theta_k - h_{jk},$$

$$\theta_j^{-+} = \theta_j - h_{jk}, \qquad \theta_k^{-+} = \theta_k + h_{jk},$$

$$\theta_j^{--} = \theta_j - h_{jk}, \qquad \theta_k^{--} = \theta_k - h_{jk}.$$

When $j = k$,

$$h_{jj} = \min\left\{\frac{1}{\sqrt{nT}}, \frac{1}{2}\left(\sqrt{1-\sum_{l\neq 0,j}\theta_l^2} - |\theta_j|\right)\right\},$$

$$\theta_j^{++} = \theta_j + 2h_{jj},$$

$$\theta_0^{++} = \sqrt{1-\sum_{l\neq 0,j}\theta_l^2 - (\theta_j^{++})^2}, \tag{3.23}$$

$$\theta_l^{++} = \theta_l \quad \text{for } l \neq 0, j, k, \tag{3.24}$$

and $\boldsymbol{\theta}^{+-}, \boldsymbol{\theta}^{-+}, \boldsymbol{\theta}^{--}$ can be defined similarly with

$$\theta_j^{+-} = \theta_j^{-+} = \theta_j, \qquad \theta_j^{--} = \theta_j - 2h_{jj}.$$

If $\theta_0 = 0$ and $\theta_j, \theta_k \neq 0$, we instead use Newton's difference quotient to estimate the Hessian matrix as

$$\widehat{\nabla}_{\theta_j\theta_k}\ell(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{\ell(\boldsymbol{\theta}^{++}, \boldsymbol{\eta}) - \ell(\boldsymbol{\theta}^{+\cdot}, \boldsymbol{\eta}) - \ell(\boldsymbol{\theta}^{\cdot+}, \boldsymbol{\eta}) + \ell(\boldsymbol{\theta}, \boldsymbol{\eta})}{(h_{jk}\operatorname{sign}\{\theta_j\})(h_{jk}\operatorname{sign}\{\theta_k\})}.$$

When $j \neq k$,

$$h_{jk} = \min\left\{\frac{1}{\sqrt{nT}}, 2\left|\theta_j\right|, 2\left|\theta_k\right|\right\},$$
$$\theta_j^{++} = \theta_j - h_{jk}\,\mathrm{sign}\{\theta_j\}, \qquad \theta_k^{++} = \theta_k - h_{jk}\,\mathrm{sign}\{\theta_k\},$$

and the other coordinates θ_l^{++} for $l \neq 0$ are calculated as (3.21) and (3.22). Similarly, $\boldsymbol{\theta}^{+-}, \boldsymbol{\theta}^{-+}$ can be defined with

$$\theta_j^{+\cdot} = \theta_j + h_{jk}, \qquad \theta_k^{+\cdot} = \theta_k,$$
$$\theta_j^{\cdot+} = \theta_j, \qquad \theta_k^{\cdot+} = \theta_k + h_{jk}.$$

When $j = k$,

$$h_{jj} = \min\left\{\frac{1}{\sqrt{nT}}, \left|\theta_j\right|\right\},$$
$$\theta_j^{++} = \theta_j - 2h_{jj}\,\mathrm{sign}\{\theta_j\},$$

and the other coordinates θ_l^{++} for $l \neq 0$ are calculated as (3.23) and (3.24). Similarly, $\boldsymbol{\theta}^{+\cdot}, \boldsymbol{\theta}^{\cdot+}$ can be defined with

$$\theta_j^{+\cdot} = \theta_j^{\cdot+} = \theta_j - h_{jj}\,\mathrm{sign}\{\theta_j\}.$$

When $\theta_0 = \theta_j = 0$ and $\theta_k \neq 0$, we use the gradient of $\boldsymbol{\theta}'$ for approximation, where $\boldsymbol{\theta}'$ is defined in (3.19). When $\theta_0 = \theta_j = \theta_k = 0$, we use the gradient of $\boldsymbol{\theta}''$ for approximation, where

$$\theta_0'' = 0, \qquad \theta_j'' = \frac{1}{\sqrt{nT}}, \qquad \theta_k'' = \frac{1}{\sqrt{nT}}, \qquad \theta_l'' = \theta_l\sqrt{1-(\theta_k'')^2-(\theta_k'')^2} \text{ for } l \neq 0, j, k.$$

3.6.2 Proof of Theorem 3.3.1

Note that the gradient and Hessian matrix of the loss function are

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\frac{\nabla_{\boldsymbol{\theta}}[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})]}{\prod_{k=1}^{t}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}[R_{i,t}-\hat{Q}_t(\boldsymbol{X}_{i,t},A_{i,t})]\right.$$
$$+\frac{\nabla_{\boldsymbol{\theta}}[\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})]}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}\sum_{a\in\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\hat{Q}_t(\boldsymbol{X}_{i,t},a),$$
$$\left.+\frac{\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}\sum_{a\in\mathcal{A}}\nabla_{\boldsymbol{\theta}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\hat{Q}_t(\boldsymbol{X}_{i,t},a)\right\},$$

$$\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{\frac{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})]}{\prod_{k=1}^{t}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}[R_{i,t}-\hat{Q}_t(\boldsymbol{X}_{i,t},A_{i,t})]\right.$$
$$+\frac{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}[\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})]}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}\sum_{a\in\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\hat{Q}_t(\boldsymbol{X}_{i,t},a),$$
$$+\frac{\nabla_{\boldsymbol{\theta}}[\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})]}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}\sum_{a\in\mathcal{A}}\nabla_{\boldsymbol{\theta}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\hat{Q}_t(\boldsymbol{X}_{i,t},a),$$
$$\left.+\frac{\prod_{k=1}^{t-1}\pi^{\boldsymbol{\theta}}(A_{i,k}|\boldsymbol{X}_{i,k})}{\prod_{k=1}^{t-1}\mu_k(A_{i,k}|\boldsymbol{H}_{i,k})}\sum_{a\in\mathcal{A}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{X}_{i,t})\hat{Q}_t(\boldsymbol{X}_{i,t},a)\right\},$$

where

$$\nabla_{\boldsymbol{\theta}}\left[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(a_k|\boldsymbol{x}_k)\right] = \left[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(a_k|\boldsymbol{x}_k)\right]\left[\sum_{k=1}^{t}\frac{a_k\boldsymbol{x}_k/\tau}{1+e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}}\right],$$
$$\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\left[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(a_k|\boldsymbol{x}_k)\right] = \left[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(a_k|\boldsymbol{x}_k)\right]\left[\sum_{k=1}^{t}\frac{a_k\boldsymbol{x}_k/\tau}{1+e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}}\right]\left[\sum_{k=1}^{t}\frac{a_k\boldsymbol{x}_k/\tau}{1+e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}}\right]^T$$
$$-\left[\prod_{k=1}^{t}\pi^{\boldsymbol{\theta}}(a_k|\boldsymbol{x}_k)\right]\sum_{k=1}^{t}e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}\left[\frac{a_k\boldsymbol{x}_k/\tau}{1+e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}}\right]\left[\frac{a_k\boldsymbol{x}_k/\tau}{1+e^{a_k\boldsymbol{x}_k^T\boldsymbol{\theta}/\tau}}\right]^T.$$

Since R_t is bounded for all t by Assumption 11, we know that $\bar{\boldsymbol{\eta}}$ is bounded by r with probability one. For a real valued function $f:\mathcal{D}\mapsto\mathbb{R}$, write the empirical process as

$$\mathbb{G}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f(\boldsymbol{D}_i)-\mathbb{E}f(\boldsymbol{D}_i)].$$

Denote $[\cdot]_j$ as the jth dimension of a vector and $[\cdot]_{j\cdot}$ the jth row of a matrix.

For $\bar{\mathcal{H}} := \{\boldsymbol{\eta} - \bar{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H}\}$ define the pathwise derivative of the nuisance parameter $D_q : \bar{\mathcal{H}} \mapsto \mathbb{R}^d$,

$$D_q[\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}] := \nabla_q \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}} + q(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}))\right], \qquad \boldsymbol{\eta} \in \mathcal{H},$$

for all $q \in [0, 1)$. This derivative exists by our construction of ℓ. We will use the Neyman orthogonality to decorrelate the high-dimensional nuisance parameters. The definition is taken from Chernozhukov et al. (2018a).

Definition 3.6.1 (Neyman orthogonality)**.** The score $\nabla_{\boldsymbol{\theta}} \ell$ obeys the orthogonality condition at $(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})$ with respect to the nuisance realization set $\mathcal{H}_n \in \mathcal{H}$ if

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})] = 0$$

and the pathwise derivative map $D_q[\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}]$ exists for all $q \in [0, 1)$ and $\boldsymbol{\eta} \in \mathcal{H}_n$ and vanishes at $q = 0$; namely,

$$D_0[\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}] = \mathbf{0}, \qquad \text{for all } \boldsymbol{\eta} \in \mathcal{H}_n.$$

Lemma 3.6.1. *Under Assumptions 8, 9, 10 and 11, the gradient $\nabla_{\boldsymbol{\theta}} \ell$ satisfies the Neyman orthogonality. In addition, the following results hold for the nuisance parameters:*

$$\sup_{\boldsymbol{\eta} \in \mathcal{H}_n, q \in (0,1)} \left\| \nabla^2_{qq} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}} + q(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}))] \right\|_\infty = 0, \tag{3.25}$$

$$\sup_{\boldsymbol{\eta} \in \mathcal{H}_n, q \in (0,1)} \left\| \nabla_q \mathbb{E}[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}} + q(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}))] \right\|_\infty = 0. \tag{3.26}$$

Proof. Since $\mathbb{E}[\ell(\boldsymbol{\theta}, \boldsymbol{\eta})] = -V(\boldsymbol{\theta})$ for any $\boldsymbol{\eta}$ and $\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}^*) = 0$ by the definition of $\boldsymbol{\theta}^*$, we have $\mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta})] = 0$ for any $\boldsymbol{\eta}$, when the regularity conditions are satisfied. Therefore,

$$D_0[\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}] = \nabla_{\boldsymbol{Q}} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta})]|_{\boldsymbol{\eta} = \bar{\boldsymbol{\eta}}} \cdot (\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}) = \mathbf{0}$$

since $\nabla_{\boldsymbol{Q}} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta})] = \mathbf{0}$.

Similarly, for each dimension j we also have $\nabla^2_{\boldsymbol{QQ}}\mathbb{E}[\nabla_{\theta_j}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta})] = \mathbf{0}$ since $\mathbb{E}[\nabla_{\theta_j}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta})]$ is a constant zero. Consequently,

$$\begin{aligned}
&\left[\nabla^2_{qq}\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}+q(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}))]\right]_j \\
=&\nabla^2_{qq}\mathbb{E}[\nabla_{\theta_j}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}+q(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}))] \\
=&(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}})^T\nabla^2_{\boldsymbol{QQ}}\mathbb{E}[\nabla_{\theta_j}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}+q(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}))](\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}) \\
=&0
\end{aligned}$$

and (3.25) follows. Since $\mathbb{E}[\nabla_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta})] = \nabla_{\boldsymbol{\theta\theta}}V(\boldsymbol{\theta}^*)$ for any $\boldsymbol{\eta}$, we have that

$$\nabla_q\mathbb{E}[\nabla_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}+q(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}))] = \nabla_q\nabla_{\boldsymbol{\theta\theta}}V(\boldsymbol{\theta}^*) = 0$$

and thus (3.26) follows.

□

Lemma 3.6.2. *Under Assumptions 8, 9, 10, 11 and 12, we have*

$$\sup_{\boldsymbol{\eta}\in\mathcal{H}_n}\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})\|_\infty = o_{\mathbb{P}}(1/\sqrt{n}), \tag{3.27}$$

$$\sup_{\boldsymbol{\eta}\in\mathcal{H}_n}\left\|\boldsymbol{v}^{*T}\nabla^2_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \boldsymbol{v}^{*T}\nabla^2_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})\right\|_\infty = o_{\mathbb{P}}(\sqrt{\log d/n}). \tag{3.28}$$

Proof. To show (3.27), first note that

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})\|_\infty \leq& \frac{1}{\sqrt{n}}\|\mathbb{G}_n[\nabla_{\boldsymbol{\theta}}l(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}l(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})]\|_\infty &(3.29)\\
&+ \|\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})]\|_\infty &(3.30)
\end{aligned}$$

for any $\boldsymbol{\eta}\in\mathcal{H}_n$. We will bound (3.30) using Neyman orthogonality and bound (3.29) using the results of empirical process.

To bound (3.30), define

$$\boldsymbol{h}_{\boldsymbol{\eta}}(q) := \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}+q(\boldsymbol{\eta}-\bar{\boldsymbol{\eta}})) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})]$$

for $q \in [0, 1)$. By Taylor's expansion, there exists $\tilde{q} \in (0, 1)$ such that

$$\boldsymbol{h}_{\boldsymbol{\eta}}(1) = \boldsymbol{h}_{\boldsymbol{\eta}}(0) + \nabla_q \boldsymbol{h}_{\boldsymbol{\eta}}(0) + \frac{1}{2}\nabla^2_{qq}\boldsymbol{h}_{\boldsymbol{\eta}}(\tilde{q}).$$

We have $\boldsymbol{h}_{\boldsymbol{\eta}}(0) = \mathbf{0}$ by definition and $\nabla_q \boldsymbol{h}_{\boldsymbol{\eta}}(0) = \mathbf{0}$ by Lemma 3.6.1. In addition, the second derivative satisfies $\|\nabla^2_{qq}\boldsymbol{h}_{\boldsymbol{\eta}}(\tilde{q})\|_\infty = 0$ by Lemma 3.6.1. Therefore, $\|\boldsymbol{h}_{\boldsymbol{\eta}}(1)\|_\infty = 0$ for any $\boldsymbol{\eta} \in \mathcal{H}_n$ and thus

$$\sup_{\boldsymbol{\eta}\in\mathcal{H}_n} \|\mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]\|_\infty = o(1/\sqrt{n}).$$

To bound (3.29), define

$$\boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(q) = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1 + q(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1)$$

for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{H}_n$. Then we have

$$\boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(1) = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_2) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1).$$

By Taylor's expansion, there exists $\tilde{q} \in (0, 1)$ such that

$$\boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(1) = \boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(0) + \nabla_q \boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(\tilde{q}) = \nabla^2_{\boldsymbol{\theta},\boldsymbol{\eta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1 + \tilde{q}(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1))(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)$$

since $\boldsymbol{g}_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(0) = 0$ by definition. Hence by Cauchy-Schwartz inequality we have

$$[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_2) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1)]_j \leq \|[\nabla^2_{\boldsymbol{\theta},\boldsymbol{\eta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1 + \tilde{q}(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1))]_{j\cdot}\|_2 \, \|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\|_2 \tag{3.31}$$

for the jth dimension, which implies that the functions $[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta})]_j$ and $[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]_j$ are Lipschitz in the parameter $\boldsymbol{\eta}$. Note that with Assumption 11 and the boundedness of $\hat{Q}_t$, we have

$$\|\nabla^2_{\boldsymbol{\theta},\boldsymbol{\eta}}\ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}_1 + \tilde{q}(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1))\|_2 \leq C,$$

where C is a constant.

Therefore, the bracketing number $N_{[]}(\epsilon, \mathcal{G}_{n,j}, L_2(\mathbb{P}))$ of the function set

$$\mathcal{G}_{n,j} := \{[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]_j : \boldsymbol{\eta} \in \mathcal{H}_n\}$$

is upper bounded by the covering number $N(\epsilon/(2C), \mathcal{H}_n, \|\cdot\|_2)$ of the nuisance parameter set $\mathcal{H}_n$ (Van der Vaart and Wellner, 1996, Theorem 2.7.11). Since $\|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|_2 = o_{\mathbb{P}}(1)$, the covering number of the nuisance set $N(\epsilon/(2C), \mathcal{H}_n, \|\cdot\|_2)$ is finite. Let

$$J_{[]}(\delta, \mathcal{G}_{n,j}, \|\cdot\|) := \int_0^{\delta} \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{G}_{n,j}, \|\cdot\|)} d\epsilon$$

be the bracketing integral. The minimum envelop function of class $\mathcal{G}_{n,j}$ is defined as $G_{n,j}(x) := \sup_{g \in G_{n,j}} |g(x)|$. Suppose $\mathcal{G}_{n,j}$ is covered by the brackets $[l_1, u_1], \ldots, [l_{N_\epsilon}, u_{N_\epsilon}]$, where $N_\epsilon := N_{[]}(\epsilon, \mathcal{G}_{n,j}, L_2(\mathbb{P}))$ is the bracketing number for any $\epsilon > 0$. Then we can write the minimum envelop function as

$$G_{n,j}(x) = \max_{j=1,\ldots,N_\epsilon} \{|l_{N_\epsilon}(x)|, |u_{N_\epsilon}(x)|\}.$$

The $L_{\mathbb{P},2}$ norm of $G_{n,j}(x)$ is then

$$\|G_{n,j}\|_{\mathbb{P},2}^2 = \mathbb{E} G_{n,j}^2(\boldsymbol{D}) \leq \mathbb{E} \sum_{j=1}^{N_\epsilon} (l_j^2(\boldsymbol{D}) + u_j^2(\boldsymbol{D})).$$

Note that

$$\begin{aligned}
\mathbb{E} l_j^2(\boldsymbol{D}) =& \mathbb{E}[g(\boldsymbol{D}) + (l_j(\boldsymbol{D}) - g(\boldsymbol{D}))]^2 \\
\leq& 2[\mathbb{E} g^2(\boldsymbol{D}) + \mathbb{E}(l_j(\boldsymbol{D}) - g(\boldsymbol{D}))^2] \\
\leq& 2[\mathbb{E} g^2(\boldsymbol{D}) + \epsilon^2]
\end{aligned}$$

for some $g \in \mathcal{G}_{n,j}$ contained in the jth bracket for any $j = 1, \ldots, N_\epsilon$. According to (3.31),

$$\mathbb{E} g^2(\boldsymbol{D}) \leq C^2 \|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|_{\mathbb{P},2} \leq \delta_n$$

by Assumption 12 and thus

$$\mathbb{E}l_j^2(\boldsymbol{D}) \leq 2[\epsilon^2 + \delta_n].$$

The same holds for upper brackets u_j. Since

$$N_\epsilon \leq N\left(\frac{\epsilon}{2C}, \mathcal{H}_n, \|\cdot\|_2\right) \leq \delta_n \frac{2C}{\epsilon},$$

we get that

$$\|G_{n,j}\|_{\mathbb{P},2}^2 \leq 4[\epsilon^2 + \delta_n]N_\epsilon \leq 4[\epsilon^2 + \delta_n] \cdot 2C\delta_n \frac{1}{\epsilon}.$$

So $\|G_{n,j}\|_{\mathbb{P},2}^2 \to 0$ when $n \to \infty$. Since

$$\mathbb{E}^* \sup_{g \in \mathcal{G}_{n,j}} \mathbb{G}_n(g) \lesssim J_{[]}(\|G_{n,j}\|_{\mathbb{P},2}, \mathcal{G}_{n,j}, L_2(\mathbb{P})) \tag{3.32}$$

by Van der Vaart and Wellner (1996, Theorem 2.14.2), the left-hand side of (3.32) is in the order of $o(1)$. Finally, by Markov's inequality,

$$\sup_{\boldsymbol{\eta} \in \mathcal{H}_n} \frac{1}{\sqrt{n}} \mathbb{G}_n[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]_j = o_{\mathbb{P}}(1/\sqrt{n}).$$

The bound on (3.29) follows by taking the maximum over all the dimensions.

Combining the upper bounds of (3.30) and (3.29) we can show (3.27).

For (3.28), note that

$$\left\|\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\right\|_\infty \tag{3.33}$$

$$\leq \frac{1}{\sqrt{n}} \left\|\mathbb{G}_n[\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]\right\|_\infty \tag{3.34}$$

$$+ \left\|\mathbb{E}[\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]\right\|_\infty \tag{3.35}$$

for any $\boldsymbol{\eta} \in \mathcal{H}_n$.

To bound (3.35), define

$$\tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(q) := \mathbb{E}[\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}} + q(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})) - \boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})]$$

for $q \in [0,1)$. By Taylor's expansion, there exists $\tilde{q} \in (0,1)$ such that

$$\tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(1) = \tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(0) + \nabla_q \tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(\tilde{q}).$$

We have $\tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(0) = \mathbf{0}$ by definition and $\nabla_q \tilde{\boldsymbol{h}}_{\boldsymbol{\eta}}(\tilde{q}) = o(\sqrt{\log d/n})$ by Lemma 3.6.1.

Using similar arguments as that for (3.29), we can conclude that (3.34) is in the order of $o_{\mathbb{P}}(1/\sqrt{n})$. Combining the upper bounds of (3.35) and (3.34) gives the results. □

Lemma 3.6.3 (Concentration of the gradient and Hessian)**.** *Under Assumptions 8, 9, 10, 11 and 12, we have*

$$\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n}), \tag{3.36}$$

$$\left\|\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) - \mathbb{E}(\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}))\right\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n}). \tag{3.37}$$

Proof. Since $\mathbb{P}(\hat{\boldsymbol{\eta}} \notin \mathcal{H}_n) \leq \Delta_n$ and Δ_n converges to zero, we can focus on the event when $\boldsymbol{\eta} \in \mathcal{H}_n$. To prove (3.36), note that

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\|_\infty \\
\leq &\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\|_\infty + \|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\|_\infty .
\end{aligned}$$

By (3.27), we only need to show that

$$\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n}). \tag{3.38}$$

Write

$$\boldsymbol{h}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n) := -\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}),$$

which is a d-dimensional real-valued function, and denote the j-th dimension of $\boldsymbol{h}$ as h_j. Remember that $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}})$ can be divided into 3 parts, so

$$\begin{aligned}
\boldsymbol{h}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \Big\{ &\nabla_{\boldsymbol{\theta}} \rho^{\boldsymbol{\theta}, \bar{\boldsymbol{\mu}}}_{i,1:t} [R_{i,t} - \bar{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})] + [\nabla_{\boldsymbol{\theta}} \rho^{\boldsymbol{\theta}, \bar{\boldsymbol{\mu}}}_{i,1:(t-1)}] \bar{U}_t(\boldsymbol{X}_{i,t}) \\
&+ \rho^{\boldsymbol{\theta}, \bar{\boldsymbol{\mu}}}_{i,1:(t-1)} [\nabla_{\boldsymbol{\theta}} \bar{U}_t(\boldsymbol{X}_{i,t})] \Big\}.
\end{aligned}$$

Since

$$\rho_{i,1:t}^{\boldsymbol{\theta},\bar{\boldsymbol{\mu}}} \le 1/p_0^t \qquad \text{and} \qquad \frac{A_{i,k}X_{i,k,j}/\tau}{1+e^{A_{i,k}\boldsymbol{X}_{i,k}^T\boldsymbol{\theta}/\tau}} \le z/\tau,$$

we have that the jth dimension of $\nabla_{\boldsymbol{\theta}}\rho_{i,1:t}^{\boldsymbol{\theta},\bar{\boldsymbol{\mu}}}$ is upper bounded by $\frac{zt}{\tau p_0^t}$. Besides, we know that $|R_{i,t} - \bar{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})| \le 2r$ since $\bar{Q}_t$ is bounded by r according to Assumption 11 and the definition of $\bar{Q}_t$. Thus the jth dimension of the first part is changed by

$$\frac{1}{nT}\sum_{t=1}^{T}\left|[\nabla_{\boldsymbol{\theta}}\rho_{i,1:t}^{\boldsymbol{\theta},\bar{\boldsymbol{\mu}}}]_j[R_{i,t} - \bar{Q}_t(\boldsymbol{X}_{i,t}, A_{i,t})] - [\nabla_{\boldsymbol{\theta}}\rho_{i,1:t}^{\prime\boldsymbol{\theta},\bar{\boldsymbol{\mu}}}]_j[R'_{i,t} - \bar{Q}_t(\boldsymbol{X}'_{i,t}, A'_{i,t})]\right| \le \frac{2}{n}\sum_{t=1}^{T}\frac{2rzt}{\tau p_0^t}$$

if the ith trajectory $\boldsymbol{D}_i$ is changed into $\boldsymbol{D}'_i$. Similarly, since $|[\nabla_{\boldsymbol{\theta}}\pi^{\boldsymbol{\theta}}(a|\boldsymbol{x})]_j| \le \frac{z}{4\tau}$, we have

$$|[\nabla_{\boldsymbol{\theta}}\rho_{i,1:(t-1)}^{\boldsymbol{\theta},\bar{\boldsymbol{\mu}}}]_j\bar{U}_t(\boldsymbol{X}_{i,t})| \le \frac{rtz}{p_0\tau} \qquad \text{and} \qquad |\rho_{i,1:(t-1)}^{\boldsymbol{\theta},\bar{\boldsymbol{\mu}}}[\nabla_{\boldsymbol{\theta}}\bar{U}_t(\boldsymbol{X}_{i,t})]_j| \le \frac{rz}{2p_0\tau}.$$

Therefore, the upper bound on the change of h_j when changing $\boldsymbol{D}_i$ is $c_i := C/n$ for some constant $C > 0$ depending on r, z, p_0, τ, T. By McDiarmid's inequality, we get that

$$\mathbb{P}\left(|h_j(\boldsymbol{D}_1,\ldots,\boldsymbol{D}_n) - \mathbb{E}h_j(\boldsymbol{D}_1,\ldots,\boldsymbol{D}_n)| \ge \epsilon\right) \le 2\exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right\} \le 2\exp\left\{-C'n\epsilon^2\right\}$$

for some constant $C' > 0$. Now the union bound inequality yields

$$\mathbb{P}\left(\|\boldsymbol{h}(\boldsymbol{D}_1,\ldots,\boldsymbol{D}_n) - \mathbb{E}\boldsymbol{h}(\boldsymbol{D}_1,\ldots,\boldsymbol{D}_n)\|_\infty \ge \epsilon\right) \le 2d\exp\left\{-C'n\epsilon^2\right\}.$$

With $\epsilon = C''\sqrt{\log d/n}$ for some $C'' > 0$, we have $\boldsymbol{h}(\boldsymbol{D}_1,\ldots,\boldsymbol{D}_n) = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n})$ and equation (3.38) follows.

Similarly, by (3.28), we only need to prove

$$\left\|\boldsymbol{v}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) - \mathbb{E}(\boldsymbol{v}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}))\right\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n})$$

for (3.37). Now $\boldsymbol{v}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}})$ can be divided into 5 parts. When $\bar{Q}_t$ is upper bounded by the constant r, each dimension for each part is bounded by C''/n for some constant $C'' > 0$, since v^* is a constant. The result follows from the same arguments as before. □

Lemma 3.6.4 (Central limit theorem for the score function)**.** *Under Assumptions 8, 9, 10 and 13, it holds that*

$$\sqrt{n}\boldsymbol{v}^{*T}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) \Rightarrow N(0,\sigma_S^*),$$

where $\sigma_S^* \geq C$ *for some constant* $C > 0$.

Proof. Since $\mathbb{E}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}) = 0$, note that

$$\sqrt{n}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) = \sqrt{n}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})] + \sqrt{n}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})]. \tag{3.39}$$

Th equation (3.27) shows that the first difference on the right-hand side of (3.39) is in the order of $o_{\mathbb{P}}(1)$ when $\hat{\boldsymbol{\eta}} \in \mathcal{H}_n$. Besides, the probability of $\hat{\boldsymbol{\eta}} \notin \mathcal{H}_n$ converges to zero. For the second different in (3.39), when $\boldsymbol{\Sigma}^*$ is finite by Assumption 13, the multivariate central limit theorem (Ferguson, 2017, Theorem 5) shows that

$$\sqrt{n}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}}) - \mathbb{E}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\bar{\boldsymbol{\eta}})] \Rightarrow N(0,\boldsymbol{\Sigma}^*).$$

The convergence follows since $N(0,\boldsymbol{v}^{*T}\boldsymbol{\Sigma}^*\boldsymbol{v}^*)+o_{\mathbb{P}}(1) = N(0,\boldsymbol{v}^{*T}\boldsymbol{\Sigma}^*\boldsymbol{v}^*)$. In addition, Assumption 13 guarantees that $\boldsymbol{v}^{*T}\boldsymbol{\Sigma}^*\boldsymbol{v}^* \geq C$, since $\boldsymbol{v}^{*T}$ is nonzero at least in its first argument. □

Lemma 3.6.5. *Under Assumptions 8, 9, 10, 11, 12 and 14, when n satisfies* $(\tau_{n,\delta} + L\delta_n)s_{\boldsymbol{\theta}} \leq \kappa/2$ *and* $\lambda_{\boldsymbol{\theta}} \simeq \sqrt{\log d/n}$*, we have*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\sqrt{\log d/n}), \tag{3.40}$$

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n). \tag{3.41}$$

Proof. Denote $\hat{\Delta}_{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. According to Chernozhukov et al. (2018b, inequality (88)), we have

$$\begin{aligned}\left\langle \nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}), \hat{\Delta}_{\boldsymbol{\theta}} \right\rangle \leq& \lambda_{\boldsymbol{\theta}}(\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1) = \lambda_{\boldsymbol{\theta}}(\|\boldsymbol{\theta}^*_{\mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\boldsymbol{\theta}^*_{\mathcal{S}_{\boldsymbol{\theta}}} + \hat{\Delta}_{\boldsymbol{\theta},S}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}{}^c}\|_1) \\ \leq& \lambda_{\boldsymbol{\theta}}(\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}{}^c}\|_1).\end{aligned} \tag{3.42}$$

Define the empirical symmetric Bregman distance as

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \boldsymbol{\eta}) = \langle \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \boldsymbol{\eta}), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle .$$

Since all variables are bounded, by Taylor's expansion on each dimension we have

$$\sup_{\boldsymbol{\theta}} \left\| \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \boldsymbol{\eta}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \boldsymbol{\eta}') \right\|_\infty \leq L \left\| \boldsymbol{\eta} - \boldsymbol{\eta}' \right\|_2$$

for any $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{H}_n$. Therefore, the empirical symmetric Bregman distance is Lipschitz in the nuisance parameters:

$$\left| H(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \boldsymbol{\eta}) - H(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \boldsymbol{\eta}') \right| \leq L \left\| \boldsymbol{\eta} - \boldsymbol{\eta}' \right\|_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1^2 \tag{3.43}$$

for any $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{H}_n$ and for any $\boldsymbol{\theta}$ (Chernozhukov et al., 2018b, Lemma 3).

By Assumption 14, we know $\mathbb{E}\ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}) = -V(\boldsymbol{\theta})$ is κ-strongly convex at $\boldsymbol{\theta}^*$. We will use Chernozhukov et al. (2018b, Lemma 2) to prove the restricted strong convexity:

$$H(\boldsymbol{\theta}^* + \Delta_{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) \geq \kappa \|\Delta_{\boldsymbol{\theta}}\|_2^2 - \tau_{n,\delta} \|\Delta_{\boldsymbol{\theta}}\|_1^2 \tag{3.44}$$

holds with probability $1 - \delta$ for all $\Delta_{\boldsymbol{\theta}} \in \mathbb{B}(R)$ and $\tau_{n,\delta} \simeq 1/(\delta\sqrt{n})$. To verify the condition, we need to show that

$$\sup_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^* + \Delta_{\boldsymbol{\theta}} : \Delta_{\boldsymbol{\theta}} \in \mathbb{B}(R)\}} \left\| \boldsymbol{v}^{*T} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}) - \boldsymbol{v}^{*T} \mathbb{E} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}) \right\|_\infty \leq \tau_{n,\delta}$$

with probability $1 - \delta$. Similar as the bound for (3.29), for each dimension j, we can use Taylor's expansion to show that $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}})$ is Lipschitz in the parameter $\boldsymbol{\theta}$ with constant C. Consequently, the bracketing number $N_{[]}(\epsilon, \mathcal{G}_{n,j}, L_2(\mathbb{P}))$ of

$$\mathcal{G}_{n,j} := \{ [\boldsymbol{v}^{*T} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^* + \Delta_{\boldsymbol{\theta}}, \bar{\boldsymbol{\eta}})]_j : \Delta_{\boldsymbol{\theta}} \in \mathbb{B}(R) \}$$

is upper bounded by the covering number $N(\epsilon/(2C), \mathbb{B}(R), \|\cdot\|_2)$. For a finite R, the set $\mathbb{B}(R)$ is bounded and $N(\epsilon, \mathbb{B}(R), \|\cdot\|_2) \simeq (1/\epsilon)^d$. Therefore,

$$\mathbb{E}^* \sup_{g \in \mathcal{G}_{n,j}} \mathbb{G}_n(g) \lesssim J_{[]}(\|G_{n,j}\|_{\mathbb{P},2}, \mathcal{G}_{n,j}, L_2(\mathbb{P})) \leq J_{[]}(\infty, \mathcal{G}_{n,j}, L_2(\mathbb{P})) < \infty$$

by Van der Vaart and Wellner (1996, Theorem 2.14.2). By Markov's inequality and union bound inequality, we get that

$$\tau_{n,\delta} \simeq d J_{[]}(\|G_{n,j}\|_{\mathbb{P},2}, \mathcal{G}_{n,j}, L_2(\mathbb{P}))/(\delta\sqrt{n}).$$

Combining the Lipschitz bound (3.43) and the restricted strong convexity bound (3.44), we have

$$\begin{aligned} H(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) \geq & H(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) - L\|\hat{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}\|_2 \|\hat{\Delta}_{\boldsymbol{\theta}}\|_1^2 \\ \geq & \kappa \|\hat{\Delta}_{\boldsymbol{\theta}}\|_2^2 - (\tau_{n,\delta} + L\|\hat{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}\|_2)\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1^2. \end{aligned}$$

On the other hand,

$$\begin{aligned} H(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) \leq & \left\langle \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}), \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\rangle \\ \leq & \lambda_{\boldsymbol{\theta}} (\|\hat{\Delta}_{\boldsymbol{\theta}, \mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta}, \mathcal{S}_{\boldsymbol{\theta}}^c}\|_1) + \|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}})\|_\infty \|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 \end{aligned}$$

according to (3.42). The assumption that $\lambda_{\boldsymbol{\theta}} \simeq \sqrt{\log d/n}$ and Lemma 3.6.3 implies

$$\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}})\|_\infty \lesssim \lambda_{\boldsymbol{\theta}}/2$$

with high probability. Conditioning on this event, the above two bounds yield

$$\begin{aligned} \kappa \|\hat{\Delta}_{\boldsymbol{\theta}}\|_2^2 \leq & \frac{\lambda_{\boldsymbol{\theta}}}{2} (3\|\hat{\Delta}_{\boldsymbol{\theta}, \mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta}, \mathcal{S}_{\boldsymbol{\theta}}^c}\|_1) + (\tau_{n,\delta} + L\|\hat{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}\|_2)\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1^2 \\ \leq & \frac{3\lambda_{\boldsymbol{\theta}}}{2} \sqrt{s_{\boldsymbol{\theta}}} \|\hat{\Delta}_{\boldsymbol{\theta}}\|_2 + (\tau_{n,\delta} + L\|\hat{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}\|_2) s_{\boldsymbol{\theta}} \|\hat{\Delta}_{\boldsymbol{\theta}}\|_2^2. \end{aligned}$$

Since $\mathbb{P}(\boldsymbol{\eta} \notin \mathcal{H}_n) \to 0$, we focus on the events when $\boldsymbol{\eta} \in \mathcal{H}_n$. When n is large enough such that $(\tau_{n,\delta} + L\delta_n) s_{\boldsymbol{\theta}} \leq \kappa/2$, we have

$$\|\hat{\Delta}_{\boldsymbol{\theta}}\|_2 \leq \frac{3\lambda_{\boldsymbol{\theta}}}{4} \sqrt{s_{\boldsymbol{\theta}}} \simeq \sqrt{s_{\boldsymbol{\theta}} \log d/n}.$$

Besides, since

$$\frac{\lambda_{\boldsymbol{\theta}}}{2}(3\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}{}^c}\|_1) \geq \frac{\kappa}{2}\|\hat{\Delta}_{\boldsymbol{\theta}}\|_2^2 \geq 0,$$

we also have $\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}{}^c}\|_1 \leq 3\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1$ and thus

$$\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 \leq 4\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1 \leq 4\sqrt{s_{\boldsymbol{\theta}}}\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_2 \lesssim s_{\boldsymbol{\theta}}\sqrt{\log d/n}.$$

Now considering the randomness of $\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\|_\infty$ and we have $\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\sqrt{\log d/n})$.

To show (3.41), note that

$$\begin{aligned}
&\left|H(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \hat{\Delta}_{\boldsymbol{\theta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right| \\
=&\left|\hat{\Delta}_{\boldsymbol{\theta}}^T\left[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\right]\hat{\Delta}_{\boldsymbol{\theta}}\right| \\
=&\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1^2\|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\|_\infty
\end{aligned}$$

by Taylor's expansion for some $q \in [0,1]$, where $\hat{\Delta}_{\theta_i} = \hat{\theta}_i - \theta_i^*$. Use Taylor's expansion again and we have

$$\begin{aligned}
&\nabla^2_{\theta_i\theta_j}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\theta_i\theta_j}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) \\
\leq&\|\nabla_{\boldsymbol{\theta}}(\nabla^2_{\theta_i\theta_j}\ell)(q'[q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*]+(1-q')\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\|_\infty\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1
\end{aligned} \tag{3.45}$$

for some $q' \in [0,1]$. Since the first term on the right-hand side of (3.45) is bounded, the left-hand side is in the order of $\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\sqrt{\log d/n})$. By Assumption 12, we have

$$\left|H(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \hat{\Delta}_{\boldsymbol{\theta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right| = \mathcal{O}_{\mathbb{P}}((s_{\boldsymbol{\theta}}\sqrt{\log d/n})^3) = o_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n) \tag{3.46}$$

if $s_{\boldsymbol{\theta}}$ is bounded. Finally, since

$$H(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) \leq \frac{\lambda_{\boldsymbol{\theta}}}{2}(3\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1 - \|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}{}^c}\|_1) \lesssim \lambda_{\boldsymbol{\theta}}\|\hat{\Delta}_{\boldsymbol{\theta},\mathcal{S}_{\boldsymbol{\theta}}}\|_1 = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n),$$

we conclude that $\hat{\Delta}_{\boldsymbol{\theta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}} = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n)$.

□

Lemma 3.6.6. *Under Assumptions 8, 9, 10, 11 and 12, when* $\lambda_{\boldsymbol{w}} \simeq \sqrt{\log d/n}$*, we have*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}((s_{\boldsymbol{\theta}} \vee s_{\boldsymbol{w}})\sqrt{\log d/n}), \tag{3.47}$$

$$(\hat{\boldsymbol{v}} - \boldsymbol{v}^*)^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}})(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = \mathcal{O}_{\mathbb{P}}((s_{\boldsymbol{\theta}} \vee s_{\boldsymbol{w}}) \log d/n). \tag{3.48}$$

Proof. Let $\hat{\Delta}_{\boldsymbol{w}} := \hat{\boldsymbol{w}} - \boldsymbol{w}^*$. Since $\|\boldsymbol{w}^*\|_1 \geq \|\hat{\boldsymbol{w}}\|_1$, we have

$$\sum_{j \in \mathcal{S}_{\boldsymbol{w}}} |w_j^*| \geq \sum_{j \in \mathcal{S}_{\boldsymbol{w}}} |\hat{w}_j| + \sum_{j \in \mathcal{S}_{\boldsymbol{w}}^c} |\hat{w}_j| \geq \sum_{j \in \mathcal{S}_{\boldsymbol{w}}} |w_j^*| - \sum_{j \in \mathcal{S}_{\boldsymbol{w}}} |\hat{\Delta}_{\boldsymbol{w},j}| + \sum_{j \in \mathcal{S}_{\boldsymbol{w}}^c} |\hat{w}_j|.$$

Hence

$$\sum_{j \in \mathcal{S}_{\boldsymbol{w}}} |\hat{\Delta}_{\boldsymbol{w},j}| \geq \sum_{j \in \mathcal{S}_{\boldsymbol{w}}^c} |\hat{w}_j| = \sum_{j \in \mathcal{S}_{\boldsymbol{w}}^c} |\hat{\Delta}_{\boldsymbol{w},j}|,$$

that is, $\|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}^c}\|_1 \leq \|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}}\|_1$. Therefore, $\|\hat{\Delta}_{\boldsymbol{w}}\|_1 \leq 2\|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}}\|_1$. Consider the following function

$$\begin{aligned}
&\hat{\Delta}_{\boldsymbol{w}}^T \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}} \\
=&\left[\nabla^2_{\theta_j \boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \boldsymbol{w}^{*T} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\right] \hat{\Delta}_{\boldsymbol{w}} - \left[\nabla^2_{\theta_j \boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{w}}^T \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\right] \hat{\Delta}_{\boldsymbol{w}} \\
=:& I_1 + I_2.
\end{aligned}$$

According to the definition of $\hat{\Delta}_{\boldsymbol{w}}$,

$$I_2 \leq \|\nabla^2_{\theta_j \boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{w}}^T \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})\|_\infty \|\hat{\Delta}_{\boldsymbol{w}}\|_1 \leq \lambda_{\boldsymbol{w}} \|\hat{\Delta}_{\boldsymbol{w}}\|_1.$$

Note that for I_1,

$$I_1 = \boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}} + \boldsymbol{v}^{*T} [\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}})] \hat{\Delta}_{\boldsymbol{w}} =: I_{11} + I_{12}.$$

Since

$$\mathbb{E}[\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*, \bar{\boldsymbol{\eta}}) \boldsymbol{v}^*] = \mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j} - \mathbf{I}^*_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \boldsymbol{w}^* = 0,$$

we can use the similar proof of Lemma 3.6.3 to show that

$$|I_{11}| \leq \|\boldsymbol{v}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\eta}})\|_\infty \|\hat{\Delta}_{\boldsymbol{w}}\|_1 \lesssim \sqrt{\log d/n} \|\hat{\Delta}_{\boldsymbol{w}}\|_1.$$

For I_{22}, use (3.45) again and we have

$$I_{12} \leq \|\boldsymbol{v}^{*T}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\|_\infty \|\hat{\Delta}_{\boldsymbol{w}}\|_1 \lesssim s_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{\log d/n} \|\hat{\Delta}_{\boldsymbol{w}}\|_1.$$

Combine the bounds for I_{11}, I_{12}, I_2,

$$\hat{\Delta}^T_{\boldsymbol{w}} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}} \lesssim s_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{\log d/n} \|\hat{\Delta}_{\boldsymbol{w}}\|_1.$$

According to Lemma 3.6.7,

$$s_{\boldsymbol{w}}^{-1/2} \|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}}\|_1 \lesssim [\hat{\Delta}^T_{\boldsymbol{w}} \ell_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}}]^{1/2}.$$

Combine the upper and lower bounds of $\hat{\Delta}^T_{\boldsymbol{w}} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}}$ with the fact that $\|\hat{\Delta}_{\boldsymbol{w}}\|_1 \leq 2\|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}}\|_1$,

$$[\hat{\Delta}^T_{\boldsymbol{w}} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}}]^{1/2} \lesssim s_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{s_{\boldsymbol{w}} \log d/n}.$$

Therefore, we have

$$\|\hat{\Delta}_{\boldsymbol{w}}\|_1 \leq 2\|\hat{\Delta}_{\boldsymbol{w},\mathcal{S}_{\boldsymbol{w}}}\|_1 \lesssim s^2_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{\log d/n}.$$

Finally,

$$\begin{aligned}
&(\hat{\boldsymbol{v}} - \boldsymbol{v}^*)^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = \hat{\Delta}^T_{\boldsymbol{w}} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}} \\
=&\hat{\Delta}^T_{\boldsymbol{w}} \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \hat{\Delta}_{\boldsymbol{w}} + \hat{\Delta}^T_{\boldsymbol{w}} [\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})] \hat{\Delta}_{\boldsymbol{w}} \\
\lesssim&(s_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{s_{\boldsymbol{w}} \log d/n})^2 + (s^2_{\boldsymbol{w}} s_{\boldsymbol{\theta}} \sqrt{\log d/n})^2 (s_{\boldsymbol{\theta}} \sqrt{\log d/n}) \\
=&s^4_{\boldsymbol{w}} s^2_{\boldsymbol{\theta}} \log d/n
\end{aligned}$$

by similar proof of (3.46). The conclusion follows by the assumption that $s_{\boldsymbol{w}}$ and $s_{\boldsymbol{\theta}}$ are constants.

□

Lemma 3.6.7. *Denote*

$$\kappa_D(s_{\boldsymbol{w}}) = \min \left\{ \frac{s^{1/2}_{\boldsymbol{w}} [\boldsymbol{w}^T \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}} \ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) \boldsymbol{w}]^{1/2}}{\|\boldsymbol{w}_{\mathcal{S}_{\boldsymbol{w}}}\|_1} : \boldsymbol{w} \in \mathbb{R}^{d-1} \backslash \{0\}, \|\boldsymbol{w}_{\mathcal{S}^c_{\boldsymbol{w}}}\|_1 \leq \xi \|\boldsymbol{w}_{\mathcal{S}_{\boldsymbol{w}}}\|_1 \right\},$$

where ξ is a positive constant. Under Assumptions 8, 9, 10, 11, 12, 13 and 14, $\kappa_D(s_{\boldsymbol{w}}) \geq \kappa/\sqrt{2}$ with probability tending to one.

Proof. Since $\|\boldsymbol{v}_{\mathcal{S}_{\boldsymbol{w}}}\|_1 \leq s_{\boldsymbol{w}}^{1/2}\|\boldsymbol{v}_{\mathcal{S}_{\boldsymbol{w}}}\|_2 \leq s_{\boldsymbol{w}}^{1/2}\|\boldsymbol{v}\|_2$, we have

$$\kappa_D^2(s_{\boldsymbol{w}}) \geq \min\left\{\frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} : \boldsymbol{w}\in\mathbb{R}^{d-1}\backslash\{0\}, \|\boldsymbol{w}_{\mathcal{S}_{\boldsymbol{w}}^c}\|_1 \leq \xi\|\boldsymbol{w}_{\mathcal{S}_{\boldsymbol{w}}}\|_1\right\}.$$

Note that

$$\frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} = \frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} + \frac{\boldsymbol{w}^T[\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}$$

and

$$\boldsymbol{w}^T[\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\boldsymbol{w} = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\sqrt{\log d/n}) = o_{\mathbb{P}}(1)$$

by similar proof as (3.46). Therefore, with probability tending to one,

$$\begin{aligned}
&\frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} \geq \frac{3}{4}\frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}\\
=&\frac{3}{4}\frac{\boldsymbol{w}^T\mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} + \frac{3}{4}\frac{\boldsymbol{w}^T[\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}]\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}\\
\geq&\frac{3}{4}\lambda_{\min}(\mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}) - \frac{3}{4}\left|\frac{\boldsymbol{w}^T[\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}]\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}\right|\\
\geq&\frac{3}{4}\left[\kappa^2 - \frac{\|\boldsymbol{w}\|_1^2\|\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\|_\infty}{\|\boldsymbol{w}\|_2^2}\right]
\end{aligned}$$

In addition, from

$$\|\boldsymbol{w}\|_1^2 \leq (\xi+1)^2\|\boldsymbol{w}_{\mathcal{S}_{\boldsymbol{w}}}\|_1^2 \leq s_{\boldsymbol{w}}(\xi+1)^2\|\boldsymbol{w}\|_2^2$$

we get

$$\frac{\boldsymbol{w}^T\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} \geq \frac{3}{4}\left[\kappa^2 - s_{\boldsymbol{w}}(\xi+1)^2\|\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\|_\infty\right].$$

Similar to the proof of Lemma 3.6.3, we can obtain

$$\|\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n}).$$

Hence we have $\|\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\|_\infty = o_{\mathbb{P}}(1)$ by Assumption 12. When n is large enough,

$$\|\nabla^2_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\|_\infty \leq \kappa^2/[3(\xi+1)^2].$$

Therefore, $\kappa_D(s_{\boldsymbol{w}}) \geq \kappa/\sqrt{2}$ with probability tending to one. □

Lemma 3.6.8 (Local smoothness conditions on the loss function)**.** *Under Assumptions , we have*

$$\boldsymbol{v}^{*T}[\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)] = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n), \tag{3.49}$$

$$(\hat{\boldsymbol{v}}-\boldsymbol{v}^*)^T[\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})] = \mathcal{O}_{\mathbb{P}}(s_{\boldsymbol{\theta}}\log d/n). \tag{3.50}$$

The same results hold for $\hat{\boldsymbol{\theta}}_0 = (0, \hat{\boldsymbol{\theta}}_{-j}^T)^T$, where θ_j is the parameter we are interested in.

Proof. Using the similar proof as (3.46),

$$\begin{aligned}
&\left|\boldsymbol{v}^{*T}[\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)]\right| \\
=&\left|\boldsymbol{v}^{*T}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}} - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}]\right| \\
\lesssim&\|\boldsymbol{v}^*\|_1\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 \lesssim s^2_{\boldsymbol{\theta}}\log d/n = o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

for some $q \in [0,1]$ by Assumption 12. Let $\hat{\Delta}_{\boldsymbol{v}} := \hat{\boldsymbol{v}} - \boldsymbol{v}^*$. For (3.50), we have

$$\begin{aligned}
&\left|(\hat{\boldsymbol{v}}-\boldsymbol{v}^*)^T[\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\right| \\
=&\left|\hat{\Delta}_{\boldsymbol{v}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right| \\
\leq&\left|\hat{\Delta}_{\boldsymbol{v}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right| + \left|\hat{\Delta}_{\boldsymbol{v}}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\hat{\Delta}_{\boldsymbol{\theta}}\right|
\end{aligned}$$

by Taylor's expansion. For the first term on the right-hand side of the inequality, the Cauchy-Schwartz inequality yields

$$\begin{aligned}
&\left|\hat{\Delta}_{\boldsymbol{v}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right| \\
\leq&\left|\hat{\Delta}_{\boldsymbol{v}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{v}}\right|^{1/2}\left|\hat{\Delta}_{\boldsymbol{\theta}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})\hat{\Delta}_{\boldsymbol{\theta}}\right|^{1/2} \\
\lesssim&\sqrt{s_{\boldsymbol{\theta}}\log d/n}\sqrt{(s_{\boldsymbol{\theta}}\vee s_{\boldsymbol{w}})\log d/n} = \mathcal{O}_{\mathbb{P}}((s_{\boldsymbol{\theta}}\vee s_{\boldsymbol{w}})\log d/n) = o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

by (3.41) and (3.48). For the second term on the right-hand side, similar proof as (3.46) yields

$$\begin{aligned}
&\left|\hat{\Delta}_{\boldsymbol{v}}[\nabla^2_{\boldsymbol{\theta\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})-\nabla^2_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\hat{\Delta}_{\boldsymbol{\theta}}\right| \\
\lesssim&\|\hat{\Delta}_{\boldsymbol{v}}\|_1\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1\|\hat{\Delta}_{\boldsymbol{\theta}}\|_1 \lesssim ((s_{\boldsymbol{\theta}}\vee s_{\boldsymbol{w}})\sqrt{\log d/n})^3 \\
=&o_{\mathbb{P}}(1/\sqrt{n})\mathcal{O}_{\mathbb{P}}((s_{\boldsymbol{\theta}}\vee s_{\boldsymbol{w}})\sqrt{\log d/n}) = o_{\mathbb{P}}(1/\sqrt{n}).
\end{aligned}$$

□

Finally, we give the proof of Theorem 3.3.1.

Proof. For the assumptions needed in Ning and Liu (2017, Theorem 3.2), Lemma 3.6.5 and 3.6.6 satisfy Assumption 1, Lemma 3.6.3 satisfies Assumption 2, Lemma 3.6.8 satisfies Assumption 3, and Lemma 3.6.4 satisfies Assumption 4. Now we only need to verify that $\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}} - I^*_{\theta_j|\boldsymbol{\theta}_{-j}} = o_{\mathbb{P}}(1)$.

First note that

$$\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}} = \nabla^2_{\theta_j\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{w}}^T\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})$$

and

$$I^*_{\theta_j|\boldsymbol{\theta}_{-j}} = I^*_{\theta_j\theta_j} - \boldsymbol{w}^{*T}\mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j}, \qquad \text{where } \boldsymbol{w}^* = \mathbf{I}^{*-1}_{\boldsymbol{\theta}_{-j}\boldsymbol{\theta}_{-j}}\mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j}.$$

For the second part of $\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}}$ and $I^*_{\theta_j|\boldsymbol{\theta}_{-j}}$,

$$\begin{aligned}
&\hat{\boldsymbol{w}}^T\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \boldsymbol{w}^{*T}\mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j} \\
\le&\left|(\hat{\boldsymbol{w}}-\boldsymbol{w}^*)^T\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\right| + \left|\boldsymbol{w}^{*T}[\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})]\right| \\
&+\left|\boldsymbol{w}^{*T}[\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}}) - \mathbf{I}^*_{\boldsymbol{\theta}_{-j}\theta_j}]\right| \\
=:&J_1+J_2+J_3.
\end{aligned}$$

It can be shown that

$$J_1 \le \|\hat{\boldsymbol{w}}-\boldsymbol{w}^*\|_1\|\nabla^2_{\boldsymbol{\theta}_{-j}\theta_j}\ell(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\|_\infty \lesssim (s_{\boldsymbol{\theta}}\vee s_{\boldsymbol{w}})\sqrt{\log d/n}$$

since the variables are bounded. By Taylor's expansion,

$$J_2 = \left|\boldsymbol{w}^{*T}\nabla^3_{\boldsymbol{\theta}_{-j}\theta_j\boldsymbol{\theta}}\ell(q\hat{\boldsymbol{\theta}}+(1-q)\boldsymbol{\theta}^*,\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right| \lesssim \|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| \lesssim s_{\boldsymbol{\theta}}\sqrt{\log d/n}$$

for some $q \in [0,1]$. Using similar arguments as the proof of (3.37), we have that

$$J_3 \lesssim \sqrt{\log d/n}.$$

In addition, similar arguments also imply that

$$|\nabla^2_{\theta_j\theta_j}\ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - I^*_{\theta_j\theta_j}| \lesssim \sqrt{\log d/n}.$$

Therefore, $\hat{I}_{\theta_j|\boldsymbol{\theta}_{-j}} - I^*_{\theta_j|\boldsymbol{\theta}_{-j}} = o_{\mathbb{P}}(1)$ and thus $\sqrt{n}(\tilde{\theta}_j - \theta^*_j)I^*_{\theta_j|\boldsymbol{\theta}_{-j}} \Rightarrow N(0, \sigma^*_S)$. □

CHAPTER 4

Fusing Individualized Treatment Rules Using Auxiliary Outcomes

4.1 Introduction

An individualized treatment rule (ITR) is the decision rule that recommends a treatment to a patient based on his or her pre-treatment covariates such as demographics, medical history and genotypes. When the primary outcome is of interest, the goal of ITR learning is to estimate the optimal ITR that yields the maximal outcome if patients follow the treatment recommendations. During the past decade, many methods have been proposed to estimate the optimal ITR using data from a randomized controlled trial (RCT) or an observational study (Qian and Murphy, 2011; Zhao et al., 2012; Liu et al., 2018b; Zhao et al., 2019). All these methods consider only a single primary outcome when estimating ITRs.

In practice, when making treatment decisions additional auxiliary outcomes also need to be considered, since they are often affected by treatments and non-favorable auxiliary outcomes can potentially represent worsened overall health or increase the chance of non-compliance. Hence, when estimating the optimal ITRs for the primary outcome of interest, the derived ITRs should also optimize auxiliary outcomes to the best extent and do not incur harm. For example, when treating patients with major depressive disorder (MDD; Trivedi et al., 2016), one common outcome to measure depressive symptoms is the Quick Inventory of Depressive Symptomatology (QIDS) score, which is a rating system based on the patient's feelings in the past 7 days. In addition, another important outcome is the Clinical Global Improvement (CGI) Scale, which is often used to assess a patient's symptoms, behavior, the impact on the patient's ability to function, and is an indicator of the overall clinical improvement. Although the primary goal is to find the best treatment strategy to improve the QIDS score, it is important for such strategy to be also effective in terms of CGI scale.

Several approaches are proposed to learn ITRs that maximize multiple outcomes simultaneously. Wang et al. (2018) and Laber et al. (2018) estimated the optimal ITRs for the primary outcome while restricting the auxiliary outcome (or risk outcome) to be no larger than a threshold. These approaches require pre-specification of the threshold value. Moreover, they only guarantee the average risk to be small, but for a given individual, the treatment decision can be different from the one yielding the minimal risk. Luckett et al. (2021) proposed to construct patient-specific composite outcomes for learning ITRs, and the composition weights were obtained using observed clinical decisions in practice. However, the estimated ITRs are no longer the optimal treatment decisions for the primary outcome of interest.

Our goal is to estimate the optimal ITRs for the primary outcome and at the same time, ensuring the derived treatments are consistent with the optimal rules for other auxiliary outcomes as much as possible. In other words, we aim to fuse the treatment rules for these different outcomes to obtain a desirable fused ITR that performs optimally for the primary outcome and effectively for non-primary outcomes, although not necessarily the best. We emphasize that this goal is fundamentally different from all existing works for combining multiple studies in analysis, such as meta-analysis (Haidich, 2010; Lin and Zeng, 2010; Claggett et al., 2014; Liu et al., 2015), where the estimators for common parameters from multiple studies are combined into a statistically efficient estimator, or integrative data analysis (Curran and Hussong, 2009; Brown et al., 2018), where multiple data sources or summary statistics are analyzed together through some shared parameter models, or transferring learning (Li et al., 2021; Tian and Feng, 2022; Cai and Wei, 2021) in which one uses existing knowledge from another task to assist the learning of a new task. All existing methods either require individual-level data or assumes models or distributions for each data to achieve integration, so are not applicable to combine different treatment rules.

In order to integrate treatment rules, we propose a fused learning framework to estimate the optimal ITR, which we name as the fused individualized treatment rule (FITR). Specifically, we maximize the value function for the primary outcome, but at the same time, we introduce a fusion penalty to encourage the similarity between the estimated ITR and the optimal ITRs for the auxiliary outcomes. The latter are assumed to be estimated apriori using either external data or the same study. The fusion penalty is chosen to be a weighted sum of the disagreement rates between the treatment rules, where the weights depend on the similarity of treatment response for

these outcomes such as their correlations. The fusion penalty encourages the ITRs for the different outcomes to be as consistent as possible. Computationally, we propose a ramp loss to approximate the fusion penalty and use a surrogate loss, e.g., logistic loss, to substitute the value function in the objective function. We obtain the convergence rate for the value function of FITR. Furthermore,we prove theoretically that the agreement rate, which is between the estimated FITR and the auxiliary outcome treatment rule, will converge to the agreement rate between their corresponding optimal ITRs. More importantly, the convergence rate is faster than the one without using the fusion penalty.

The rest of this chapter is organized as follows. In Section 4.2, we introduce basic assumptions and the proposed method, FITR. We also propose optimization algorithms to solve for the FITR. In Section 4.3, we derive the convergence rates of FITR in terms of the value functions, treatment selection accuracy and agreement rates. In Section 4.4, we demonstrate the performance of our method using simulation studies. In addition, we analyze how close the true optimal ITRs for different outcomes should be in order for FITRs to have higher value functions. Finally, in Section 4.5, we illustrate the proposed method through analysis of a clinical trial for MDD patients (Trivedi et al., 2016).

4.2 Methodology

Let R_1 denote the primary outcome. Without loss of generality, assume a higher outcome indicates a better health condition. We use a vector $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ to denote pre-treatment covariates for tailoring treatment decisions, where d is the dimension of $\boldsymbol{X}$, and assume that treatment $A \in \mathcal{A} = \{1, -1\}$ is binary. For the primary outcome R_1, an ITR $\mathcal{D}_1 : \mathcal{X} \to \mathcal{A}$ maps the covariate space of a patient to a treatment. Equivalently, we can express $\mathcal{D}_1(\boldsymbol{X}) = \text{sign}\{f_1(\boldsymbol{X})\}$ for some decision function f_1. Letting $\mathcal{V}_1(f_1) = \mathbb{E}^{\mathcal{D}_1}(R_1)$ be the value function associated with $\mathcal{D}_1$, our goal is to estimate the optimal ITR that maximizes this value function.

Let $\mathbb{P}$ be the joint distribution of $\boldsymbol{Z} := (\boldsymbol{X}, A, R_1)$ and $\mathbb{E}$ be the corresponding expectation for $k = 1, \ldots, K$. If the treatment is assigned according to some ITR $\mathcal{D}_1$ with $A = \mathcal{D}_1(\boldsymbol{X})$, the distribution and expectation are denoted as $\mathbb{P}^{\mathcal{D}_1}$ and $\mathbb{E}^{\mathcal{D}_1}$, respectively. Then the optimal ITR $\mathcal{D}_1^*$ is given as $\text{sign}\{f_1^*(\boldsymbol{x})\}$, where $f_1^*(\boldsymbol{x}) = \mathbb{E}[R_1(1)|\boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[R_1(-1)|\boldsymbol{X} = \boldsymbol{x}]$ and $R_1(a)$ denotes the

potential outcome for treatment a. We assume the following conditions so that the optimal ITR is estimable using data n i.i.d copies of $(\boldsymbol{X}, A, R_1)$.

Assumption 15 (Ignorability)**.** The treatment A is independent of the potential outcomes $R_1^*(a)$ given covariates $\boldsymbol{X}$.

Assumption 16 (Consistency)**.** The observed outcome R_1 under a treatment $A = a$ equals the potential outcome $R_1(a)$ for all $a \in \mathcal{A}$.

Assumption 17 (Positivity)**.** There exists $p_0 > 0$ such that $\pi(a; \boldsymbol{x}) \equiv P(A = a|\boldsymbol{X} = \boldsymbol{x}) > p_0$ for all $a \in \mathcal{A}$ and all $\boldsymbol{x} \in \mathcal{X}$.

Under these conditions, according to Qian and Murphy (2011), $\mathcal{V}_1(f_1) = \mathbb{E}[R_1 I(Af_1(\boldsymbol{X}) > 0)/\pi(A; \boldsymbol{X})]$. Thus, the optimal ITR can be estimated by solving

$$\min_{f_1 \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{R_{i1} I(A_i f_1(\boldsymbol{X}_i) < 0)}{\pi(A_i; \boldsymbol{X}_i)} + \lambda_{1n} \|f_1\|^2 , \tag{4.1}$$

where $\mathcal{F}$ is some function class, $\|f_1\|$ is the semi-norm for f_1 in its function space, and λ_{1n} is a tuning parameter depending on the sample size n. Since the 0-1 loss is computationally challenging, it can be substituted by some convex surrogate loss (Bartlett et al., 2006), denoted by $\phi(x)$, and we can solve a convex optimization problem:

$$\widetilde{f}_{1n} = \arg\min_{f_1 \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{R_{i1}}{\pi(A_i; \boldsymbol{X}_i)} \phi(A_i f_1(\boldsymbol{X}_i)) + \lambda_{1n} \|f_1\|^2 . \tag{4.2}$$

For example, Zhao et al. (2012) proposed to use the hinge loss where $\phi(x) = \max(1 - x, 0)$. In our following implementation, we propose to use the logistic loss, $\phi(t) = \log(1 + e^{-t})$, due to its differentiability property. We refer to this method as separate learning (*SepL*).

4.2.1 Learning Fused ITR using Optimal Rules for Auxiliary Outcomes

Suppose that for auxiliary outcomes $R_2, \ldots, R_K$, corresponding ITRs denoted as $\widetilde{f}_2, \ldots, \widetilde{f}_K$ have been obtained from external data or the same study. Our goal is to estimate the optimal ITR for the primary outcome but encourage it to be consistent to these auxiliary ITRs as much as possible. To this end, we propose a *fusion penalty* on the disagreement rates between f_1 and

$\widetilde{f}_2, \ldots, \widetilde{f}_K$. Specifically, the *fused individualized treatment rule (FITR)* f_1 is estimated by

$$\begin{aligned}\widehat{f}_{1n} = \arg\min_{f_1 \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{R_{i1}}{\pi(A_i; \boldsymbol{X}_i)} I(A_i f_1(\boldsymbol{X}_i) < 0) + \lambda_{1n} \|f_1\|^2 \\ + \frac{\mu_{1n}}{n} \sum_{i=1}^{n} \sum_{j=2}^{K} \Omega_{1j} I(f_1(\boldsymbol{X}_i) \widetilde{f}_j(\boldsymbol{X}_i) < 0),\end{aligned} \tag{4.3}$$

where Ω_{1j} is a pre-specified constant to reflect the similarity between ITRs $\widetilde{f}_{1n}$ and $\widetilde{f}_j$ for $j \geq 2$. For example, Ω_{1j} can be defined as the correlation of R_1 and R_j if the treatment decisions are expected to be similar between two highly-correlated outcomes. In the objective function (4.3), μ_{1n} is a tuning parameter to be selected data-adaptively.

The optimization in (4.3) is a NP-hard problem, so we substitute the 0-1 losses with another smooth loss for optimization. First, the 0-1 loss in the first part of the expression is substituted by a logistic loss as in SepL. For the 0-1 loss in the fusion penalty, we use a ramp loss $\psi_\kappa(t) = \min\{1, \max\{0, 1 - t/\kappa\}\}$, where κ is a tuning parameter, for approximation since the latter converges to the 0-1 loss when κ decreases to zero. As a result, we solve the following problem to estimate f_1. where we allow κ_{1n} to depend on the sample size n. To further reduce the variability, we can replace R_{i1} by $R_{i1} - \mathbb{E}(R_{i1}|\boldsymbol{X}_i)$. However, we assume that all rewards are nonnegative in Section 4.3. To deal with negative rewards, we can take the absolute value of $R_{i1} - \mathbb{E}(R_{i1}|\boldsymbol{X}_i)$ and flip the sign of A (c.f., (Liu et al., 2018b)). Therefore, FITR-Ramp finds $\widehat{f}_{1n}$ by minimizing

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} \frac{|R_{i1} - \mathbb{E}(R_{i1}|\boldsymbol{X}_i)|}{\pi(A_i; \boldsymbol{X}_i)} \phi(A_i \operatorname{sign}\{R_{i1} - \mathbb{E}(R_{i1}|\boldsymbol{X}_i)\} f_1(\boldsymbol{X}_i)) \\ + \lambda_{1n} \|f_1\|^2 + \frac{\mu_{1n}}{n} \sum_{i=1}^{n} \sum_{j=2}^{K} \Omega_{1j} \psi_{\kappa_{1n}}(f_1(\boldsymbol{X}_i) \widetilde{f}_j(\boldsymbol{X}_i))\end{aligned}$$

with a shift of a constant. Here $\mathbb{E}(R_{i1}|\boldsymbol{X}_i)$ can be estimated by a simple linear regression. We call this optimization problem *FITR-Ramp* since it substitutes the fusion penalty with the ramp loss.

Alternatively, we can solve (4.3) using the following procedure. First note that for $j \geq 2$,

$$\begin{aligned}
&I(f_1(\boldsymbol{X}_i)\widetilde{f}_j(\boldsymbol{X}_i) < 0) \\
=&I(A_i f_1(\boldsymbol{X}_i) < 0)I(A_i\widetilde{f}_j(\boldsymbol{X}_i) > 0) + (1 - I(A_i f_1(\boldsymbol{X}_i) < 0))I(A_i\widetilde{f}_j(\boldsymbol{X}_i) < 0) \\
=&I(A_i f_1(\boldsymbol{X}_i) < 0)\,\mathrm{sign}\{A_i\widetilde{f}_j(\boldsymbol{X}_i)\} + \frac{1 - \mathrm{sign}\{A_i\widetilde{f}_j(\boldsymbol{X}_i)\}}{2}.
\end{aligned}$$

Therefore, the problem in (4.3) is equivalent to

$$\widehat{f}_{1n} = \arg\min_{f_1 \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{R}_{i1}}{\pi(A_i; \boldsymbol{X}_i)} I(A_i f_1(\boldsymbol{X}_i) < 0) + \lambda_{1n} \|f_1\|^2,$$

where $\widetilde{R}_{i1} = R_{i1} + \mu_{1n}\pi(A_i; \boldsymbol{X}_i)\sum_{j=2}^{K} \Omega_{1j}\,\mathrm{sign}\{A_i\widetilde{f}_j(\boldsymbol{X}_i)\}$ is the pseudo outcome. After substitute the indicator function by the logistic loss to obtain

$$\widehat{f}_{1n} = \arg\min_{f_1 \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{R}_{i1}}{\pi(A_i; \boldsymbol{X}_i)} \phi(A_i f_1(\boldsymbol{X}_i)) + \lambda_{1n} \|f_1\|^2. \tag{4.4}$$

Similarly, we can replace $\widetilde{R}_{i1}$ by $\widetilde{R}_{i1} - \mathbb{E}(\widetilde{R}_{i1}|\boldsymbol{X}_i)$ and estimate $\widehat{f}_{1n}$ by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\left|\widetilde{R}_{i1} - \mathbb{E}(\widetilde{R}_{i1}|\boldsymbol{X}_i)\right|}{\pi(A_i; \boldsymbol{X}_i)} \phi(A_i\,\mathrm{sign}\{\widetilde{R}_{i1} - \mathbb{E}(\widetilde{R}_{i1}|\boldsymbol{X}_i)\} f_1(\boldsymbol{X}_i)) + \lambda_{1n} \|f_1\|^2.$$

We call this optimization method *FITR-IntL*. As a note, similar procedure was originally proposed in Qiu et al. (ress) to integrate treatment rules from multiple studies.

In both procedures, the semi-norm for f_1 is usually chosen as the one from a reproducing kernel Hilbert space (RKHS) associated with a real valued kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The choice of k can be $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$, which yields a linear decision function for f_1, or the Gaussian kernel, $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma^2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$, where σ is a parameter that can depend on n, to give a nonlinear decision function. By the representer theorem, the minimizer for f_1 takes the form $f(\boldsymbol{X}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{X}, \boldsymbol{X}_i)$ so solving for *FITR-Ramp* or *FITR-IntL* can be restricted to class

$$\mathcal{H} := \left\{ f : f(\boldsymbol{X}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{X}, \boldsymbol{X}_i), (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\}.$$

The tuning parameters λ_{1n} and μ_{1n} can be selected by cross-validation. For example, we can first select λ_{1n} in SepL and then select μ_{1n} for FITR-Ramp or FITR-IntL with λ_{1n} fixed. Computationally, FITR-Ramp can be solved by the difference of convex functions algorithm (DCA) (Le Thi and Pham Dinh, 2018) or the Powell algorithm (Powell, 1964; Press et al., 2007). While DCA re-expresses the objective function as the difference of two convex functions and obtains solution iteratively, the Powell algorithm is suitable for non-differentiable objective functions. In contrast, FITR-IntL has a differentiable convex objective function so it can be easily solved by gradient-based algorithms like variations of the gradient descent algorithm (Ruder, 2016) or the BFGS algorithm (Fletcher, 1987).

4.3 Theoretical Results

We first study the nonasymptotic properties of FITR-Ramp for $K = 2$. That is, there exits only one auxiliary treatment rule $\widetilde{f}_2$, which is already estimated either from an external dataset or from the same dataset. Without loss of generality, we assume that $\widetilde{f}_2 : \mathcal{X} \to \{1, -1\}$ is binary, since every decision function f can be transformed into a binary function by taking its sign. For the first outcome, we define $\ell_1 \circ f_1(\boldsymbol{Z}) := \frac{R_1}{\pi(A|\boldsymbol{X})}\phi(Af_1(\boldsymbol{X}))$ and $\ell_2 \circ f_1(\boldsymbol{Z}) := \mu_{1n}\Omega_{12}\psi_{\kappa_{1n}}[f_1(\boldsymbol{X})\widetilde{f}_2(\boldsymbol{X})]$. Let the risk based on the surrogate losses to be

$$\mathcal{R}(f_1) := \mathbb{E}\left[\ell_1 \circ f_1 + \ell_2 \circ f_1\right].$$

Assumption 18. Suppose that $0 \leq R_1 \leq r$ for all R_1 and for some constant $r > 0$.

To give additional assumptions, we define

$$\eta(\boldsymbol{X}) := \begin{cases} \frac{\mathbb{E}(R_1|1,\boldsymbol{X})}{\mathbb{E}(R_1|1,\boldsymbol{X})+\mathbb{E}(R_1|-1,\boldsymbol{X})}, & \text{if } \mathbb{E}(R_1|1,\boldsymbol{X}) \neq \mathbb{E}(R_1|-1,\boldsymbol{X}) \\ \frac{1}{2}, & \text{otherwise} \end{cases}$$

and define the classes of a probability measure $\mathbb{P}$ by $\mathcal{X}_{-1} := \left\{\boldsymbol{x} \in \mathcal{X} : \eta(\boldsymbol{x}) < \frac{1}{2}\right\}$, $\mathcal{X}_1 := \left\{\boldsymbol{x} \in \mathcal{X} : \eta(\boldsymbol{x}) > \frac{1}{2}\right\}$ and $\mathcal{X}_0 := \left\{\boldsymbol{x} \in \mathcal{X} : \eta(\boldsymbol{x}) = \frac{1}{2}\right\}$ for some choice of η. Finally, we define a

distance function $\boldsymbol{x} \mapsto \omega_{\boldsymbol{x}}$ as

$$\omega_{\boldsymbol{x}} := \begin{cases} d(\boldsymbol{x}, \mathcal{X}_0 \cup \mathcal{X}_1), & \text{if } \boldsymbol{x} \in \mathcal{X}_{-1}, \\ d(\boldsymbol{x}, \mathcal{X}_0 \cup \mathcal{X}_{-1}), & \text{if } \boldsymbol{x} \in \mathcal{X}_1, \\ 0, & \text{otherwise,} \end{cases}$$

where $d(\boldsymbol{x}, \mathcal{S})$ denotes the distance of $\boldsymbol{x}$ to a set $\mathcal{S}$ with respect to the Euclidean norm.

Assumption 19. Assume that the distribution of $\boldsymbol{X}$ satisfies the following Tsybakov's noise assumption: there exists a constant $C > 0$ such that for all sufficiently small $t > 0$ we have

$$\mathbb{P}(\{\boldsymbol{X} \in \mathcal{X} : |2\eta(\boldsymbol{X}) - 1| \leq t\}) \leq Ct^{\beta}$$

for some $\beta > 0$. Let $\alpha = \beta/(1+\beta)$ so that $\alpha \in (0, 1]$.

Assumption 20. Assume that there exists a constant $q \in (0, \infty]$ for the distribution of $\boldsymbol{X}$ such that

$$\int_{\mathcal{X}} \exp\left(-\frac{\omega_{\boldsymbol{x}}^2}{t}\right) \mathbb{P}_{\mathcal{X}}(d\boldsymbol{x}) \leq Ct^{qd/2}, \qquad t > 0, \tag{4.5}$$

for some constant $C > 0$.

Note that Assumption 20 is used to bound the approximation error when we find the estimated ITR in a specific function class. It is different from the geometric noise assumption (Steinwart and Scovel, 2007) in the sense that the term $|2\eta(\boldsymbol{x}) - 1|$ is not included in left-hand side of (4.5). Since $|2\eta(\boldsymbol{x}) - 1| \leq 1$, Assumption 20 is stronger than the geometric noise assumption. However, when $t \to 0$, we can still ensure that the left-hand side of (4.5) goes to zero. For Assumption 19, it has been shown that the boundary assumption of $\eta(\boldsymbol{x})$ regarding β is equivalent to the misclassification assumption of f_1 regarding α (Bartlett et al., 2006). This assumption is needed when bounding the risk with 0-1 loss using the risk with logistic loss. In the following inequalities, "$\lesssim$" indicates that the left-hand side is no larger than the right-hand side for all n up to a universal constant.

Lemma 4.3.1. *Under Assumptions 15-17, 18 and 20, when* $\mu_{1n}, \kappa_{1n}, \sigma_{1n}^2 \to 0$, $\mu_{1n}/\kappa_{1n} \to \infty$, *and* $\kappa_n \leq 1$ *for all* n, *for any* $\delta > 0$, $0 < \nu < 2$ *and for all* $\tau \geq 1$, *we have* $\mathbb{P}(\mathcal{R}(\widehat{f}_{1n}) - \mathcal{R}(f_1^*) \lesssim$

$\delta_{1n}(\tau)) \geq 1 - e^{-\tau}$, *where*

$$\delta_{1n}(\tau) := \lambda_{1n}^{-\frac{1}{2}} n^{-\frac{1}{2}} \left[\sqrt{\tau} + \sigma_{1n}^{(1-\nu/2)(1+\delta)d} \gamma_n^{-\nu}\right] + \lambda_{1n}\sigma_{1n}^{d} + \left(1 + \frac{\mu_{1n}}{\kappa_{1n}}\right)(2d)^{qd/2}\sigma_{1n}^{-qd}, \tag{4.6}$$

and $\gamma_n := [1 + \mu_{1n}\kappa_{1n}^{-1}]^{-1}$.

We assume $\mu_{1n}/\kappa_{1n} \to \infty$ to ensure that μ_{1n} does not decrease too quickly and guarantee the effect of the fusion penalty on the agreement rate as in Zou (2006). The first term on the right-hand side of (4.6) is the estimation error and the sum of the last two terms is the approximation error. A larger class $\mathcal{H}$ (with small penalty parameters λ_{1n} and μ_{1n}) generally leads to a larger estimation error and a smaller approximation error. The parameters $\lambda_{1n}, \mu_{1n}, \kappa_{1n}$ should balance them to achieve the minimal upper bound $\delta_{1n}(\tau)$. The minimal approximation error is $(\mu_{1n}/\kappa_{1n})^{\frac{1}{1+q}} \lambda_{1n}^{\frac{q}{1+q}}$ when $\sigma_{1n} = (\mu_{1n}/(\kappa_{1n}\lambda_{1n}))^{\frac{1}{(1+q)d}}$. Then the best λ_{1n} that balances the estimation error and the approximation error is

$$\lambda_{1n} = \left(\frac{\mu_{1n}}{\kappa_{1n}}\right)^{\frac{2(\Delta-1)+2(1+q)v}{3q+2\Delta+1}} n^{-\frac{1+q}{3q+2\Delta+1}},$$

and the corresponding convergence rate is

$$\delta_{1n}(\tau) = \left(\frac{\mu_{1n}}{\kappa_{1n}}\right)^{\frac{2qv+2\Delta+1}{3q+2\Delta+1}} n^{-\frac{q}{3q+2\Delta+1}} \tag{4.7}$$

where $\Delta = (1 - \nu/2)(1 + \delta)$.

Note that when $\mu_{1n} = 0$, FITR degenerates to SepL learnt with the logistic loss. In this case,

$$\delta_{1n}^{(0)}(\tau) := \lambda_{1n}^{-\frac{1}{2}} n^{-\frac{1}{2}} \left[\sqrt{\tau} + \sigma_{1n}^{(1-\nu/2)(1+\delta)d}\right] + \lambda_{1n}\sigma_{1n}^{d} + (2d)^{qd/2}\sigma_{1n}^{-qd}.$$

The minimum approximation error $\lambda_{1n}^{\frac{q}{1+q}}$ is obtained when $\sigma_{1n} = \lambda_{1n}^{-\frac{1}{(1+q)d}}$. Then the best λ_{1n} that balances the estimation error and the approximation error is $\lambda_{1n} = n^{-\frac{1+q}{3q+2\Delta+1}}$ and the corresponding convergence rate is

$$\delta_{1n}^{(0)}(\tau) = n^{-\frac{q}{3q+2\Delta+1}}. \tag{4.8}$$

To quantify the agreement rate between $\widetilde{f}_2$ and its corresponding optimal ITR f_2^*, we further assume that $\widetilde{f}_2$ is a binary mapping learnt from a dataset of size N with certain convergence rate.

Assumption 21. Assume that the estimator of the secondary ITR $\widetilde{f}_2$ converges to f_2^* in the sense that $\mathbb{P}(\widetilde{f}_2 f_2^* < 0) \leq \widetilde{\delta}_{2N}(\tau)$ with probability greater than or equal to $1 - e^{\tau}$, where N is the sample size of the dataset and $\widetilde{\delta}_{2N}(\tau) = o(1)$ as $N \to \infty$.

Now we are able to present the convergence rates of the value function $\mathcal{V}_1(\widehat{f}_{1n})$ and the agreement rate $\mathbb{P}(\widehat{f}_{1n} f_2^* > 0)$.

Theorem 4.3.2. *Under Assumptions 15-21, the value function of the estimated FITR satisfies*

$$\mathcal{V}_1(f_1^*) - \mathcal{V}_1(\widehat{f}_{1n}) \lesssim (\delta_{1n}(\tau) + \mu_{1n})^{\frac{1}{2-\alpha}}, \tag{4.9}$$

with probability greater than or equal to $1 - 2e^{\tau}$.

Notice that the convergence rate $\widetilde{\delta}_{2N}(\tau)$ of $\widetilde{f}_2$ does not appear in (4.9), since the term containing $\widetilde{\delta}_{2N}(\tau)$ is not dominant in the proof due to the assumption that $\widetilde{\delta}_{2N}(\tau) = o(1)$. By comparing $\delta_{1n}(\tau) + \mu_{1n}$ and $\delta_{1n}^{(0)}(\tau)$, it is clear that FITR has a slower convergence rate for the value function than SepL. This is due to the fact that the fusion penalty introduces bias for maximizing the primary outcome.

However, as discussed before, one major advantage of FITR is to improve the approximity between the estimated ITR and the optimal ITR, $\widetilde{f}_2$, for the auxiliary outcome.

Theorem 4.3.3. *Under Assumptions 15-21, the agreement rate between* $\widehat{f}_{1n}$ *and* f_2^* *satisfies*

$$\mathbb{P}(f_1^* f_2^* > 0) - \mathbb{P}(\widehat{f}_{1n} f_2^* > 0) \lesssim \frac{\delta_{1n}(\tau)}{\mu_{1n}} + \widetilde{\delta}_{2N}(\tau) \tag{4.10}$$

with probability greater than or equal to $1 - 2e^{\tau}$.

To compare the agreement rate with or without the fusion penalty, first note that

$$\begin{aligned}\mathbb{P}(\widehat{f}_{1n} f_2^* < 0) =& \mathbb{P}(\widehat{f}_{1n} f_1^* < 0, f_1^* f_2^* > 0) + \mathbb{P}(\widehat{f}_{1n} f_1^* > 0, f_1^* f_2^* < 0) \\ \leq& \mathbb{P}(\widehat{f}_{1n} f_1^* < 0) + \mathbb{P}(f_1^* f_2^* < 0),\end{aligned}$$

which bound the agreement rate with the decision accuracy. This can be used for the bound without the fusion penalty. In Corollary 4.7.1, we present the convergence rate of the decision accuracy with an additional assumption. For SepL with $\mu_{1n} = 0$, $\mathbb{P}(\widehat{f}_{1n} f_1^* < 0) \lesssim n^{-\frac{\alpha}{2-\alpha}\frac{q}{3q+2\Delta+1}}$ with probability

greater than or equal to $1-2e^{\tau}$ by (4.8) when $\sigma_{1n}=\lambda_{1n}^{-\frac{1}{(1+q)d}}$ and $\lambda_{1n}=n^{-\frac{1+q}{3q+2\Delta+1}}$. Then we can conclude that for SepL

$$\mathbb{P}(f_1^*f_2^*>0)-\mathbb{P}(\widehat{f}_{1n}f_2^*>0)\leq\mathbb{P}(\widehat{f}_{1n}f_1^*<0)\lesssim n^{-\frac{\alpha}{2-\alpha}\frac{q}{3q+2\Delta+1}}. \tag{4.11}$$

On the other hand, for FITR-Ramp, (4.10) and (4.7) implies that

$$\mathbb{P}(f_1^*f_2^*>0)-\mathbb{P}(\widehat{f}_{1n}f_2^*>0)\lesssim\frac{1}{\mu_{1n}}\left(\frac{\mu_{1n}}{\kappa_{1n}}\right)^{\frac{2qv+2\Delta+1}{3q+2\Delta+1}}n^{-\frac{q}{3q+2\Delta+1}}+\widetilde{\delta}_{2N}(\tau) \tag{4.12}$$

when $\sigma_{1n}=(\mu_{1n}/(\kappa_{1n}\lambda_{1n}))^{\frac{1}{(1+q)d}}$ and $\lambda_{1n}=\left(\frac{\mu_{1n}}{\kappa_{1n}}\right)^{\frac{2(\Delta-1)+2(1+q)v}{3q+2\Delta+1}}n^{-\frac{1+q}{3q+2\Delta+1}}$. When the data are fully separated by the decision boundary and $\alpha=1$, the right-hand side of (4.11) $n^{-\frac{q}{3q+2\Delta+1}}$ is smaller than that of (4.12). In this case, SepL can learn the disagreement rate quick enough and adding the fusion penalty will reduce the convergence speed. However, the right-hand side of (4.12) is smaller than that of (4.11) when

$$\frac{\mu_{1n}^{2qv-3q}}{\kappa_{1n}^{2qv+2\Delta+1}}\leq n^{q\frac{2-2\alpha}{2-\alpha}}$$

if $\widetilde{\delta}_{2N}(\tau)=O\left(\frac{1}{\mu_{1n}}\left(\frac{\mu_{1n}}{\kappa_{1n}}\right)^{\frac{2qv+2\Delta+1}{3q+2\Delta+1}}n^{-\frac{q}{3q+2\Delta+1}}\right)$. This relationship holds when, for example, $v\to 2,\Delta\to 0,\alpha=1/2$ and $\mu_{1n}=n^{-1/18},\kappa_{1n}=n^{-1/9},q\geq 2/5$. Besides, (4.12) does not rely on Assumption 22 which assumes that for all patients at least one action can generate a positive mean reward given individual covariates, and thus allow the existence of nonrespondents.

The assumptions and results can be easily generalized to any $K\geq 2$. Assume that $\widetilde{f}_k:\mathcal{X}\to\{1,-1\}$ is a binary mapping learnt from a dataset of size N_k for all $k=2,\ldots,K$.

Assumption 21′. For any $k=2,\ldots,K$, assume that the estimator of the secondary ITR $\widetilde{f}_k$ converges to f_k^* in the sense that $\mathbb{P}(\widetilde{f}_kf_k^*<0)\leq\widetilde{\delta}_{kN_k}(\tau)$ with probability greater than or equal to $1-e^{\tau}$, where N_k is the sample size of the dataset and $\widetilde{\delta}_{kN_k}(\tau)=o(1)$ as $N_k\to\infty$.

Then Theorem 4.3.2 and 4.3.3 can be generalized as follows.

Theorem 4.3.4. *Under Assumptions 15-20 and 21′, the value function of the estimated FITR $\widehat{f}_{1n}$ satisfies*

$$\mathcal{V}_1(f_1^*)-\mathcal{V}_1(\widehat{f}_{1n})\lesssim(\delta_{1n}(\tau)+\mu_{1n})^{\frac{1}{2-\alpha}}, \tag{4.13}$$

with probability greater than or equal to $1 - Ke^{\tau}$*, and the agreement rate between* $\widehat{f}_{1n}$ *and any* f_k^* *for* $k \geq 2$ *satisfies*

$$\mathbb{P}(f_1^* f_k^* > 0) - \mathbb{P}(\widehat{f}_{1n} f_k^* > 0) \lesssim \frac{\delta_{1n}(\tau)}{\mu_{1n}} + \sum_{j=2}^{K} \widetilde{\delta}_{jN_j}(\tau) \tag{4.14}$$

with probability greater than or equal to $1 - Ke^{\tau}$*.*

4.4 Simulation Study

We conduct extensive simulation studies to demonstrate our proposed method (FITR) and compare its performance with separate learning (SepL) that does not use the auxiliary outcome ITRs. Each simulated dataset of size n contains K outcomes, where $K = 2$ or 3, and we treat one of them as the primary so the others are auxiliary outcomes. We let the dimension of covariates be $d = 10$ and there are n patients in total. The first two covariates are important variables and are generated as $X_{ij} \overset{i.i.d.}{\sim} Unif(-1, 1)$ for all $i = 1, \ldots, n$ and $j = 1, 2$. For the rest noise variables, let $X_{i3} = 0.8X'_{i3} + X_{i1}$, where $X'_{i3} \overset{i.i.d.}{\sim} Unif(-1, 1)$ so that X_{i3} is correlated with X_{i1}. The other variables are generated independently as $X_{ij} \overset{i.i.d.}{\sim} Unif(-1, 1)$ for all $j = 4, \ldots, d$. We assume the data are collected from a randomized controlled trial and $\pi(1; \boldsymbol{X}_i) = \pi(-1; \boldsymbol{X}_i) = 0.5$ for all $i = 1, \ldots, n$. The kth outcome is defined as $R_{ik} = m_k(\boldsymbol{X}_i) + T_k(\boldsymbol{X}_i, A_i) + \epsilon_k(\boldsymbol{X}_i, A_i)$, where m_k is the main effect, T_k is the interaction effect between the covariates and the treatment, and ϵ_k is the noise term. By choosing different T_k, we allow both linear and nonlinear treatment rules. We repeat the simulation process 400 times under each scenario.

To implement our method, the ITRs for the auxiliary outcomes are estimated using SepL in the same dataset and we then learn FITR for the primary one using the proposed algorithm. The similarity matrix $\boldsymbol{\Omega}$ is the Pearson correlation matrix based on K outcomes. Our experiments show that Spearman's rank correlation generates similar results. We implement both FITR-Ramp and FITR-IntL with the linear kernel and the Gaussian kernel. We use `scipy` package in `Python` to solve the optimization problem. For FITR-Ramp, our experiments show that the Powell algorithm usually achieves better optimization results than DCA, so we use the function `minimize(method='Powell')` to obtain solution. For FITR-Int, we use the function `minimize(method='BFGS')` with specified gradients for optimization. The tuning parameter λ_{kn} is first chosen with cross-validation when estimating the ITR using SepL for each reward $k = 1, \ldots, K$.

Then the parameter μ_{kn} in FITR-IntL or the pamameters μ_{kn} and κ_{kn} in FITR-Ramp are tuned simultaneously with cross-validation while λ_{kn} is kept fixed. The parameter σ in the Gaussian kernel is chosen as the median of the distances between all covariate pairs.

The value functions are calculated numerically from an independent test set of size 100,000. For the optimal value $\mathcal{V}_k^*$, the treatment is taken as $\arg\max_{A_i} T_k(\boldsymbol{X}_i, A_i)$ for the kth outcome of the ith patient. For the value function of each learned ITR, the treatment follows the corresponding ITR. The correct decision ratio of $\widehat{f}_k$, or accuracy, is estimated by averaging $I(\widehat{f}_k(\boldsymbol{X}) = f_k^*(\boldsymbol{X}))$ on this test set. The agreement rate between any two decision functions f and f' is estimated by averaging $I(f(\boldsymbol{X}) = f'(\boldsymbol{X}))$ on this test set.

4.4.1 Learning FITRs

In this section, we examine performance with $K = 2$ outcomes. The experiment with $K = 3$ is included in the Supplementary Material. Consider the following two scenarios. In both scenarios, the main effects are set to be

$$m_1(\boldsymbol{X}) = 1 + 2X_1 + X_2^2 + X_1X_2, \quad m_2(\boldsymbol{X}) = 1 + 2X_1^2 + 1.5X_2 + 0.5X_1X_2.$$

The residuals $(\epsilon_1, \ldots, \epsilon_K)$ of each patient follow a mean zero multivariate normal distribution, where the covariance matrix has 0.2 on its diagonal and 0.1 on its off-diagonal entries. The interaction terms are defined in two scenarios as follows:

1. Linear scenario $T_1(\boldsymbol{X}, A) = \gamma_1 A(0.2 - X_1 - 2X_2)$, $T_2(\boldsymbol{X}, A) = \gamma_2 A(0.2 - X_1 - 1.8X_2)$;
2. Nonlinear scenario $T_1(\boldsymbol{X}, A) = \gamma_1 A(0.9 - X_1^2 - X_2^2)$, $T_2(\boldsymbol{X}, A) = \gamma_2 A(1 - X_1^2 - 1.2X_2^2)$.

Here γ_1, γ_2 are fixed parameters controlling the ratio between the strength of heterogeneous treatment effect and the noise. We test different kernels and models for various sample sizes n and parameters γ_1, γ_2. The true optimal values of each scenario is summarized in Table 4.5. The computation time of one replication when $n = 200$ is about 0.22 seconds for FITR-IntL, 23 seconds for FITR-Ramp when using the linear kernel, and about 6 seconds for FITR-IntL, 860 seconds for FITR-Ramp when using the Gaussian kernel.

We report the root mean square error (RMSE) of each model across all replications and compare it with SepL in Tables 4.1 and 4.2 for scenarios 1 and 2 correspondingly. The accuracy is shown in Figures 4.1 for scenarios 1 and 2. The results are presented for each outcome $R_k, k = 1, \ldots, K$ as if it is treated as the primary outcome and the other $R_j, j \neq k$ are treated as the auxiliary outcomes. Since there are outliers that affect the visual display, we truncate the vertical axis to demonstrate the results clearly (approximately 5% were outliers not shown in each subfigure).

Table 4.1 suggests that for the value function in scenario 1, FITR-Ramp is better than FITR-IntL, and the linear kernel is better than the Gaussian kernel. Table 4.2 shows in scenario 2, FITR-IntL is better than FITR-Ramp, and the Gaussian kernel is better than the linear kernel. The outcome with larger treatment effect compared to the noise, for example R_2 with $\gamma_2 = 0.75$, has relatively small improvement from SepL to FITR, but as the auxiliary outcome it can help improve $\widehat{\mathcal{D}}_1$ significantly. For the accuracy, FITR-Ramp with linear kernel in scenario 1 and FITR-IntL with Gaussian kernel in scenario 2 generally has larger mean and smaller variance in Figure 4.1. To summarize, in scenario 1, FITR-Ramp with linear kernel can reduce the RMSE of SepL up to 30.33%, and reduce the standard error of the accuracy up to 28.91%. In scenario 2, FIRT-IntL with Gaussian kernel can reduce RMSE up to 7.92%, and reduce the standard error of the accuracy up to 47.98%.

The agreement rates between $\widehat{f}_{1n}$ and $\widetilde{f}_{2n}$ or $\widehat{f}_{2n}$ and $\widetilde{f}_{1n}$ and their standard deviations across all replications are also included in Tables 4.1 and 4.2. The two agreement rates are equal for SepL since $\widehat{f}_{kn} = \widetilde{f}_{kn}$ for all k for SepL. We can see that the fusion penalty in FITR-Ramp and FITR-IntL indeed increases the agreement rate between an FITR and the corresponding auxiliary outcome ITR, but never exceeds the true value $\mathbb{P}(f_1^* = f_2^*)$, which is 98.51% for scenario 1 and 95.80% for scenario 2. In addition, the standard deviations of the agreement rates are also reduced.

4.4.2 Sensitivity Analysis

To demonstrate the influence of the similarity between outcomes on the value function and classification accuracy of FITRs, we fix the first outcome and vary the second outcome when $K = 2$. Specifically, we use the same main effect and noise term as Section 4.4.1 and let the

$n/(\gamma_1, \gamma_2)$	Kernel	Model	RMSE_1	RMSE_2	$\frac{\text{RMSE}_1}{\text{RMSE}_{1,SepL}}$	$\frac{\text{RMSE}_2}{\text{RMSE}_{2,SepL}}$	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{2n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{1n})$
200 (0.5, 0.75)	Linear	SepL	0.0236	0.0265	1.0000	1.0000	0.9163 (0.0253)	0.9163 (0.0253)
		FITR-IntL	0.0223	0.0274	0.9434	1.0311	0.9376 (0.0224)	0.9335 (0.0252)
		FITR-Ramp	0.0164	0.0217	0.6967	0.8168	0.9430 (0.0205)	0.9393 (0.0240)
	Gaussian	SepL	0.0764	0.0922	1.0000	1.0000	0.8757 (0.0620)	0.8757 (0.0620)
		FITR-IntL	0.0812	0.0880	1.0622	0.9549	0.8998 (0.0547)	0.8944 (0.0561)
		FITR-Ramp	0.0646	0.0671	0.8460	0.7279	0.8946 (0.0471)	0.8925 (0.0503)
200 (0.5, 0.5)	Linear	SepL	0.0236	0.0346	1.0000	1.0000	0.8965 (0.0307)	0.8965 (0.0307)
		FITR-IntL	0.0258	0.0286	1.0914	0.8261	0.9224 (0.0277)	0.9222 (0.0297)
		FITR-Ramp	0.0199	0.0240	0.8430	0.6941	0.9263 (0.0274)	0.9299 (0.0285)
	Gaussian	SepL	0.0764	0.1218	1.0000	1.0000	0.8365 (0.0852)	0.8365 (0.0852)
		FITR-IntL	0.0836	0.1103	1.0938	0.9055	0.8662 (0.0699)	0.8674 (0.0817)
		FITR-Ramp	0.0689	0.0827	0.9021	0.6796	0.8591 (0.0725)	0.8735 (0.0583)
300 (0.5, 0.75)	Linear	SepL	0.0165	0.0188	1.0000	1.0000	0.9342 (0.0196)	0.9342 (0.0196)
		FITR-IntL	0.0174	0.0234	1.0556	1.2452	0.9487 (0.0185)	0.9465 (0.0186)
		FITR-Ramp	0.0127	0.0160	0.7690	0.8541	0.9544 (0.0168)	0.9519 (0.0187)
	Gaussian	SepL	0.0353	0.0353	1.0000	1.0000	0.9056 (0.0332)	0.9056 (0.0332)
		FITR-IntL	0.0305	0.0457	0.8620	1.2930	0.9210 (0.0305)	0.9192 (0.0355)
		FITR-Ramp	0.0231	0.0357	0.6531	1.0117	0.9202 (0.0299)	0.9207 (0.0288)
300 (0.5, 0.5)	Linear	SepL	0.0165	0.0218	1.0000	1.0000	0.9181 (0.0231)	0.9181 (0.0231)
		FITR-IntL	0.0181	0.0203	1.1014	0.9310	0.9351 (0.0219)	0.9386 (0.0229)
		FITR-Ramp	0.0139	0.0160	0.8455	0.7335	0.9419 (0.0221)	0.9418 (0.0233)
	Gaussian	SepL	0.0353	0.0781	1.0000	1.0000	0.8760 (0.0597)	0.8760 (0.0597)
		FITR-IntL	0.0422	0.0743	1.1955	0.9511	0.8944 (0.0527)	0.9002 (0.0605)
		FITR-Ramp	0.0348	0.0476	0.9854	0.6090	0.8929 (0.0536)	0.9023 (0.0392)

Table 4.1: The RMSEs of value functions, their ratios between SepL and FITR, and the agreement rates between $\widehat{f}_{1n}$ and $\widetilde{f}_{2n}$ or $\widehat{f}_{2n}$ and $\widetilde{f}_{1n}$ under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 1.

$n/(\gamma_1, \gamma_2)$	Kernel	Model	RMSE_1	RMSE_2	$\frac{\text{RMSE}_1}{\text{RMSE}_{1,SepL}}$	$\frac{\text{RMSE}_2}{\text{RMSE}_{2,SepL}}$	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{2n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{1n})$
200 (0.5, 0.75)	Linear	SepL	0.1125	0.1631	1.0000	1.0000	0.8434 (0.1113)	0.8434 (0.1113)
		FITR-IntL	0.0966	0.1532	0.8588	0.9392	0.9127 (0.0827)	0.8817 (0.0936)
		FITR-Ramp	0.1028	0.1657	0.9136	1.0161	0.9050 (0.0788)	0.8946 (0.0808)
	Gaussian	SepL	0.0936	0.1399	1.0000	1.0000	0.8196 (0.1403)	0.8196 (0.1403)
		FITR-IntL	0.0862	0.1364	0.9208	0.9755	0.8682 (0.1210)	0.8541 (0.1322)
		FITR-Ramp	0.0890	0.1386	0.9501	0.9909	0.8505 (0.1268)	0.8371 (0.1344)
200 (0.5, 0.5)	Linear	SepL	0.1125	0.1283	1.0000	1.0000	0.8036 (0.1171)	0.8036 (0.1171)
		FITR-IntL	0.1041	0.1122	0.9249	0.8751	0.8631 (0.1003)	0.8721 (0.0983)
		FITR-Ramp	0.1093	0.1203	0.9712	0.9377	0.8671 (0.0939)	0.8794 (0.0882)
	Gaussian	SepL	0.0936	0.1078	1.0000	1.0000	0.8032 (0.1575)	0.8032 (0.1575)
		FITR-IntL	0.0882	0.0993	0.9415	0.9213	0.8509 (0.1400)	0.8479 (0.1388)
		FITR-Ramp	0.0905	0.1040	0.9669	0.9653	0.8296 (0.1432)	0.8308 (0.1409)
300 (0.5, 0.75)	Linear	SepL	0.1015	0.1519	1.0000	1.0000	0.8964 (0.0884)	0.8964 (0.0884)
		FITR-IntL	0.0915	0.1456	0.9011	0.9586	0.9404 (0.0641)	0.9206 (0.0764)
		FITR-Ramp	0.0957	0.1531	0.9425	1.0078	0.9355 (0.0557)	0.9271 (0.0677)
	Gaussian	SepL	0.0846	0.1289	1.0000	1.0000	0.8208 (0.1204)	0.8208 (0.1204)
		FITR-IntL	0.0811	0.1276	0.9590	0.9896	0.8619 (0.1138)	0.8576 (0.1143)
		FITR-Ramp	0.0830	0.1283	0.9813	0.9946	0.8447 (0.1170)	0.8378 (0.1182)
300 (0.5, 0.5)	Linear	SepL	0.1015	0.1186	1.0000	1.0000	0.8536 (0.1049)	0.8536 (0.1049)
		FITR-IntL	0.0957	0.1046	0.9430	0.8816	0.8936 (0.0855)	0.9098 (0.0834)
		FITR-Ramp	0.1019	0.1104	1.0040	0.9307	0.9028 (0.0705)	0.9117 (0.0708)
	Gaussian	SepL	0.0846	0.0989	1.0000	1.0000	0.8093 (0.1333)	0.8093 (0.1333)
		FITR-IntL	0.0821	0.0926	0.9708	0.9362	0.8533 (0.1251)	0.8521 (0.1205)
		FITR-Ramp	0.0842	0.0966	0.9956	0.9764	0.8348 (0.1289)	0.8329 (0.1264)

Table 4.2: The RMSEs of value functions, their ratios between SepL and FITR, and the agreement rates between $\widehat{f}_{1n}$ and $\widetilde{f}_{2n}$ or $\widehat{f}_{2n}$ and $\widetilde{f}_{1n}$ under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 2.

ρ	$\mathbb{P}(f_1^* = f_2^*)$	Model	RMSE_1	RMSE_2	$\frac{\text{RMSE}_1}{\text{RMSE}_{1,SepL}}$	$\frac{\text{RMSE}_2}{\text{RMSE}_{2,SepL}}$	Accuracy_1	Accuracy_2
1	100%	SepL	0.0236	0.0244	1.0000	1.0000	0.9234 (0.0256)	0.9345 (0.0192)
		FITR-IntL	0.0208	0.0267	0.8805	1.0928	0.9287 (0.0223)	0.9328 (0.0214)
		FITR-Ramp	0.0157	0.0218	0.6644	0.8937	0.9377 (0.0178)	0.9389 (0.0178)
0.75	95.55%	SepL	0.0236	0.0311	1.0000	1.0000	0.9234 (0.0256)	0.9135 (0.0264)
		FITR-IntL	0.0266	0.0303	1.1272	0.9751	0.9217 (0.0300)	0.9153 (0.0256)
		FITR-Ramp	0.0195	0.0241	0.8258	0.7775	0.9294 (0.0208)	0.9240 (0.0211)
0.5	87.56%	SepL	0.0236	0.0382	1.0000	1.0000	0.9234 (0.0256)	0.8904 (0.0335)
		FITR-IntL	0.0340	0.0353	1.4391	0.9229	0.9112 (0.0351)	0.8908 (0.0296)
		FITR-Ramp	**0.0298**	0.0304	**1.2641**	0.7957	**0.9112 (0.0282)**	0.8964 (0.0241)
0.25	75.65%	SepL	0.0236	0.0457	1.0000	1.0000	0.9234 (0.0256)	0.8733 (0.0437)
		FITR-IntL	0.0396	0.0488	1.6793	1.0664	0.9036 (0.0404)	0.8631 (0.0427)
		FITR-Ramp	0.0446	**0.0502**	1.8914	**1.0973**	0.8953 (0.0421)	**0.8563 (0.0387)**
0	62.94%	SepL	0.0236	0.0515	1.0000	1.0000	0.9234 (0.0256)	0.8696 (0.0523)
		FITR-IntL	0.0392	0.0552	1.6596	1.0724	0.9059 (0.0407)	0.8601 (0.0486)
		FITR-Ramp	0.0505	0.0644	2.1385	1.2514	0.8939 (0.0499)	0.8505 (0.0538)
-0.25	50.64%	SepL	0.0236	0.0425	1.0000	1.0000	0.9234 (0.0256)	0.8762 (0.0410)
		FITR-IntL	0.0324	0.0453	1.3743	1.0657	0.9136 (0.0337)	0.8725 (0.0400)
		FITR-Ramp	0.0489	0.0558	2.0702	1.3118	0.8990 (0.0497)	0.8650 (0.0485)

Table 4.3: The change of RMSE and accuracy when the similarity between outcomes is changed.

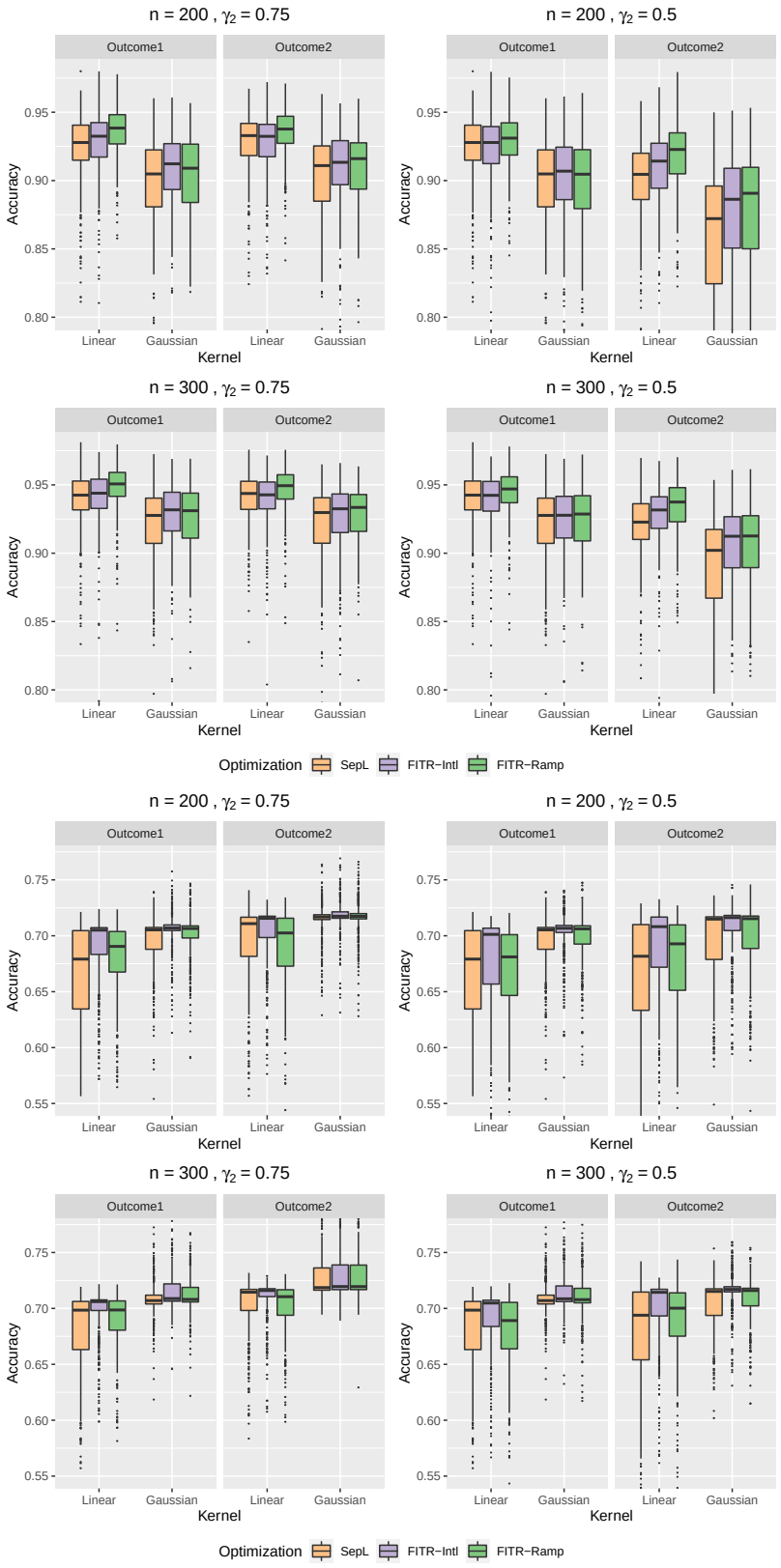

Figure 4.1: The accuracy of SepL, FITR-Ramp and FITR-Intl under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 1 and 2 when $K = 2$.

interaction effect be

$$T_1(\boldsymbol{X}, A) = 0.5A(0.2 - X_1 - 2X_2), \quad T_2(\boldsymbol{X}, A) = 0.75A(0.2 - X_1 - 2\rho X_2),$$

where ρ controls the similarity between R_1 and R_2. In Table 4.3, we show different values of ρ and their corresponding agreement rates between f_1^* and f_2^*. The sample size is $n = 200$.

The value function and accuracy of FITRs is larger than SepL when $\rho \geq 0.75$ for R_1 and when $\rho \geq 0.5$ for R_2. This indicates that the agreement rate should be no less than 87% in order for the fusion penalty to have a positive effect on the FITRs when the signal is weak ($\gamma_1 = 0.5$), and be no less than 75% when the signal is strong ($\gamma_2 = 0.75$).

4.5 Real Data Analysis

We apply our proposed methods to the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC; Trivedi et al., 2016) study, which randomized patients with major depressive disorder (MDD) to serotonin selective reuptake inhibitor sertraline (SERT) or placebo (PBO). Our primary outcome for depressive symptoms is the Quick Inventory of Depressive Symptomatology (QIDS) score, where a lower score represents better relief of symptoms. Two covariates were shown to be informative for tailoring treatments in a prior study (Chen et al., 2021b). The first is the NEO-Five Factor Inventory score, where NEO Personality Inventory is a 240-item measurement designed to assess personality in the domains of neuroticism, extraversion, openness and so on, we focus on the neuroticism domain. The second informative measure is the Flanker Interference Accuracy score, where a higher value indicates reduced cognitive control. Five additional baseline variables are used, including sex, age, education years, Edinburgh Handedness Inventory (EHI) score, and the QIDS score at the beginning of the study.

We consider three different outcomes as rewards. The primary outcome is the difference of QIDS at the beginning and the end of study (QIDS-change), with a larger value more desirable. The auxiliary outcome is clinician assessed Clinical Global Improvement scale (CGI) as an overall assessment of treatment effect, and a smaller score suggests better improvement (Busner and Targum, 2007). Another auxiliary measure is the Social Adjustment Scale (SAS), which evaluates the impact of a person's mental health difficulties, with a higher score indicating greater impairment.

Value functions from 400 replicates				
All SERT	All PBO	SepL	FITR-IntL	FITR-Ramp
8.0000	6.4375	8.7750 (1.0291)	8.9387 (1.0474)	8.9419 (1.0505)

Agreement rates between primary ITR and auxiliary ITRs			
Auxiliary ITR		$\widetilde{f}_{\text{CGI}}$	$\widetilde{f}_{\text{SAS}}$
Primary ITR from different methods	SepL	0.8065	0.6935
	FITR-IntL	0.9032	0.7366
	FITR-Ramp	0.8871	0.7204

Table 4.4: The upper half table shows the estimated value functions of the OSFA strategy, SepL, FITR-IntL and FITR-Ramp for the primary outcome QIDS-change in the EMBARC study. The lower half table shows the agreement rates between SepL, FITR-Ramp, FITR-Ramp and the auxiliary outcome ITRs $\widetilde{f}_{\text{CGI}}$, $\widetilde{f}_{\text{SAS}}$ for QIDS-change.

Consequently, we use QIDS-change along with the negative values of CGI and SAS as the three outcomes. There are 186 patients with complete information.

We compare the estimated value functions of FITR-IntL, FITR-Ramp, SepL and the one-size-for-all (OSFA) strategy in Table 4.4. For SepL, FITR-IntL and FITR-Ramp, we split the dataset into the training and test set with probability 0.7 and 0.3 respectively. The training set is used to estimate the ITRs, where the tuning parameters are selected using 4-fold cross-validation. The test set is used to estimate the value functions $\mathcal{V}_k(f_k)$ with the unbiased estimator

$$\sum_{i=1}^{n} \frac{R_{ik} I(A_i = \text{sign}\{f_k(\boldsymbol{X}_i)\})}{\pi_i(A_i; \boldsymbol{X}_i)} \Bigg/ \sum_{i=1}^{n} \frac{I(A_i = \text{sign}\{f_k(\boldsymbol{X}_i)\})}{\pi_i(A_i; \boldsymbol{X}_i)}. \tag{4.15}$$

The process is repeated 400 times. The mean and standard error of the value functions across the 400 replications are reported. For OSFA, we directly calculate the mean reward on the subpopulation with the desired treatment (also equals (4.15) since $\pi_i = 0.5$ for all i). Our experiments show that the linear kernel is better than the Gaussian kernel on this dataset, so only the results of the linear kernel are presented for SepL, FITR-IntL and FITR-Ramp.

The upper half of Table 4.4 suggests that FITR-IntL and FITR-Ramp can improve the ITR learning for QIDS-change assisted by the other two auxiliary outcomes, when comparing with OSFA and SepL. FITR-Ramp and FITR-IntL have similar performance. To compare the learned ITR using different methods, we present the coefficient of each variable in Table 4.11 when the linear kernel is used. For the primary outcome QIDS, the signs of the coefficients of SepL, FITR-Ramp and FITR-IntL are all the same. To see the impact of the coefficient differences on SepL and

FITR, we present the agreement rates between SepL, FITR-Ramp, FITR-Ramp and the auxiliary outcome ITRs $\widetilde{f}_{\text{CGI}}$, $\widetilde{f}_{\text{SAS}}$ in the lower half of Table 4.4. We can see that the agreement rates between FITR-Ramp or FITR-Ramp and the auxiliary outcome ITRs are greater than that for SepL, which is consistent with the faster convergence rate proved in Theorem 4.3.3.

4.6 Discussion

In this work, we have proposed a method to borrow auxiliary outcome ITRs to optimize the primary outcome while also optimizing auxiliary outcomes as much as possible. A fusion penalty was introduced to regularize the average number of inconsistent treatments suggested by the ITRs of each outcome pairs, which encourages the ITRs to yield similar treatment recommendations. We showed theoretically and numerically that the agreement rate between the proposed FITR for the primary outcome and the ITR for the auxiliary outcome converges faster to its true value than SepL. Besides, the simulation and real data study suggested that the learned FITRs have better value function and accuracy than SepL when the true optimal ITRs are close and the sample size is relatively small.

The fusion penalty share the same spirits with the Laplacian penalty Huang et al. (2011), which is used to learn multiple models simultaneously while encouraging similarity. Specifically, the Laplacian matrix is defined as $\mathbf{L} := \mathbf{D} - \mathbf{\Omega}$ with size $K \times K$, where $\mathbf{\Omega}$ is the adjacency matrix of the rewards and $\mathbf{D}$ is the degree matrix with $D_{kk} = \sum_{l \neq k} A_{kl}$ and $D_{jk} = 0$ for $j \neq k$, $k, j = 1, \ldots, K$. The adjacency matrix can be constructed based on the Euclidean distance, correlation, or power adjacency function (Huang et al., 2011). Then the Laplacian penalty can be defined as

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ I(\boldsymbol{f}(\boldsymbol{X}_i) \geq 0) \cdot \mathbf{L} \cdot I(\boldsymbol{f}(\boldsymbol{X}_i) \geq 0)^T + I(\boldsymbol{f}(\boldsymbol{X}_i) \leq 0) \cdot \mathbf{L} \cdot I(\boldsymbol{f}(\boldsymbol{X}_i) \leq 0)^T \right\}$$

It prompts $f_k(\boldsymbol{X}_i)$ and $f_j(\boldsymbol{X}_i)$ to have the same sign when R_k are R_j are similar. This is equivalent to (4.3) if we use the same data for learning ITRs simultaneously, regardless of which outcome is primary.

There are several promising directions that worth further research. Currently the tuning parameters λ and μ only depends on the sample size n and is homogeneous for all patients. A possible extension is to let them depend on the covariates $\boldsymbol{X}$, since the fusion level should decrease if the

optimal ITRs for the primary outcome and auxiliary outcomes are far away for these patients. Another interesting direction is to explore the fusion penalty when estimating the dynamic treatment rules, which are a sequence of rules for multi-stage settings. We can apply the fusion penalty backwards from the last stage to the first stage. Since our method relies on the auxiliary outcome only through its treatment rules, we do not need individual-level data if such rules are already available. The covariates used for constructing the auxiliary outcome ITR may not even need to be the same as for FITR. Finally, the same fusion idea can be used to combine treatment rules for different outcomes from different studies, which can be treated as a form of meta-analysis for learning ITRs.

4.7 Supplementary Materials

In the supplementary section, we provide some addition details about the simulation study and real data analysis in Section 4.7.1 and prove the theoretical results in Section 4.7.2.

4.7.1 Additional Simulation and Real Data Experiment Results

In this section, we provide some addition details about the simulation in Section 4.4.1 and show the simulation results when there are $K = 3$ outcomes. In addition, we provide more results about the real data analysis in Section 4.5.

4.7.1.1 Similation Details for Section 4.4.1

For the two scenarios considered in Section 4.4.1, here we present their true optimal values.

(γ_1, γ_2)	Scenario 1		Scenario 2	
	$\mathcal{V}_1^*$	$\mathcal{V}_2^*$	$\mathcal{V}_1^*$	$\mathcal{V}_2^*$
(0.5, 0.75)	1.88	2.42	1.54	2.01
(0.5, 0.5)	1.88	2.17	1.54	1.89

Table 4.5: True optimal values of scenarios 1 and 2 when $K = 2$.

4.7.1.2 Learning ITRs for $K = 3$

In this section, we experiment with the setting where we have $K = 3$ outcomes.

The main effects are set to be

$$m_1(\boldsymbol{X}) = 1 + 2X_1 + X_2^2 + X_1X_2, \quad m_2(\boldsymbol{X}) = 1 + 2X_1^2 + 1.5X_2 + 0.5X_1X_2,$$
$$m_3(\boldsymbol{X}) = 1 + X_1 + X_2$$

in the two scenarios that we are considering. The noise terms are the same as that in Section 4.4.1. The interaction terms are defined differently in the two scenarios as follows:

3. Linear scenario $T_1(\boldsymbol{X}, A) = \gamma_1 A(0.2 - X_1 - 2X_2)$, $T_2(\boldsymbol{X}, A) = \gamma_2 A(0.2 - X_1 - 1.8X_2)$, $T_3(\boldsymbol{X}, A) = \gamma_3 A(0.2 - 0.8X_1 - 1.5X_2)$;

4. Nonlinear scenario $T_1(\boldsymbol{X}, A) = \gamma_1 A(0.9 - X_1^2 - X_2^2)$, $T_2(\boldsymbol{X}, A) = \gamma_2 A(1 - X_1^2 - 1.2X_2^2)$, $T_3(\boldsymbol{X}, A) = \gamma_3 A(0.9 - 0.8X_1^2 - 0.9X_2^2)$.

The first two outcomes are the same as in Section 4.4.1 when $K = 2$. Here $\gamma_1, \gamma_2, \gamma_3$ are fixed parameters controlling the ratio between the signal and the noise. The true optimal values of each scenario is summarized in Table 4.6.

(γ_1, γ_2)	Scenario 3			Scenario 4		
	$\mathcal{V}_1^*$	$\mathcal{V}_2^*$	$\mathcal{V}_3^*$	$\mathcal{V}_1^*$	$\mathcal{V}_2^*$	$\mathcal{V}_3^*$
(0.5, 0.75, 0.5)	1.88	2.42	1.42	1.54	2.01	1.21
(0.5, 0.5, 0.5)	1.88	2.17	1.42	1.54	1.89	1.21

Table 4.6: True optimal values of scenarios 3 and 4 when $K = 3$.

Table 4.7 and 4.9 shows the RMSE of scenario 3 and scenario 4 correspondingly. Figures 4.2 and 4.3 show the accuracy of each replication, with the y-axis truncated to display the boxes clearly. The general conclusion is similar to that of Section 4.4.1. However, we can see that the increase in the number of correlated outcomes can improve the learning of ITRs. Indeed, in scenario 3, FITR-Ramp with linear kernel can reduce the RMSE of SepL up to 36.59% among the first two outcomes and up to 49.34% for the third outcome. In addition, it reduces the standard error of the accuracy

$n/(\gamma_1, \gamma_2)$	Kernel	Model	RMSE_1	RMSE_2	RMSE_3	$\frac{\text{RMSE}_1}{\text{RMSE}_{1,SepL}}$	$\frac{\text{RMSE}_2}{\text{RMSE}_{2,SepL}}$	$\frac{\text{RMSE}_3}{\text{RMSE}_{3,SepL}}$
200 (0.5, 0.75, 0.5)	Linear	SepL	0.0271	0.0254	0.0208	1.0000	1.0000	1.0000
		FITR-IntL	0.0278	0.0311	0.0190	1.0261	1.2243	0.9158
		FITR-Ramp	0.0173	0.0218	0.0116	0.6374	0.8570	0.5573
	Gaussian	SepL	0.0845	0.0627	0.0702	1.0000	1.0000	1.0000
		FITR-IntL	0.0869	0.0857	0.0622	1.0277	1.3663	0.8862
		FITR-Ramp	0.0545	0.0555	0.0405	0.6444	0.8852	0.5775
200 (0.5, 0.5, 0.5)	Linear	SepL	0.0271	0.0330	0.0208	1.0000	1.0000	1.0000
		FITR-IntL	0.0326	0.0320	0.0204	1.2038	0.9700	0.9822
		FITR-Ramp	0.0202	0.0209	0.0136	0.7451	0.6341	0.6533
	Gaussian	SepL	0.0845	0.1038	0.0702	1.0000	1.0000	1.0000
		FITR-IntL	0.0927	0.1056	0.0679	1.0961	1.0170	0.9678
		FITR-Ramp	0.0581	0.0706	0.0448	0.6876	0.6800	0.6389
300 (0.5, 0.75, 0.5)	Linear	SepL	0.0169	0.0184	0.0147	1.0000	1.0000	1.0000
		FITR-IntL	0.0193	0.0242	0.0131	1.1412	1.3122	0.8920
		FITR-Ramp	0.0121	0.0149	0.0074	0.7157	0.8077	0.5066
	Gaussian	SepL	0.0367	0.0382	0.0412	1.0000	1.0000	1.0000
		FITR-IntL	0.0288	0.0426	0.0364	0.7850	1.1140	0.8830
		FITR-Ramp	0.0242	0.0320	0.0225	0.6578	0.8366	0.5473
300 (0.5, 0.5, 0.5)	Linear	SepL	0.0169	0.0209	0.0147	1.0000	1.0000	1.0000
		FITR-IntL	0.0202	0.0201	0.0131	1.1937	0.9597	0.8939
		FITR-Ramp	0.0129	0.0135	0.0085	0.7653	0.6438	0.5774
	Gaussian	SepL	0.0426	0.0878	0.0353	1.0000	1.0000	1.0000
		FITR-IntL	0.0412	0.0844	0.0345	0.9674	0.9614	0.9765
		FITR-Ramp	0.0368	0.0508	0.0287	0.8640	0.5789	0.8136

Table 4.7: The RMSEs of value functions and their ratios between SepL and FITR under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 3.

up to 32.66% among the first two outcomes and up to 47.25% for the third outcome. In scenario 4, FITR-IntL with Gaussian kernel can reduce the RMSE of SepL up to 10.29% among the first two outcomes and up to 11.37% for the third outcome. In addition, it reduces the standard error of the accuracy up to 52.29% among the first two outcomes and up to 34.14% for the third outcome. The improvement in RMSE and accuracy of R_1 and R_2 is larger than that in Section 4.4.1, suggesting that additional outcomes can further help improve the ITRs. Table 4.8 and 4.10 demonstrate the increase of the agreement rate between an FITR and the corresponding auxiliary outcome ITRs compared to SepL.

$n/(\gamma_1, \gamma_2)$	Kernel	Model	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{2n})$	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{3n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{1n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{3n})$	$\widehat{\mathbb{P}}(\widehat{f}_{3n} = \widetilde{f}_{1n})$	$\widehat{\mathbb{P}}(\widehat{f}_{3n} = \widetilde{f}_{2n})$
200 (0.5, 0.75, 0.5)	Linear	SepL	0.9186 (0.0253)	0.9175 (0.0300)	0.9186 (0.0253)	0.9206 (0.0287)	0.9175 (0.0300)	0.9206 (0.0287)
		FITR-IntL	0.9351 (0.0244)	0.9355 (0.0236)	0.9319 (0.0239)	0.9345 (0.0225)	0.9343 (0.0251)	0.9351 (0.0243)
		FITR-Ramp	0.9418 (0.0197)	0.9373 (0.0261)	0.9380 (0.0231)	0.9386 (0.0251)	0.9402 (0.0215)	0.9431 (0.0187)
	Gaussian	SepL	0.8779 (0.0593)	0.8702 (0.0778)	0.8779 (0.0593)	0.8770 (0.0655)	0.8702 (0.0778)	0.8770 (0.0655)
		FITR-IntL	0.8972 (0.0618)	0.8966 (0.0634)	0.8963 (0.0509)	0.8977 (0.0550)	0.8965 (0.0686)	0.9009 (0.0601)
		FITR-Ramp	0.8945 (0.0443)	0.8924 (0.0566)	0.8892 (0.0514)	0.8900 (0.0540)	0.8972 (0.0510)	0.9013 (0.0429)
200 (0.5, 0.5, 0.5)	Linear	SepL	0.9012 (0.0311)	0.9175 (0.0300)	0.9012 (0.0311)	0.9045 (0.0315)	0.9175 (0.0300)	0.9045 (0.0315)
		FITR-IntL	0.9185 (0.0286)	0.9342 (0.0245)	0.9254 (0.0279)	0.9276 (0.0271)	0.9339 (0.0256)	0.9199 (0.0254)
		FITR-Ramp	0.9268 (0.0259)	0.9373 (0.0239)	0.9312 (0.0256)	0.9311 (0.0269)	0.9391 (0.0204)	0.9256 (0.0262)
	Gaussian	SepL	0.8502 (0.0775)	0.8702 (0.0778)	0.8502 (0.0775)	0.8504 (0.0794)	0.8702 (0.0778)	0.8504 (0.0794)
		FITR-IntL	0.8719 (0.0705)	0.8954 (0.0658)	0.8794 (0.0763)	0.8852 (0.0725)	0.8960 (0.0684)	0.8753 (0.0714)
		FITR-Ramp	0.8644 (0.0651)	0.8904 (0.0572)	0.8774 (0.0477)	0.8770 (0.0557)	0.8952 (0.0529)	0.8707 (0.0651)
300 (0.5, 0.75, 0.5)	Linear	SepL	0.9336 (0.0217)	0.9316 (0.0255)	0.9336 (0.0217)	0.9338 (0.0240)	0.9316 (0.0255)	0.9338 (0.0240)
		FITR-IntL	0.9453 (0.0199)	0.9450 (0.0207)	0.9435 (0.0217)	0.9446 (0.0213)	0.9464 (0.0202)	0.9467 (0.0172)
		FITR-Ramp	0.9520 (0.0157)	0.9480 (0.0243)	0.9502 (0.0193)	0.9482 (0.0232)	0.9530 (0.0169)	0.9538 (0.0145)
	Gaussian	SepL	0.9053 (0.0383)	0.9007 (0.0509)	0.9053 (0.0383)	0.9038 (0.0437)	0.9007 (0.0509)	0.9038 (0.0437)
		FITR-IntL	0.9214 (0.0317)	0.9175 (0.0413)	0.9191 (0.0355)	0.9193 (0.0370)	0.9234 (0.0458)	0.9252 (0.0368)
		FITR-Ramp	0.9191 (0.0286)	0.9135 (0.0439)	0.9149 (0.0320)	0.9142 (0.0376)	0.9198 (0.0307)	0.9215 (0.0303)
300 (0.5, 0.5, 0.5)	Linear	SepL	0.9190 (0.0250)	0.9316 (0.0255)	0.9190 (0.0250)	0.9201 (0.0265)	0.9316 (0.0255)	0.9201 (0.0265)
		FITR-IntL	0.9321 (0.0224)	0.9439 (0.0217)	0.9381 (0.0241)	0.9396 (0.0225)	0.9462 (0.0198)	0.9338 (0.0204)
		FITR-Ramp	0.9394 (0.0210)	0.9468 (0.0233)	0.9442 (0.0219)	0.9414 (0.0247)	0.9516 (0.0164)	0.9395 (0.0196)
	Gaussian	SepL	0.8766 (0.0686)	0.9011 (0.0442)	0.8766 (0.0686)	0.8813 (0.0591)	0.9011 (0.0442)	0.8813 (0.0591)
		FITR-IntL	0.8946 (0.0586)	0.9188 (0.0375)	0.8982 (0.0702)	0.9056 (0.0576)	0.9217 (0.0394)	0.9035 (0.0504)
		FITR-Ramp	0.8889 (0.0573)	0.9147 (0.0350)	0.9019 (0.0458)	0.9045 (0.0408)	0.9177 (0.0358)	0.8955 (0.0558)

Table 4.8: The agreement rates between $\widehat{f}_{1n}, \widehat{f}_{2n}, \widehat{f}_{3n}$ and their corresponding auxiliary outcome ITRs under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 3.

$n/(\gamma_1, \gamma_2)$	Kernel	Model	RMSE_1	RMSE_2	RMSE_3	$\frac{\text{RMSE}_1}{\text{RMSE}_{1,SepL}}$	$\frac{\text{RMSE}_2}{\text{RMSE}_{2,SepL}}$	$\frac{\text{RMSE}_3}{\text{RMSE}_{3,SepL}}$
200 (0.5, 0.75, 0.5)	Linear	SepL	0.1164	0.1621	0.0495	1.0000	1.0000	1.0000
		FITR-IntL	0.0931	0.1467	0.0469	0.7999	0.9050	0.9470
		FITR-Ramp	0.0984	0.1540	0.0581	0.8455	0.9501	1.1723
	Gaussian	SepL	0.0937	0.1390	0.0520	1.0000	1.0000	1.0000
		FITR-IntL	0.0847	0.1344	0.0461	0.9043	0.9669	0.8863
		FITR-Ramp	0.0883	0.1360	0.0499	0.9423	0.9784	0.9599
200 (0.5, 0.5, 0.5)	Linear	SepL	0.1164	0.1282	0.0495	1.0000	1.0000	1.0000
		FITR-IntL	0.0937	0.1042	0.0493	0.8052	0.8130	0.9960
		FITR-Ramp	0.1002	0.1098	0.0623	0.8605	0.8562	1.2573
	Gaussian	SepL	0.0937	0.1054	0.0520	1.0000	1.0000	1.0000
		FITR-IntL	0.0855	0.0946	0.0476	0.9126	0.8971	0.9153
		FITR-Ramp	0.0893	0.0990	0.0499	0.9538	0.9388	0.9596
300 (0.5, 0.75, 0.5)	Linear	SepL	0.1023	0.1507	0.0432	1.0000	1.0000	1.0000
		FITR-IntL	0.0884	0.1411	0.0422	0.8645	0.9364	0.9761
		FITR-Ramp	0.0919	0.1444	0.0487	0.8983	0.9582	1.1262
	Gaussian	SepL	0.0856	0.1281	0.0444	1.0000	1.0000	1.0000
		FITR-IntL	0.0813	0.1277	0.0426	0.9501	0.9971	0.9602
		FITR-Ramp	0.0825	0.1264	0.0436	0.9634	0.9870	0.9818
300 (0.5, 0.5, 0.5)	Linear	SepL	0.1023	0.1147	0.0432	1.0000	1.0000	1.0000
		FITR-IntL	0.0894	0.0984	0.0435	0.8733	0.8584	1.0067
		FITR-Ramp	0.0930	0.1010	0.0516	0.9089	0.8806	1.1940
	Gaussian	SepL	0.0856	0.0984	0.0444	1.0000	1.0000	1.0000
		FITR-IntL	0.0818	0.0905	0.0430	0.9553	0.9201	0.9685
		FITR-Ramp	0.0826	0.0932	0.0437	0.9643	0.9479	0.9834

Table 4.9: The RMSEs of value functions and their ratios between SepL and FITR under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 4.

$n/(\gamma_1, \gamma_2)$	Kernel	Model	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{2n})$	$\widehat{\mathbb{P}}(\widehat{f}_{1n} = \widetilde{f}_{3n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{1n})$	$\widehat{\mathbb{P}}(\widehat{f}_{2n} = \widetilde{f}_{3n})$	$\widehat{\mathbb{P}}(\widehat{f}_{3n} = \widetilde{f}_{1n})$	$\widehat{\mathbb{P}}(\widehat{f}_{3n} = \widetilde{f}_{2n})$
200 (0.5, 0.75, 0.5)	Linear	SepL	0.8389 (0.1091)	0.8505 (0.1142)	0.8389 (0.1091)	0.9183 (0.0863)	0.8505 (0.1142)	0.9183 (0.0863)
		FITR-IntL	0.9138 (0.0848)	0.9459 (0.0703)	0.8586 (0.1090)	0.9577 (0.0607)	0.8611 (0.1111)	0.9306 (0.0770)
		FITR-Ramp	0.9080 (0.0750)	0.9237 (0.0689)	0.8747 (0.0897)	0.9400 (0.0526)	0.8744 (0.0896)	0.9210 (0.0654)
	Gaussian	SepL	0.8218 (0.1362)	0.8491 (0.1381)	0.8218 (0.1362)	0.8696 (0.1207)	0.8491 (0.1381)	0.8696 (0.1207)
		FITR-IntL	0.8782 (0.1164)	0.9084 (0.1032)	0.8605 (0.1346)	0.9126 (0.0995)	0.8741 (0.1348)	0.8939 (0.1134)
		FITR-Ramp	0.8517 (0.1231)	0.8785 (0.1179)	0.8423 (0.1336)	0.8899 (0.1088)	0.8583 (0.1363)	0.8785 (0.1178)
200 (0.5, 0.5, 0.5)	Linear	SepL	0.8000 (0.1108)	0.8505 (0.1142)	0.8000 (0.1108)	0.8509 (0.1163)	0.8505 (0.1142)	0.8509 (0.1163)
		FITR-IntL	0.8544 (0.1069)	0.9448 (0.0717)	0.8522 (0.1055)	0.9351 (0.0787)	0.8653 (0.1061)	0.8651 (0.1075)
		FITR-Ramp	0.8653 (0.0861)	0.9187 (0.0672)	0.8640 (0.0877)	0.9167 (0.0690)	0.8709 (0.0902)	0.8667 (0.0927)
	Gaussian	SepL	0.8052 (0.1469)	0.8491 (0.1381)	0.8052 (0.1469)	0.8476 (0.1396)	0.8491 (0.1381)	0.8476 (0.1396)
		FITR-IntL	0.8556 (0.1347)	0.9070 (0.1039)	0.8577 (0.1374)	0.9070 (0.1065)	0.8738 (0.1335)	0.8701 (0.1348)
		FITR-Ramp	0.8333 (0.1378)	0.8757 (0.1184)	0.8370 (0.1365)	0.8811 (0.1163)	0.8596 (0.1350)	0.8569 (0.1374)
300 (0.5, 0.75, 0.5)	Linear	SepL	0.8976 (0.0908)	0.9120 (0.0936)	0.8976 (0.0908)	0.9522 (0.0625)	0.9120 (0.0936)	0.9522 (0.0625)
		FITR-IntL	0.9456 (0.0654)	0.9725 (0.0478)	0.9125 (0.0911)	0.9793 (0.0376)	0.9171 (0.0912)	0.9575 (0.0590)
		FITR-Ramp	0.9367 (0.0626)	0.9538 (0.0556)	0.9194 (0.0757)	0.9680 (0.0350)	0.9229 (0.0777)	0.9536 (0.0493)
	Gaussian	SepL	0.8258 (0.1217)	0.8625 (0.1186)	0.8258 (0.1217)	0.8726 (0.1130)	0.8625 (0.1186)	0.8726 (0.1130)
		FITR-IntL	0.8723 (0.1105)	0.9205 (0.0865)	0.8622 (0.1144)	0.9168 (0.0878)	0.8798 (0.1162)	0.8867 (0.1100)
		FITR-Ramp	0.8532 (0.1143)	0.8956 (0.0974)	0.8446 (0.1200)	0.8938 (0.1016)	0.8689 (0.1183)	0.8788 (0.1129)
300 (0.5, 0.5, 0.5)	Linear	SepL	0.8625 (0.1014)	0.9120 (0.0936)	0.8625 (0.1014)	0.9012 (0.1002)	0.9120 (0.0936)	0.9012 (0.1002)
		FITR-IntL	0.8997 (0.0931)	0.9691 (0.0535)	0.9052 (0.0908)	0.9631 (0.0631)	0.9197 (0.0871)	0.9075 (0.0936)
		FITR-Ramp	0.9001 (0.0821)	0.9493 (0.0581)	0.9078 (0.0781)	0.9500 (0.0523)	0.9215 (0.0741)	0.9112 (0.0803)
	Gaussian	SepL	0.8166 (0.1344)	0.8625 (0.1186)	0.8166 (0.1344)	0.8602 (0.1336)	0.8625 (0.1186)	0.8602 (0.1336)
		FITR-IntL	0.8624 (0.1280)	0.9213 (0.0872)	0.8618 (0.1212)	0.9174 (0.0973)	0.8784 (0.1154)	0.8732 (0.1306)
		FITR-Ramp	0.8450 (0.1301)	0.8945 (0.0992)	0.8416 (0.1248)	0.8911 (0.1098)	0.8706 (0.1166)	0.8670 (0.1324)

Table 4.10: The agreement rates between $\widehat{f}_{1n}, \widehat{f}_{2n}, \widehat{f}_{3n}$ and their corresponding auxiliary outcome ITRs under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 4.

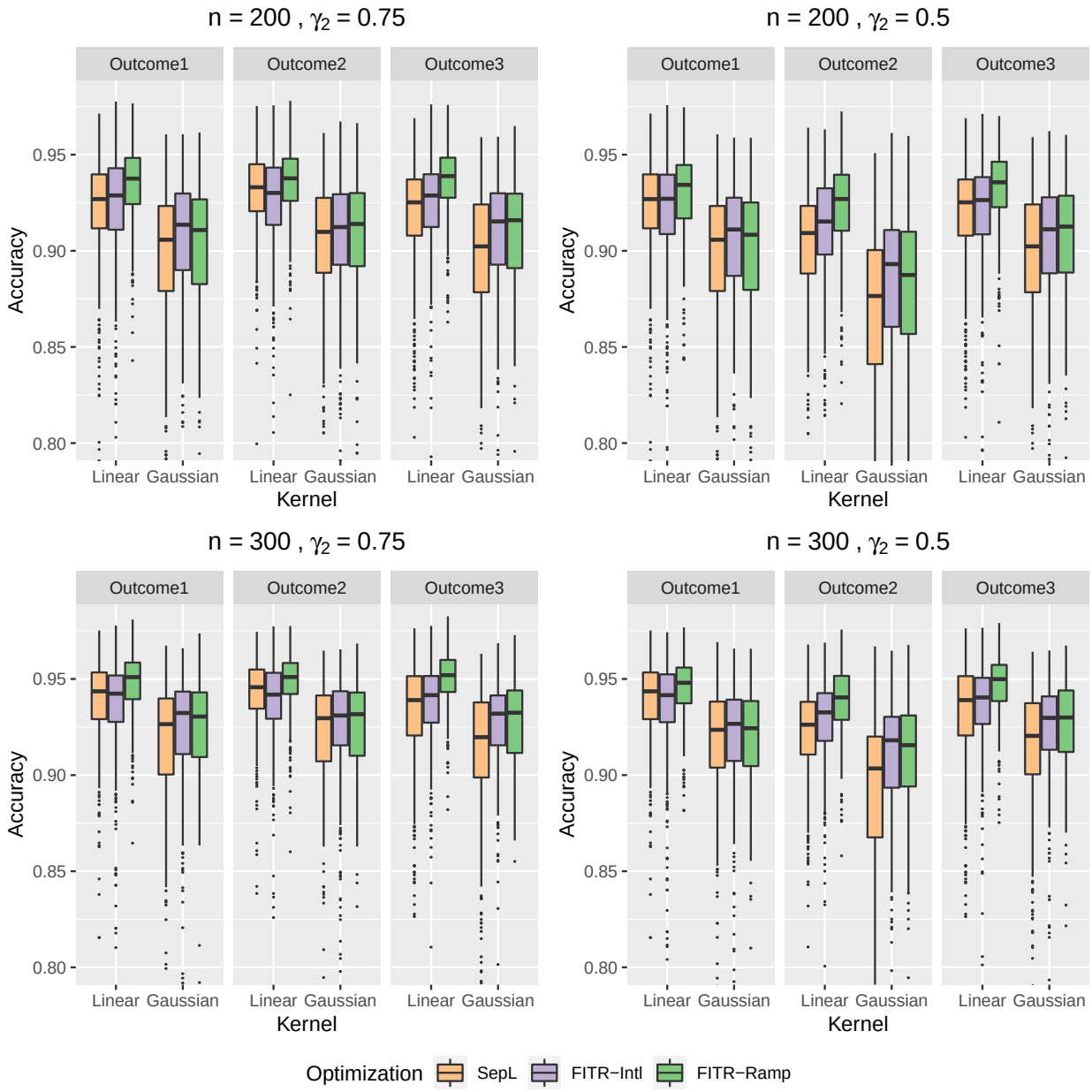

Figure 4.2: The accuracy of SepL, FITR-Ramp and FITR-Intl under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 3 when $K = 3$.

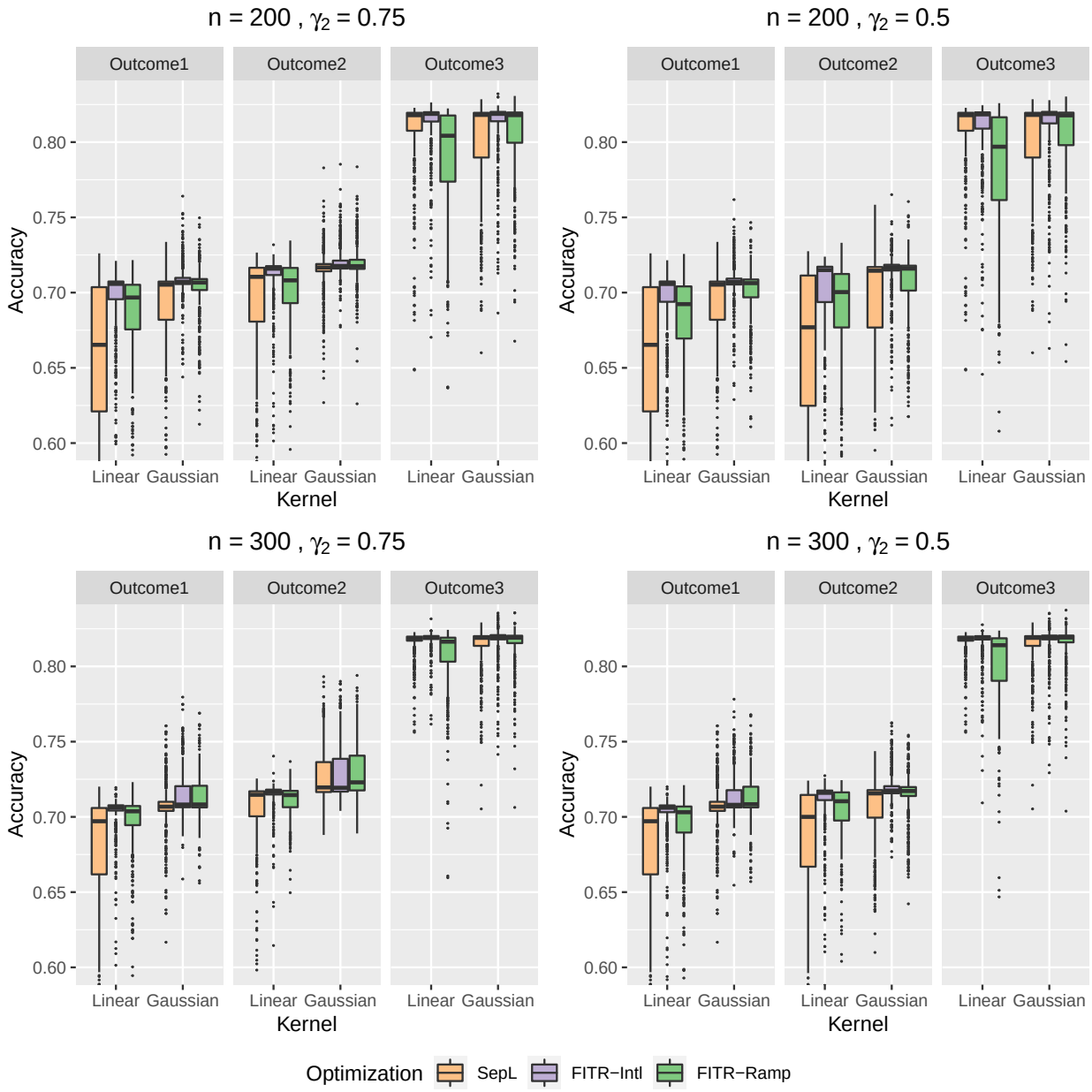

Figure 4.3: The accuracy of SepL, FITR-Ramp and FITR-Intl under different sample sizes n, parameters (γ_1, γ_2), models and kernels in scenario 4 when $K = 3$.

4.7.1.3 Additional Results for Section 4.5

We list the coefficients of the estimated ITRs by SepL, FITR-IntL and FITR-Ramp in Table 4.11. They are learnt with the complete dataset and the linear kernel. We can conclude that the coefficients fitted by different methods are generally close, which suggests that the fusion penalty will not dramatically change an ITR compared to SepL. The coefficients are more similar for QIDS-change and CGI, which is expected since SAS measures impacts of depression on social functioning. For the outcome SAS, Flanker variable has a different sign in terms of the ITR coefficients from different methods. While SepL estimates a positive coefficient, FITR-IntL and FITR-Ramp estimate the opposite.

	Variable	intercept	sex	age	education	EHI	QIDS	NEO	Flanker
QIDS-change	SepL	0.1302	0.1055	0.2313	-0.1256	0.0655	0.0308	0.2514	-0.3454
	FITR-IntL	0.1367	0.0983	0.1257	-0.0204	0.0318	0.0650	0.0867	-0.1738
	FITR-Ramp	0.0628	0.0383	0.0776	-0.0364	0.0255	0.0268	0.0709	-0.1079
CGI	SepL	0.3579	0.0194	0.2745	-0.0429	0.0862	0.0136	0.1548	-0.2805
	FITR-IntL	0.5473	0.3066	0.5942	-0.1625	0.0884	0.3279	0.4205	-0.6595
	FITR-Ramp	0.1958	0.0622	0.2361	-0.1317	0.0670	0.1076	0.2195	-0.2043
SAS	SepL	0.2068	0.0136	0.1471	-0.1018	-0.0514	0.2019	0.1570	0.0285
	FITR-IntL	0.1410	0.0842	0.1037	-0.0296	-0.0133	0.1183	0.0888	-0.0808
	FITR-Ramp	0.0816	0.0284	0.0646	-0.0228	-0.0104	0.0645	0.0565	-0.0496

Table 4.11: The coefficients of the estimated ITRs by SepL, FITR-IntL and FITR-Ramp when the linear kernel is used.

4.7.2 Proof for Section 4.3

In this section, we provide the proof for the theoretical results in Section 4.3.

4.7.2.1 Details for the Agreement Rate Comparison

To give a bound for the agreement rate of SepL, we need to bound its decision accuracy. We first provide another assumption about the rewards and then present the general accuracy bound for $\mu_{1n} \geq 0$ in FITR-Ramp.

Assumption 22. Suppose the conditional expectation of rewards satisfies $\sum_{a\in\mathcal{A}} \mathbb{E}(R_1|a, \boldsymbol{x}) \geq c_r$ for some constant $c_r > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.

Corollary 4.7.1. *Under Assumptions 15-21 and 22, the misclassification rate satisfies*

$$\mathbb{P}(\widehat{f}_{1n} f_1^* < 0) \lesssim [\delta_{1n}(\tau) + \mu_{1n}]^{\frac{\alpha}{2-\alpha}} \tag{4.16}$$

with probability greater than or equal to $1 - 2e^{\tau}$.

Remark. If $\widetilde{f}_2$ is the sign of the decision function learnt by SepL as described in Section 4.2 with the same dataset of size n, we can directly use Corollary 4.7.1 with $\mu_{2n} = 0$ to find that $\widetilde{\delta}_{2n}(\tau) = (\delta_{2n}^{(0)}(\tau))^{\frac{\alpha}{2-\alpha}}$, where

$$\delta_{2n}^{(0)}(\tau) := \lambda_{2n}^{-\frac{1}{2}} n^{-\frac{1}{2}} \left[\sqrt{\tau} + \sigma_{2n}^{(1-\nu/2)(1+\delta)d}\right] + \lambda_{2n}\sigma_{2n}^{d} + (2d)^{qd/2}\sigma_{2n}^{-qd}$$

with probability greater than or equal to $1 - e^{\tau}$.

4.7.2.2 Proof for Lemma 4.3.1

Proof. First note that

$$\mathcal{R}(\widehat{f}_{1n}) - \mathcal{R}(f_1^*) \leq \mathcal{R}(\widehat{f}_{1n}) - \mathcal{R}(f_1^*) + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 \tag{4.17}$$

$$\leq \left\{\left[\mathcal{R}(\widehat{f}_{1n}) + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2\right] - \inf_{f_1 \in \mathcal{H}}\left[\mathcal{R}(f_1) + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2\right]\right\} \tag{4.18}$$

$$+ \left\{\inf_{f_1 \in \mathcal{H}}\left[\mathcal{R}(f_1) + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2\right] - \mathcal{R}(f_1^*)\right\}. \tag{4.19}$$

Define $f_1^{\dagger}$ as the minimizer of $\mathcal{R}(f_1) + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2$ in $\mathcal{H}$. We will bound the two terms on the right-hand side separately.

To bound (4.19), we follow the construction in the proof of Theorem of 2.7 in Steinwart and Scovel (2007). When $\mathcal{X}$ is the closed unit ball, on $\acute{\mathcal{X}} := 3\mathcal{X}$ define

$$\acute{\eta}(\boldsymbol{x}) = \begin{cases} \eta(\boldsymbol{x}), & \text{if } \|\boldsymbol{x}\|_2 \leq 1, \\ \eta(\boldsymbol{x}/\|\boldsymbol{x}\|_2), & \text{otherwise.} \end{cases}$$

Besides, let $\acute{\mathcal{X}}_{-1} := \{x \in \acute{\mathcal{X}} : \acute{\eta}(\boldsymbol{x}) < \frac{1}{2}\}$ and $\acute{\mathcal{X}}_1 := \{x \in \acute{\mathcal{X}} : \acute{\eta}(\boldsymbol{x}) > \frac{1}{2}\}$. Fix a measurable $\acute{f}_1 : \acute{\mathcal{X}} \mapsto [-1, 1]$ that satisfies $\acute{f}_1 = 1$ on $\acute{\mathcal{X}}_1$, $\acute{f}_1 = -1$ on $\acute{\mathcal{X}}_{-1}$ and $\acute{f}_1 = 0$ otherwise. The linear

operator $V_{\sigma_{1n}} : L_2(\mathbb{R}^d) \mapsto \mathcal{H}_{\sigma_{1n}}(\mathbb{R}^d)$ defined by

$$V_{\sigma_{1n}} g(\boldsymbol{x}) = \frac{(2\sigma_{1n})^{d/2}}{\pi^{d/4}} \int_{\mathbb{R}^d} e^{-2\sigma_{1n}^2 \|\boldsymbol{x}-\boldsymbol{y}\|_2^2} g(\boldsymbol{y}) d\boldsymbol{y}, \qquad g \in L_2(\mathbb{R}^d), \boldsymbol{x} \in \mathbb{R}^d,$$

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

$$\begin{aligned} &\inf_{f_1 \in \mathcal{H}} \left[\mathcal{R}(f_1) + \lambda_{1n} \|f_1\|_{\mathcal{H}}^2\right] - \mathcal{R}(f_1^*) \\ \leq &\inf_{g \in L_2(\mathbb{R}^d)} \left[\mathbb{E}(\ell_1 \circ V_{\sigma_{1n}} g - \ell_1 \circ f_1^*) + \mathbb{E}(\ell_2 \circ V_{\sigma_{1n}} g - \ell_2 \circ f_1^*) + \lambda_{1n} \|g\|_{L_2(\mathbb{R}^d)}^2\right]. \end{aligned}$$

Now take a specific $g := (\frac{\sigma_{1n}^2}{\pi})^{d/4} \acute{f}_1$ and we obtain

$$\|g\|_{L_2(\mathbb{R}^d)} \leq \left(\frac{81\sigma_{1n}^2}{\pi}\right)^{d/4} \theta(d), \tag{4.20}$$

where $\theta(d)$ denotes the volume of $\mathcal{X}$. Since $1 \leq \acute{f}_1 \leq 1$, it can be easily seen that $1 \leq V_{\sigma_{1n}} g \leq 1$. Note that $|\phi'(t)| = \left|-\frac{e^{-t}}{1+e^{-t}}\right| \leq 1$, so ℓ_1 is Lipschitz continuous with respect to f_1 with Lipschitz constant r/p_0. It has been shown in Steinwart and Scovel (2007) that

$$|V_{\sigma_{1n}} g(\boldsymbol{x}) - f_1^*(\boldsymbol{x})| \leq 8e^{-\sigma_{1n}^2 \omega_{\boldsymbol{x}}^2/2d}.$$

Therefore, Assumption 20 for $t = 2d/\sigma_{1n}^2$ yield

$$\mathbb{E}(\ell_1 \circ V_{\sigma_{1n}} g - \ell_1 \circ f_1^*) \lesssim \mathbb{E}\,|V_{\sigma_{1n}} g - f_1^*| \lesssim \mathbb{E} e^{-\sigma_{1n}^2 \omega_{\boldsymbol{x}}^2/2d} \lesssim (2d)^{qd/2} \sigma_{1n}^{-qd}. \tag{4.21}$$

Since ℓ_2 is Lipschitz continuous with respect to f_1 with Lipschitz constant $\frac{\mu_{1n}\Omega_{12}}{\kappa_{1n}} \|\widetilde{f}_2\|_\infty$,

$$\mathbb{E}(\ell_2 \circ V_{\sigma_{1n}} g - \ell_2 \circ f_1^*) \lesssim \frac{\mu_{1n}}{\kappa_{1n}} \mathbb{E}\,|V_{\sigma_{1n}} g - f_1^*| \lesssim \frac{\mu_{1n}}{\kappa_{1n}} (2d)^{qd/2} \sigma_{1n}^{-qd} \tag{4.22}$$

when $\|\widetilde{f}_2\|_\infty = 1$. Combining (4.20), (4.21) and (4.22), we can bound the approximation error (4.19) as

$$\inf_{f_1 \in \mathcal{H}} \left[\mathcal{R}(f_1) + \lambda_{1n} \|f_1\|_{\mathcal{H}}^2\right] - \mathcal{R}(f_1^*) \lesssim \lambda_{1n} \sigma_{1n}^d + \left(1 + \frac{\mu_{1n}}{\kappa_{1n}}\right) (2d)^{qd/2} \sigma_{1n}^{-qd}. \tag{4.23}$$

To bound (4.18), we will use the Talagrand's inequality quoted as follows (Steinwart and Scovel, 2007, Theorem 5.6).

Theorem 4.7.2. *Let $\mathcal{H}$ be a set of bounded measurable functions from $\boldsymbol{Z}$ to $\mathbb{R}$ which is separable with respect to $\|\cdot\|_\infty$ and satisfies $\mathbb{E}h = 0$ for all $h \in \mathcal{H}$. Furthermore, let $B > 0$ and $b \geq 0$ be constants with $\|h\|_\infty \leq B$ and $\mathbb{E}h^2 \leq b$ for all $h \in \mathcal{H}$. Then for all $\tau \geq 1$ and all $n \geq 1$ we have*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \mathbb{P}_n h > 3\mathbb{E}\sup_{h\in\mathcal{H}} \mathbb{P}_n h + \sqrt{\frac{2\tau b}{n}} + \frac{B\tau}{n}\right) \leq e^{-\tau}.$$

We first obtain a bound for $\|\widehat{f}_{1n}\|_{\mathcal{H}}^2$. Since $\mathbb{P}_n(\ell_1 \circ \widehat{f}_{1n} + \ell_2 \circ \widehat{f}_{1n}) + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 \leq \mathbb{P}_n(\ell_1 \circ f_1 + \ell_2 \circ f_1) + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2$ for any $f \in \mathcal{H}$, when taking $f = 0$ we have

$$\lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 \leq \mathbb{P}_n(\ell_1 \circ f_1 + \ell_2 \circ f_1) \leq \frac{r}{p_0} + \mu_{1n}\Omega_{12}.$$

Since $r/p_0 + \mu_{1n}\Omega_{12} \simeq M$, $\|\widehat{f}_{1n}\|_{\mathcal{H}}$ is bounded by $\sqrt{M/\lambda_{1n}}$, where $M := r/p_0$. To this end, it suffices to consider the ball of radius $\sqrt{M/\lambda_{1n}}$. Therefore, the function class that we consider here is

$$\mathcal{G} := \left\{\ell_1 \circ f_1 + \ell_2 \circ f_1 + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2 - \left[\ell_1 \circ f_1^\dagger + \ell_2 \circ f_1^\dagger + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right] : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\},$$

where $B_{\mathcal{H}}(r)$ is the ball in $\mathcal{H}$ of radius r. Since ℓ_1 is Lipschitz continuous with respect to f_1 and $\|f\|_\infty \leq \|f\|_{\mathcal{H}}$ for any $g \in \mathcal{G}$,

$$\begin{aligned}
|g| \leq& \left|\ell_1 \circ f_1 - \ell_1 \circ f_1^\dagger\right| + \left|\ell_2 \circ f_1 - \ell_2 \circ f_1^\dagger\right| + \lambda_{1n}\left|\|f_1\|_{\mathcal{H}}^2 - \|f_1^\dagger\|_{\mathcal{H}}^2\right| \\
\leq& M\left|f_1 - f_1^\dagger\right| + \mu_{1n}\Omega_{12} + M \\
\leq& 2M\sqrt{M/\lambda_{1n}} + \mu_{1n}\Omega_{12} + M.
\end{aligned}$$

Hence with $B := 2M\sqrt{M/\lambda_{1n}} + \mu_{1n}\Omega_{12} + M \simeq \lambda_{1n}^{-1/2}$, we have $\|g\|_\infty \leq B$.

Define the modulus of continuity of $\mathcal{G}$ by

$$\omega_n(\mathcal{G}, \epsilon) := \mathbb{E}\left(\sup_{g\in\mathcal{G}, \mathbb{E}g^2 \leq \epsilon} |\mathbb{E}g - \mathbb{P}_n g|\right), \quad \epsilon > 0,$$

where the supremum is measurable by the separability assumption on $\mathcal{G}$. Define the function class

$$\mathcal{E} := \{\mathbb{E}g - g : g \in \mathcal{G}\}, \tag{4.24}$$

then we have $\omega_n(\mathcal{G}, 4B^2) \geq \mathbb{E}\sup_{h\in\mathcal{E}} \mathbb{P}_n h$ since $|\mathbb{E}g - g| \leq 2B$. By Theorem 4.7.2 we obtain

$$\mathbb{P}\left(\sup_{h\in\mathcal{E}} \mathbb{P}_n h > 3\omega_n(\mathcal{G}, 4B^2) + \sqrt{\frac{2\tau b}{n}} + \frac{B\tau}{n}\right) \leq e^{-\tau}. \tag{4.25}$$

Let $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variable. Then the local Rademacher average of $\mathcal{F}$ is defined by

$$\mathrm{Rad}(\mathcal{G}, n, \epsilon) := \mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{g\in\mathcal{G}, \mathbb{E}g^2\leq\epsilon} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{z}_i)\right|.$$

It has been shown that

$$\omega_n(\mathcal{G}, \epsilon) \leq 2\,\mathrm{Rad}(\mathcal{G}, n, \epsilon), \quad \epsilon > 0$$

by symmetrization (Vaart and Wellner, 1996). Since

$$\mathrm{Rad}(\mathcal{G}, n, \epsilon) = B\,\mathrm{Rad}(B^{-1}\mathcal{G}, n, B^{-2}\epsilon)$$

for any $a > 0$ by equation (37) of Steinwart and Scovel (2007), we only need to obtain a bound for $\mathrm{Rad}(B^{-1}\mathcal{G}, n, B^{-2}\epsilon)$. To this end, we will use Proposition 5.5 of Steinwart and Scovel (2007) to bound the local Rademacher average, quoted as follows.

Theorem 4.7.3. *Let $\mathcal{F}$ be a class of measurable functions from $\boldsymbol{Z}$ to $[-1, 1]$ which is separable with respect to $\|\cdot\|_\infty$. Assume there are constants $a > 0$ and $0 < p < 2$ with*

$$\sup_{\mathbb{P}_n} \log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n)) \leq a\epsilon^{-p}$$

for all $\epsilon > 0$. Then there exists a constant $c_p > 0$ depending only on p such that for all $n \geq 1$ and all $\epsilon > 0$ we have

$$\mathrm{Rad}(\mathcal{F}, n, \epsilon) \leq c_p \max\left\{\epsilon^{\frac{1}{2}-\frac{p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}}, \left(\frac{a}{n}\right)^{\frac{2}{2+p}}\right\}.$$

Now we need to find some constants $a > 0$ and $0 < p < 2$ such that

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

for some $a \geq 1, 0 < p < 2$ and for all $\epsilon > 0$. To this end, note that

$$\begin{aligned}
&\log N(\epsilon, B^{-1}\mathcal{G}, L_2(\mathbb{P}_n)) \\
=&\log N\left(B^{-1}\left\{\ell_1 \circ f_1 + \ell_2 \circ f_1 + \lambda_{1n}\|f_1\|_{\mathcal{H}}^2 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}, \epsilon, L_2(\mathbb{P}_n)\right) \\
\leq&\log N\left(B^{-1}\left\{\ell_1 \circ f_1 + \ell_2 \circ f_1 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}, \epsilon, L_2(\mathbb{P}_n)\right) \\
&+ \log N\left(B^{-1}\left\{\lambda_{1n}\|f_1\|_{\mathcal{H}}^2 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}, \epsilon, L_2(\mathbb{P}_n)\right)
\end{aligned}$$

by the subadditivity of the entropy. For the first term on the right-hand side, for any $f_1, f_1' \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})$, let $u := B^{-1}(\ell_1 + \ell_2) \circ f_1$ and $u' := B^{-1}(\ell_1 + \ell_2) \circ f_1'$. Since ℓ_1 and ℓ_2 are Lipschitz continuous with respect to f_1,

$$\left\|u - u'\right\|_{L_2(\mathbb{P}_n)} \leq B^{-1}\left(M + \frac{\mu_{1n}\Omega_{12}}{\kappa_{1n}}\right)\|f - f'\|_{L_2(\mathbb{P}_n)}.$$

With $u, u' \in B^{-1}\left\{\ell_1 \circ f_1 + \ell_2 \circ f_1 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}$,

$$\begin{aligned}
&\log N\left(B^{-1}\left\{\ell_1 \circ f_1 + \ell_2 \circ f_1 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}, \epsilon, L_2(\mathbb{P}_n)\right) \\
\leq&\log N\left(B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}}), \frac{B\epsilon}{M + \mu_{1n}\Omega_{12}\kappa_{1n}^{-1}}, L_2(\mathbb{P}_n)\right) \\
\leq&\log N\left(B_{\mathcal{H}}, \gamma_n\epsilon, L_2(\mathbb{P}_n)\right),
\end{aligned}$$

where

$$\gamma_n := \frac{B}{\sqrt{M/\lambda_{1n}}(M + \mu_{1n}\Omega_{12}\kappa_{1n}^{-1})} \simeq \frac{1}{1 + \mu_{1n}\kappa_{1n}^{-1}}$$

For the second term on the right-hand side, it follows that

$$\log N\left(B^{-1}\left\{\lambda_{1n}\|f_1\|_{\mathcal{H}}^2 : f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})\right\}, \epsilon, L_2(\mathbb{P}_n)\right) \leq \log\frac{M}{B\epsilon}$$

since $\lambda_{1n}\|f_1\|_{\mathcal{H}}^2 \leq M$ for all $f \in B_{\mathcal{H}}(\sqrt{M/\lambda_{1n}})$. Therefore, we can conclude that

$$\log N(\epsilon, B^{-1}\mathcal{G}, L_2(\mathbb{P}_n)) \leq \log N\left(B_{\mathcal{H}}, \gamma_n\epsilon, L_2(\mathbb{P}_n)\right) + \log\frac{M}{B\epsilon}.$$

Theorem 2.1 of Steinwart and Scovel (2007) then yields that

$$\sup_{\mathbb{P}_n} \log N(\epsilon, B^{-1}\mathcal{G}, L_2(\mathbb{P}_n)) \lesssim \sigma_{1n}^{(1-\nu/2)(1+\delta)d}(\gamma_n\epsilon)^{-\nu},$$

where $\sigma_{1n} > 0$ is the parameter of the Gaussian kernel associated with $\mathcal{H}$, and $0 < \nu \leq 2, \delta > 0, \epsilon > 0$. Therefore, we have $a = \sigma_{1n}^{(1-\nu/2)(1+\delta)d}\gamma_n^{-\nu}$, $p = \nu$ and

$$\mathrm{Rad}(\mathcal{G}, n, \epsilon) \leq c_p \max\left\{B^{\frac{p}{2}}\epsilon^{\frac{1}{2}-\frac{p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}}, B\left(\frac{a}{n}\right)^{\frac{2}{2+p}}\right\}.$$

With $\epsilon = 4B^2$, we can bound the modulus of continuity as

$$\omega_n(\mathcal{G}, \epsilon) \leq 2\,\mathrm{Rad}(\mathcal{G}, n, \epsilon) \lesssim Ba^{\frac{1}{2}}n^{-\frac{1}{2}} \simeq \lambda_{1n}^{-\frac{1}{2}}\sigma_{1n}^{(1-\nu/2)(1+\delta)d}\gamma_n^{-\nu}n^{-\frac{1}{2}}. \tag{4.26}$$

The definition of $\widehat{f}_{1n}$ yields that

$$\mathbb{P}_n\left\{\ell_1 \circ \widehat{f}_{1n} + \ell_2 \circ \widehat{f}_{1n} + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 - \left[\ell_1 \circ f_1^\dagger + \ell_2 \circ f_1^\dagger + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right]\right\} \leq 0.$$

Therefore,

$$\begin{aligned}
&\left[\mathcal{R}(\widehat{f}_{1n}) + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2\right] - \left[\mathcal{R}(f_1^\dagger) + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right] \\
=&\mathbb{E}\left\{\ell_1 \circ \widehat{f}_{1n} + \ell_2 \circ \widehat{f}_{1n} + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 - \left[\ell_1 \circ f_1^\dagger + \ell_2 \circ f_1^\dagger + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right]\right\} \\
\leq&(\mathbb{E} - \mathbb{P}_n)\left\{\ell_1 \circ \widehat{f}_{1n} + \ell_2 \circ \widehat{f}_{1n} + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2 - \left[\ell_1 \circ f_1^\dagger + \ell_2 \circ f_1^\dagger + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right]\right\} \\
\leq&\sup_{h\in\mathcal{E}} \mathbb{P}_n h,
\end{aligned}$$

where $\mathcal{E}$ is defined in (4.24). Note that $\|h\|_\infty \le 2B \lesssim \lambda_{1n}^{-\frac{1}{2}}$ and $\mathbb{E}h^2 \le 4B^2 \lesssim \lambda_{1n}^{-1}$ for all $h \in \mathcal{E}$. Plugging (4.26) into (4.25) and we have

$$\begin{aligned}
&\left[\mathcal{R}(\widehat{f}_{1n}) + \lambda_{1n}\|\widehat{f}_{1n}\|_{\mathcal{H}}^2\right] - \left[\mathcal{R}(f_1^\dagger) + \lambda_{1n}\|f_1^\dagger\|_{\mathcal{H}}^2\right] \le \sup_{h\in\mathcal{E}} \mathbb{P}_n h \\
\lesssim &\lambda_{1n}^{-\frac{1}{2}} \sigma_{1n}^{(1-\nu/2)(1+\delta)d} \gamma_n^{-\nu} n^{-\frac{1}{2}} + \sqrt{\frac{2\tau\lambda_{1n}^{-1}}{n}} + \frac{\tau\lambda_{1n}^{-\frac{1}{2}}}{n} \\
\lesssim &\lambda_{1n}^{-\frac{1}{2}} n^{-\frac{1}{2}} \left[\sqrt{\tau} + \sigma_{1n}^{(1-\nu/2)(1+\delta)d} \gamma_n^{-\nu}\right]
\end{aligned} \tag{4.27}$$

with probability greater than or equal to $1 - e^{-\tau}$ for any $\tau \ge 1$.

Finally, plug the upper bounds (4.27) and (4.23) into (4.18) and (4.19) and we get the results.

□

4.7.2.3 Proof for Theorem 4.3.2

Proof. Define $\widehat{U}_{1n} := \mathbb{E}I(\widehat{f}_{1n} f_1^* < 0)$ and $\Delta\mathcal{V}_1(f_1) := \mathcal{V}_1(f_1^*) - \mathcal{V}_1(f_1)$ for simplicity.

To utilize existing results in general classification problems, we can rewrite our loss functions with a change of measure. Let $h(\cdot)$ be the probability distribution function of the covariates $\boldsymbol{X}$. Then the expectation of ℓ_1 can be written as

$$\mathbb{E}\ell_1 \circ f_1 = \mathbb{E}\left[\frac{R_1}{\pi(A;\boldsymbol{X})}\phi(Af_1(\boldsymbol{X}))\right] = \int_{\mathcal{X}_+} \sum_{a\in\mathcal{A}} \frac{\mathbb{E}(R_1|a,\boldsymbol{x})}{\pi(a;\boldsymbol{x})}\phi(af_1(\boldsymbol{x}))\pi(a;\boldsymbol{x})h(\boldsymbol{x})d\boldsymbol{x},$$

where $\mathcal{X}_+ := \mathbb{E}(R_1|1,\boldsymbol{x}) + \mathbb{E}(R_1|-1,\boldsymbol{x}) > 0$. Now define $g(\boldsymbol{x}) := \sum_{a\in\mathcal{A}} \mathbb{E}(R_1|a,\boldsymbol{x}), C_{R_1} := \int g(\boldsymbol{x})h(\boldsymbol{x})d\boldsymbol{x}$. Let $h'(\boldsymbol{x}) := g(\boldsymbol{x})h(\boldsymbol{x})/C_{R_1}$ so that h' is a new probability distribution function. Let

$$\pi'(a;\boldsymbol{x}) := \begin{cases} \frac{\mathbb{E}(R_1|a,\boldsymbol{x})}{g(\boldsymbol{x})}, & \text{if } g(\boldsymbol{x}) > 0 \\ \frac{1}{2}, & \text{otherwise,} \end{cases}$$

so that $\pi' \in [0,1]$ by Assumption 18 and can be regarded as a new policy for sampling the treatments. Then we obtain

$$\begin{aligned}\mathbb{E}\ell_1 \circ f_1 =& C_{R_1} \int_{\mathcal{X}_+} \sum_{a\in\mathcal{A}} \phi(af_1(\boldsymbol{x})) \frac{\mathbb{E}(R_1|a,\boldsymbol{x})}{g(\boldsymbol{x})} \frac{g(\boldsymbol{x})h(\boldsymbol{x})}{C_{R_1}} d\boldsymbol{x} \\ =& C_{R_1} \int \sum_{a\in\mathcal{A}} \phi(af_1(\boldsymbol{x}))\pi'(a;\boldsymbol{x})h'(\boldsymbol{x})d\boldsymbol{x}.\end{aligned}$$

Denote $\mathbb{E}'$ as the expectation corresponding to the distributions h' and π', so we get

$$\mathbb{E}\ell_1 \circ f_1 = C_{R_1}\mathbb{E}'\phi(Af_1).$$

Conversely, for the 0-1 loss, the difference between value functions can be written as

$$\begin{aligned}\Delta\mathcal{V}_1(f_1) =& \mathcal{V}_1(f_1^*) - \mathcal{V}_1(f_1) \\ =& \mathbb{E}\left[\frac{R_1}{\pi(A;\boldsymbol{X})} I(Af_1(\boldsymbol{X}) < 0)\right] - \mathbb{E}\left[\frac{R_1}{\pi(A;\boldsymbol{X})} I(Af_1^*(\boldsymbol{X}) < 0)\right] \\ =& C_{R_1}\left[\mathbb{E}'I(Af_1 < 0) - \mathbb{E}'I(Af_1^* < 0)\right].\end{aligned}$$

By Theorem 3 of Bartlett et al. (2006)

$$\begin{aligned}&\mathbb{E}\ell_1 \circ f_1 - \mathbb{E}\ell_1 \circ f_1^* = C_{R_1}\mathbb{E}'\phi(Af_1) - C_{R_1}\mathbb{E}'\phi(Af_1^*) \\ \geq& C_{R_1} c\left[\mathbb{E}'I(Af_1 < 0) - \mathbb{E}'I(Af_1^* < 0)\right]^\alpha \rho\left(\frac{\left[\mathbb{E}'I(Af_1 < 0) - \mathbb{E}'I(Af_1^* < 0)\right]^{1-\alpha}}{2c}\right) \\ =& C_{R_1} c\left[\frac{1}{C_{R_1}}\Delta\mathcal{V}_1(f_1)\right]^\alpha \rho\left[\frac{1}{2c}\left(\frac{1}{C_{R_1}}\Delta\mathcal{V}_1(f_1)\right)^{1-\alpha}\right] \\ \simeq& \left[\Delta\mathcal{V}_1(f_1)\right]^\alpha \rho\left[\left(\Delta\mathcal{V}_1(f_1)\right)^{1-\alpha}\right]\end{aligned} \tag{4.28}$$

for any f_1, where $c > 0$ is a constant and $\rho(t) = \frac{1}{2}\left[(1+t)\log(1+t) + (1-t)\log(1-t)\right]$ for the logistic loss ϕ.

For ℓ_2 with the ramp loss, note that $\psi_{\kappa_n}(f_1^* f_2^*) = I(f_1^* f_2^* < 0)$ since $f_1^* f_2^*$ takes values only in $\{-1, 1\}$ when $\kappa_n \leq 1$. Besides, $\psi_{\kappa_n}(f_1 f_2^*) \geq I(f_1 f_2^* < 0)$ for any f_1 by the definition of the ramp loss ψ. Hence we obtain the relationship between the excess risks under the ramp loss and the 0-1

loss as

$$\mathbb{E}\psi_{\kappa_n}(f_1 f_2^*) - \mathbb{E}\psi_{\kappa_n}(f_1^* f_2^*) \geq \mathbb{E}I(f_1 f_2^* < 0) - \mathbb{E}I(f_1^* f_2^* < 0). \tag{4.29}$$

Since $\widetilde{f}_2$ and f_2^* are binary decision functions,

$$|\psi_{\kappa_n}(f_1\widetilde{f}_2) - \psi_{\kappa_n}(f_1 f_2^*)| = I(\widetilde{f}_2 f_2^* < 0)|\psi_{\kappa_n}(f_1\widetilde{f}_2) - \psi_{\kappa_n}(f_1 f_2^*)| \leq I(\widetilde{f}_2 f_2^* < 0) \leq \widetilde{\delta}_{2N}(\tau) \tag{4.30}$$

with probability greater than or equal to $1 - e^\tau$ for any f_1 by Assumption 21. The first inequality comes from the fact that one of $\psi_{\kappa_n}(f_1\widetilde{f}_2)$ and $\psi_{\kappa_n}(f_1 f_2^*)$ must be zero and the other is bounded by one. Therefore, we have

$$\begin{aligned}
\mathbb{E}\ell_2 \circ f_1 - \mathbb{E}\ell_2 \circ f_1^* =& \mu_{1n}\Omega_{12}\left[\mathbb{E}\psi_{\kappa_n}(f_1\widetilde{f}_2) - \mathbb{E}\psi_{\kappa_n}(f_1^*\widetilde{f}_2)\right] \\
\geq& \mu_{1n}\Omega_{12}\left[\mathbb{E}\psi_{\kappa_n}(f_1 f_2^*) - \mathbb{E}\psi_{\kappa_n}(f_1^* f_2^*) - 2\widetilde{\delta}_{2N}(\tau)\right] \\
\geq& \mu_{1n}\Omega_{12}\left[\mathbb{E}I(f_1 f_2^* < 0) - \mathbb{E}I(f_1^* f_2^* < 0) - 2\widetilde{\delta}_{2N}(\tau)\right],
\end{aligned} \tag{4.31}$$

where the first inequality comes from (4.30) and the second inequality comes from (4.29).

Finally, combine Lemma 4.3.1 with (4.28), (4.31) and we get that

$$\begin{aligned}
&\left[\Delta\mathcal{V}_1(\widehat{f}_{1n})\right]^\alpha \rho\left[\left(\Delta\mathcal{V}_1(\widehat{f}_{1n})\right)^{1-\alpha}\right] + \mu_{1n}[\mathbb{E}I(\widehat{f}_{1n} f_2^* < 0) - \mathbb{E}I(f_1^* f_2^* < 0) - \widetilde{\delta}_{2N}(\tau)] \\
\lesssim& \mathbb{E}\ell_1 \circ \widehat{f}_{1n} - \mathbb{E}\ell_1 \circ f_1^* + \mathbb{E}\ell_2 \circ \widehat{f}_{1n} - \mathbb{E}\ell_2 \circ f_1^* \lesssim \delta_{1n}(\tau)
\end{aligned}$$

with probability greater than or equal to $1 - 2e^\tau$, that is,

$$\left[\Delta\mathcal{V}_1(\widehat{f}_{1n})\right]^\alpha \rho\left[\left(\Delta\mathcal{V}_1(\widehat{f}_{1n})\right)^{1-\alpha}\right] + \mu_{1n}[\mathbb{E}I(\widehat{f}_{1n} f_2^* < 0) - \mathbb{E}I(f_1^* f_2^* < 0)] \lesssim \delta_{1n}(\tau) + \mu_{1n}\widetilde{\delta}_{2N}(\tau). \tag{4.32}$$

By Taylor's expansion, it is easy to see that $\rho(t) \simeq t^2$. Since $\mathbb{E}I(f_1^* f_2^* < 0) \leq 1$, we can conclude that

$$[\Delta\mathcal{V}_1(f_1)]^{2-\alpha} \lesssim \delta_{1n}(\tau) + \mu_{1n}\widetilde{\delta}_{2N}(\tau) + \mu_{1n}. \tag{4.33}$$

and thus

$$\Delta\mathcal{V}_1(f_1) \lesssim (\delta_{1n}(\tau) + \mu_{1n})^{\frac{1}{2-\alpha}}$$

with probability greater than or equal to $1 - 2e^{\tau}$. □

4.7.2.4 Proof of Theorem 4.3.3

Proof. The results can be easily obtained since

$$\mu_{1n}[\mathbb{E}I(\widehat{f}_{1n}f_2^* < 0) - \mathbb{E}I(f_1^*f_2^* < 0)] \lesssim \delta_{1n}(\tau) + \mu_{1n}\widetilde{\delta}_{2N}(\tau).$$

according to (4.32). □

4.7.2.5 Proof of Corollary 4.7.1

Proof. According to Lemma 5 and (9) of Bartlett et al. (2006),

$$\mathbb{E}I(f_1f_1^* < 0) \leq c\left[\mathbb{E}(I(f_1f_1^* < 0)\left|2\eta(\boldsymbol{X}) - 1\right|)\right]^{\alpha}$$

where c is some constant. By Assumption 22, we have

$$2\eta(\boldsymbol{X}) - 1 := \frac{\mathbb{E}(R_1|1,\boldsymbol{X}) - \mathbb{E}(R_1|-1,\boldsymbol{X})}{\mathbb{E}(R_1|1,\boldsymbol{X}) + \mathbb{E}(R_1|-1,\boldsymbol{X})} \leq \frac{1}{c_r}\left|\mathbb{E}(R_1|1,\boldsymbol{X}) - \mathbb{E}(R_1|-1,\boldsymbol{X})\right|.$$

Since

$$\begin{aligned}\Delta\mathcal{V}_1(f_1) =& \mathbb{E}\left[\frac{R_1}{\pi(A;\boldsymbol{X})}I(Af_1(\boldsymbol{X}) < 0)\right] - \mathbb{E}\left[\frac{R_1}{\pi(A;\boldsymbol{X})}I(Af_1^*(\boldsymbol{X}) < 0)\right] \\ =& \mathbb{E}\left\{I(f_1f_1^* < 0)\left|\mathbb{E}(R_1|1,\boldsymbol{X}) - \mathbb{E}(R_1|-1,\boldsymbol{X})\right|\right\},\end{aligned}$$

we can bound the disagreement rate by the value difference, such that

$$\mathbb{E}I(f_1f_1^* < 0) \leq c\left[\frac{1}{c_r}\mathbb{E}(I(f_1f_1^* < 0)\left|\mathbb{E}(R_1|1,\boldsymbol{X}) - \mathbb{E}(R_1|-1,\boldsymbol{X})\right|)\right]^{\alpha} \lesssim \left[\Delta\mathcal{V}_1(f_1)\right]^{\alpha}, \tag{4.34}$$

that is, $\widehat{U}_{1n} \lesssim \left[\Delta\mathcal{V}_1(\widehat{f}_{1n})\right]^{\alpha}$. Then following (4.33) we have

$$\widehat{U}_{1n}^{\frac{2-\alpha}{\alpha}} \lesssim \delta_{1n}(\tau) + \mu_{1n}\widetilde{\delta}_{2N}(\tau) + \mu_{1n}$$

with probability greater than or equal to $1 - 2e^{\tau}$ if we take $\rho(t) \simeq t^2$. □

4.7.2.6 Proof of Theorem 4.3.4

Proof. The proof is similar to that for $K = 2$. We only highlight the main differences here.

To extend the results of Lemma 4.3.1 to $K \geq 3$, we can write the surrogate loss for the fusion penalty as $\ell_2 \circ f_1(\boldsymbol{Z}) := \mu_{1n} \sum_{j=2}^{K} \Omega_{1j} \psi_{\kappa_{1n}}[f_1(\boldsymbol{X}) \widetilde{f}_j(\boldsymbol{X})]$. Then ℓ_2 is Lipschitz continuous with respect to f_2 with Lipschitz constant $\frac{\mu_{1n} \sum_{j=2}^{K} \Omega_{1j}}{\kappa_{1n}}$. Hence we have

$$B := 2M\sqrt{M/\lambda_{1n}} + \mu_{1n} \sum_{j=2}^{K} \Omega_{1j} + M \simeq \lambda_{1n}^{-1/2}$$

and

$$\gamma_n := \frac{B}{\sqrt{M/\lambda_{1n}}(M + \mu_{1n} \sum_{j=2}^{K} \Omega_{1j} \kappa_{1n}^{-1})} \simeq \frac{1}{1 + \mu_{1n} \kappa_{1n}^{-1}},$$

which shows that the conclusion in Lemma 4.3.1 still holds.

Now inequality (4.31) should be written as

$$\begin{aligned} \mathbb{E}\ell_2 \circ f_1 - \mathbb{E}\ell_2 \circ f_1^* =& \mu_{1n} \sum_{j=2}^{K} \Omega_{1j} \left[\mathbb{E}\psi_{\kappa_n}(f_1 \widetilde{f}_j) - \mathbb{E}\psi_{\kappa_n}(f_1^* \widetilde{f}_j) \right] \\ \geq& \mu_{1n} \sum_{j=2}^{K} \Omega_{1j} \left[\mathbb{E}I(f_1 f_j^* < 0) - \mathbb{E}I(f_1^* f_j^* < 0) - 2\widetilde{\delta}_{jN_j}(\tau) \right], \end{aligned} \tag{4.35}$$

with probability greater than or equal to $1 - (K-1)e^{\tau}$ for any f_1 by Assumption 21. Therefore, inequality (4.32) is changed to

$$\begin{aligned} \left[\Delta\mathcal{V}_1(\widehat{f}_{1n}) \right]^{\alpha} \rho \left[\left(\Delta\mathcal{V}_1(\widehat{f}_{1n}) \right)^{1-\alpha} \right] + \mu_{1n} \sum_{j=2}^{K} [\mathbb{E}I(\widehat{f}_{1n} f_j^* < 0) - \mathbb{E}I(f_1^* f_j^* < 0)] & \\ \lesssim \delta_{1n}(\tau) + \mu_{1n} \sum_{j=2}^{K} \widetilde{\delta}_{jN_j}(\tau) & \end{aligned} \tag{4.36}$$

with probability greater than or equal to $1 - Ke^{\tau}$, and thus

$$[\Delta\mathcal{V}_1(f_1)]^{2-\alpha} \lesssim \delta_{1n}(\tau) + \mu_{1n} \sum_{j=2}^{K} \widetilde{\delta}_{jN_j}(\tau) + \mu_{1n}. \tag{4.37}$$

The inequality 4.13 follows from the assumption that $\sum_{j=2}^{K} \widetilde{\delta}_{jN_j}(\tau) = o(1)$. Similarly, from (4.36) we can conclude that

$$\mu_{1n}[\mathbb{E}I(\widehat{f}_{1n}f_k^* < 0) - \mathbb{E}I(f_1^* f_k^* < 0)] \lesssim \delta_{1n}(\tau) + \mu_{1n} \sum_{j=2}^{K} \widetilde{\delta}_{jN_j}(\tau)$$

for any $k = 2, \ldots, K$ with probability greater than or equal to $1 - Ke^{\tau}$. □

BIBLIOGRAPHY

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Bae, J. and Levental, S. (1995). Uniform CLT for Markov chains and its invariance principle: a martingale approach. *Journal of Theoretical Probability*, 8(3):549–570.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294.

Bastani, H., Bayati, M., and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349.

Bennett, A. and Kallus, N. (2021). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*.

Bennett, A., Kallus, N., Li, L., and Mousavi, A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 1999–2007. PMLR.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

Brown, C. H., Brincks, A., Huang, S., Perrino, T., Cruden, G., Pantin, H., Howe, G., Young, J. F., Beardslee, W., Montag, S., et al. (2018). Two-year impact of prevention programs on adolescent depression: An integrative data analysis approach. *Prevention Science*, 19(1):74–94.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer.

Busner, J. and Targum, S. D. (2007). The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)*, 4(7):28.

Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128.

Chambaz, A., Zheng, W., and van der Laan, M. J. (2017). Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Annals of Statistics*, 45(6):2537.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Chapple, A. G. and Thall, P. F. (2019). A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III. *Biometrics*, 75(2):371–381.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521.

Chen, H., Lu, W., and Song, R. (2020). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, pages 1–16.

Chen, H., Lu, W., and Song, R. (2021a). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719.

Chen, J., Fu, H., He, X., Kosorok, M. R., and Liu, Y. (2018). Estimating individualized treatment rules for ordinal treatments. *Biometrics*, 74(3):924–933.

Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209.

Chen, Y., Liu, Y., Zeng, D., and Wang, Y. (2019). *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*. R package version 1.0.

Chen, Y., Zeng, D., and Wang, Y. (2021b). Learning individualized treatment rules for multiple-domain latent outcomes. *Journal of the American Statistical Association*, 116(533):269–282.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Nekipelov, D., Semenova, V., and Syrgkanis, V. (2018b). Plug-in regularized estimation of high dimensional parameters in nonlinear semiparametric models. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chow, S.-C. (2014). Adaptive clinical trial design. *Annual Review of Medicine*, 65:405–415.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Claggett, B., Xie, M., and Tian, L. (2014). Meta-analysis with fixed, unknown, study-specific parameters. *Journal of the American Statistical Association*, 109(508):1660–1671.

Curran, P. J. and Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological methods*, 14(2):81.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR.

Ferguson, T. S. (2017). *A Course in Large Sample Theory*. Routledge.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Fletcher, R. (1987). *Practical methods of optimization*. John Wiley & Sons.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability*, pages 100–118.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. (2022). Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*.

Gao, D., Liu, Y., and Zeng, D. (2022). Non-asymptotic properties of individualized treatment rules from sequentially rule-adaptive trials. *Journal of Machine Learning Research*, 23(250):1–42.

Gao, D., Liu, Y., and Zeng, D. (2023a). Asymptotic inference for multi-stage stationary treatment policy with high dimensional features. *arXiv preprint arXiv:2301.12553*.

Gao, Y., Shi, C., and Song, R. (2023b). Deep spectral q-learning with application to mobile health. *arXiv preprint arXiv:2301.00927*.

Gunter, L., Zhu, J., and Murphy, S. (2007). Variable selection for optimal decision making. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 149–154. Springer.

Haidich, A.-B. (2010). Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. CRC Press.

Hu, F. and Rosenberger, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons.

Hu, J., Zhu, H., and Hu, F. (2015). A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, 110(509):357–367.

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2021). Personalized policy learning using longitudinal mobile health data. *Journal of the american statistical association*, 116(533):410–420.

Hu, Y., Kallus, N., and Mao, X. (2020). Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010. PMLR.

Huang, J., Ma, S., Li, H., and Zhang, C.-H. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics*, 39(4):2021.

Jeng, X. J., Lu, W., and Peng, H. (2018). High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12(1):2074.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167).

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

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., et al. (2011). The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

Laber, E. B., Wu, F., Munera, C., Lipkovich, I., Colucci, S., and Ripa, S. (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.

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

Lai, T. L., Lavori, P. W., Shih, M.-C. I., and Sikic, B. I. (2012). Clinical trial designs for testing biomarker-based personalized therapies. *Clinical Trials*, 9(2):141–154.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society. Series A*, 163(1):29–38.

Le Thi, H. A. and Pham Dinh, T. (2018). Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68.

Lei, H., Tewari, A., and Murphy, S. A. (2017). An actor-critic contextual bandit algorithm for personalized mobile health interventions. *arXiv preprint arXiv:1706.09090*.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, pages 661–670.

Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. (2022). Reinforcement learning in possibly nonstationary environments. *arXiv preprint arXiv:2203.01707*.

Li, S., Cai, T., and Li, H. (2021). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(1):149–173.

Liang, M., Choi, Y.-G., Ning, Y., Smith, M. A., and Zhao, Y.-Q. (2022). Estimation and inference on high-dimensional individualized treatment rule in observational data using split-and-pooled de-correlated score. *Journal of Machine Learning Research*, 23(262):1–65.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Liao, P., Klasnja, P., and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. (2022). Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387.

Lin, D.-Y. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018a). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018b). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, 37(26):3776–3788.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical methods in medical research*, 22(5):493–504.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706.

Luckett, D. J., Laber, E. B., Kim, S., and Kosorok, M. R. (2021). Estimation and optimization of composite outcomes. *Journal of Machine Learning Research*, 22:167–1.

Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713.

Miao, R., Qi, Z., and Zhang, X. (2022). Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *arXiv preprint arXiv:2209.10064*.

Minsker, S., Zhao, Y.-Q., and Cheng, G. (2016). Active clinical trials for personalized medicine. *Journal of the American Statistical Association*, 111(514):875–887.

Mo, W., Liu, Y., et al. (2022). Efficient learning of optimal individualized treatment rules for heteroscedastic or misspecified treatment-free effect models. *Journal of the Royal Statistical Society Series B*, 84(2):440–472.

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

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

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

Murphy, S. A. (2005b). A generalization error for Q-learning. *Journal of Machine Learning Research*, 6(Jul):1073–1097.

Murphy, S. A., Deng, Y., Laber, E. B., Maei, H. R., Sutton, R. S., and Witkiewitz, K. (2016). A batch, off-policy, actor-critic algorithm for optimizing the average reward. *arXiv preprint arXiv:1607.05047*.

Murphy, S. A., Oslin, D. W., Rush, A. J., and Zhu, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*, 32(2):257–262.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.

Nie, X., Brunskill, E., and Wager, S. (2021). Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*, 45(1):158–195.

Nishiyama, Y. (1997). Some central limit theorems for ℓ_{infty}-valued semimartingales and their applications. *Probability Theory and Related Fields*, 108(4):459–494.

Nishiyama, Y. et al. (2000). Weak convergence of some classes of martingales with jumps. *Annals of Probability*, 28(2):685–712.

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2):693–721.

Powell, M. J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162.

Precup, D. (2000a). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Precup, D. (2000b). *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Qi, Z., Liu, D., Fu, H., and Liu, Y. (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.

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

Qiu, X., Zeng, D., and Wang, Y. (in press). *Integrative Learning to Combine Individualized Treatment Rules from Multiple Randomized Trials*. Springer.

Rakhlin, A. and Sridharan, K. (2014). *Statistical learning and sequential prediction*. Book Draft.

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153.

Renfro, L. A., Mallick, H., An, M.-W., Sargent, D. J., and Mandrekar, S. J. (2016). Clinical trial designs incorporating predictive biomarkers. *Cancer Treatment Reviews*, 43:74–82.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*.

Riviere, M.-K., Yuan, Y., Jourdan, J.-H., Dubois, F., and Zohar, S. (2018). Phase I/II dose-finding design for molecularly targeted agent: plateau determination using adaptive randomization. *Statistical Methods in Medical Research*, 27(2):466–479.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., et al. (2004). Sequenced treatment alternatives to relieve depression (star*d): rationale and design. *Controlled clinical trials*, 25(1):119–142.

Shi, C., Fan, A., Song, R., and Lu, W. (2018). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of Statistics*, 46(3):925.

Shi, C., Lu, W., and Song, R. (2020a). Breaking the curse of nonregularity with subagging: inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21.

Shi, C., Luo, S., Le, Y., Zhu, H., and Song, R. (2022a). Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *Journal of the American Statistical Association*, pages 1–14.

Shi, C., Song, R., and Lu, W. (2016). Robust learning for optimal treatment decision with np-dimensionality. *Electronic journal of statistics*, 10:2894.

Shi, C., Uehara, M., Huang, J., and Jiang, N. (2022b). A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. (2021a). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. (2020b). Does the markov decision process fit the data: testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR.

Shi, C., Zhang, S., Lu, W., and Song, R. (2021b). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 84(3):765–793.

Shi, C., Zhu, J., Ye, S., Luo, S., Zhu, H., and Song, R. (2022c). Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12.

Song, R., Luo, S., Zeng, D., Zhang, H. H., Lu, W., and Li, Z. (2017). Semiparametric single-index model for estimating optimal individualized treatment strategy. *Electronic Journal of Statistics*, 11(1):364–384.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*. Omnipress.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607.

Sugiyama, M. (2015). *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. CRC Press.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Sun, Y. and Wang, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 116(533):421–432.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, pages 28–76.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

Thall, P. F. (2002). Ethical issues in oncology biostatistics. *Statistical methods in medical research*, 11(5):429–448.

Thall, P. F., Nguyen, H. Q., Braun, T. M., and Qazilbash, M. H. (2013). Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes. *Biometrics*, 69(3):673–682.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Thomas, P. S. (2015). *Safe reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

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

Tian, Y. and Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. (2022). Reward design for an online reinforcement learning algorithm supporting oral self-care. *arXiv preprint arXiv:2208.07406*.

Trivedi, M. H., McGrath, P. J., Fava, M., Parsey, R. V., Kurian, B. T., Phillips, M. L., Oquendo, M. A., Bruder, G., Pizzagalli, D., Toups, M., et al. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (embarc): Rationale and design. *Journal of psychiatric research*, 78:11–23.

Uehara, M., Shi, C., and Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*.

Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*. Springer.

Van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Annals of Statistics*, pages 1779–1801.

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

Wang, J., Qi, Z., and Shi, C. (2022). Blessing from experts: Super reinforcement learning in confounded environments. *arXiv preprint arXiv:2209.15448*.

Wang, Y., Fu, H., and Zeng, D. (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, C. J. C. H. (1989). Learning from delayed rewards. *PhD Thesis, University of Cambridge, England*.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694.

Xu, Y., Zhu, J., Shi, C., Luo, S., and Song, R. (2022). An instrumental variable approach to confounded off-policy evaluation. *arXiv preprint arXiv:2212.14468*.

Yang, Y. and Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, 30(1):100–121.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142.

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

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

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

Zhang, B. and Zhang, M. (2018a). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, 74(3):891–899.

Zhang, B. and Zhang, M. (2018b). Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates. *Annals of Applied Statistics*, 12(4):2335–2358.

Zhang, C., Chen, J., Fu, H., He, X., Zhao, Y.-Q., and Liu, Y. (2020a). Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica sinica*, 30:1857.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, J. and Bareinboim, E. (2016). Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab.

Zhang, K., Janson, L., and Murphy, S. (2020b). Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829.

Zhang, K., Janson, L., and Murphy, S. (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in neural information processing systems*, 34:7460–7471.

Zhang, K. W., Janson, L., and Murphy, S. A. (2022). Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *Annals of Statistics*, 35(3):1166–1182.

Zhao, Y., Laber, E. B., Ning, Y., Saha, S., and Sands, B. E. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, 20(1):1821–1843.

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

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

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

Zhou, D., Li, L., and Gu, Q. (2020). Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR.

Zhou, W., Zhu, R., and Qu, A. (2022). Estimating optimal infinite horizon dynamic treatment regimes via pt-learning. *Journal of the American Statistical Association*, pages 1–14.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.

Zhu, W., Zeng, D., and Song, R. (2019). Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.