Journal of Machine Learning Research 23 (2022) 1-42 Submitted 4/21; Revised 3/22; Published 8/22

# Non-asymptotic Properties of Individualized Treatment Rules from Sequentially Rule-Adaptive Trials

**Daiqi Gao** DQGAO@LIVE.UNC.EDU
*Department of Statistics and Operations Research*
*The University of North Carolina at Chapel Hill*
*Chapel Hill, NC 27599, USA*

**Yufeng Liu** YFLIU@EMAIL.UNC.EDU
*Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics*
*The University of North Carolina at Chapel Hill*
*Chapel Hill, NC 27599, USA*

**Donglin Zeng** DZENG@EMAIL.UNC.EDU
*Department of Biostatistics*
*The University of North Carolina at Chapel Hill*
*Chapel Hill, NC 27599, USA*

**Editor:** Ambuj Tewari

## Abstract

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. 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 machine learning 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.

**Keywords:** Contextual bandit, empirical process, martingale, outcome weighted learning, sequential decision making

## 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., 2018).

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 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 paper. 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 $\epsilon$-greedy methods assign the current optimal arm with a probability of $1-\epsilon$ or chooses from all arms randomly with a total probability of $\epsilon$ (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 $\epsilon$-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 $\epsilon$, where $\epsilon$ is a positive constant between 0 and 0.5. Furthermore, we let $\epsilon$ 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 $\epsilon$-greedy. Note that the $\epsilon$ defined here is one half of that in the definition of $\epsilon$-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 paper, 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 $\epsilon$-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 $\epsilon$-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 paper 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 paper is organized as follows. In Section 2, we describe our proposed algorithm that uses the OWL algorithm for learning ITRs over time. Section 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 4. In Section 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 6. The paper is concluded with some remarks in Section 7.

## 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 paper. 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 $f$.

### 2.1 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

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

where $\pi(A; \boldsymbol{X})$ is the probability of taking treatment $A$ given covariates $\boldsymbol{X}$ of a patient. After transforming (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., 2018).

### 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 $i$th 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 $i$th 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 $\epsilon$-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 $i$, $\epsilon$-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 $\text{benefit}_i$ 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 $i$th 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 $i$th ($i = 1, \ldots, n$) 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 3.1 and 3.2. Different choices of $p_i(\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ will be discussed in Section 4 and their performances will be compared in Section 5.

Following the scheme of OWL, we propose to minimize the $\phi$-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

$$\hat{f}_0 = \arg\min_{f \in \mathcal{F}} \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{R_j^{(0)}}{\pi_j(A_j^{(0)})} \phi(A_j^{(0)} f(\boldsymbol{X}_j^{(0)})) \tag{2}$$

using the pilot trial and update it to get

$$\hat{f}_i = \arg\min_{f \in \mathcal{F}} \frac{1}{n_0 + i} \left\{ \sum_{j=1}^{n_0} \frac{R_j^{(0)}}{\pi_j(A_j^{(0)})} \phi(A_j^{(0)} f(\boldsymbol{X}_j^{(0)})) + \sum_{j=1}^{i} \frac{R_j}{\pi_j(A_j; \boldsymbol{H}_{j-1}, \boldsymbol{X}_j)} \phi(A_j f(\boldsymbol{X}_j)) \right\} \tag{3}$$

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.

## 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.

### 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 paper 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

**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);

**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 (3).

**end**

Let $\hat{\mathcal{D}}_n = \text{sign}\{\hat{f}_n\}$ be the final estimated ITR.

$0 \leq E^{\text{sign}\{f^*\}}(R) - E^{\text{sign}\{f\}}(R) \leq E[g^f(Z)] - E[g^{f^*}(Z)]$. 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 $\epsilon_i$ 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*

$$\int_0^\infty \sqrt{1 + \log N_{[]}(\eta, \mathcal{F}, L_2(\mathbb{P}))} d\eta < \infty. \tag{4}$$

*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 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 $\delta > 0$,*

$$\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n) \leq \frac{C}{n_0 + n}\left[(J + \sqrt{\delta b})r\sqrt{n_0} + rb\delta + \frac{r^2 bJ}{\epsilon_n^2}\sqrt{\delta n \log^3 n}\right], \tag{5}$$

*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 $\delta, r, b, J$ and $\{\epsilon_i\}_{i=1}^n$.*

**Remark 2** *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 $\epsilon_n$ is nonincreasing and $\epsilon_n^{-2}$ 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 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 3** *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 4** *For the bound (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 $\epsilon_n$ to decay slower than $n^{-1/4}\log^{3/4}(n)$ and $J$ to be finite. Intuitively speaking, if the $\epsilon$ 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 5** *Theorem 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 $\epsilon_n$, 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 5 shows that there is no significant variance inflation for different choices of $\epsilon_n$ sequences in practice. The complexity of the function class containing $\pi_i$ increases as the lower and upper bounds of propensity score get wider, but our proof only relies on the lower bound $\epsilon_i$.*

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 C in the supplementary material. This indicates that the efficiency of learning ITR is not significantly affected by using sequentially generated data.

**Example 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 (4) is finite. The assumptions in Theorem 1 are then satisfied.*

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

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

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 2.1, 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

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

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 13). 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 6** *Assume the main trial satisfies Assumptions 1, 2, 4, 5. If we take a function class satisfying Assumption 6 in Algorithm 1, then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{W}_n(f) \lesssim \frac{r}{\sqrt{n}\epsilon_n} J_{[]}(\|F\|_{\mathbb{P},2}, \mathcal{F}, L_2(\mathbb{P})). \tag{6}$$

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 $\epsilon_n$ 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.

### 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 $i$th $(i = 1,\ldots,n)$ 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 $\epsilon_i'$ 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 7** *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 $\delta > 0$,*

$$\left|\frac{1}{n}\sum_{i=1}^n\left[\mathcal{V}(\hat{f}_{i-1}) - R_i\right]\right| \leq C' r\left[\sqrt{\frac{\delta\wedge\delta^2}{n}} + \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i'\right)\right], \tag{7}$$

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

**Remark 8** *The concentration bound implies that the training regret is upper bounded by the average of the $\epsilon_i'$ sequence plus a term of order $O_p(1/\sqrt{n})$. $\epsilon_i'$ 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 9** *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 $\epsilon$-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 10** *Let the assumptions in Theorems 1 and 7 hold. With a probability higher than $1 - e^{-\delta}$ for any $\delta > 0$,*

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

*where $C''$ is a constant depending on $r, b, J, \delta$ 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 7, which characterize the loss in the value due to exploration and increases as $\epsilon_i'$ increases. On the other hand, the last term, which comes from Theorem 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.

### 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 $\epsilon_n$ and $\epsilon'_n$. Here we assume $\epsilon_n$ and $\epsilon'_n$ decreases polynomially with $n$ since the upper bounds in (5) and (7) are dominated by polynomial terms of $n$.

**Theorem 11** *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 $\theta$ for the test regret and positive in $\theta'$ for the training regret. When $\theta$ and $\theta'$ 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 11 provides a balance point where the two rates match with each other. In Sections 5 and 6, we will further demonstrate the tradeoff between training and test values using numerical examples. Note that $\epsilon$-greedy satisfies the assumptions with $\theta = \theta'$ and $\epsilon_0 = \epsilon'_0$.

## 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 $\epsilon$-greedy. We call the full algorithm SRAT-E in this case.

To build a bridge between $\epsilon$-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 $i$th 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 $\alpha_i$ 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 $\gamma_i$ 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 $\epsilon_i$ 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

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

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

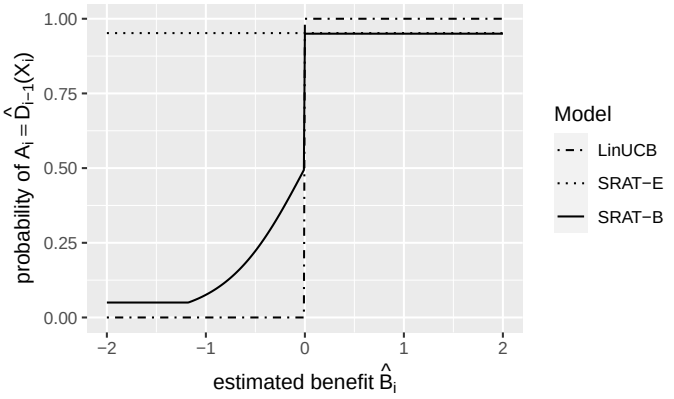


Figure 1: The randomization probability $\mathbb{P}(A_i = \hat{\mathcal{D}}_{i-1}(\boldsymbol{X}_i)|\boldsymbol{H}_{i-1}, \boldsymbol{X}_i)$ of SRAT-E, SRAT-B and LinUCB when $\epsilon_i = 0.05$ and $\gamma_i = 0.4$.

The relationship between SRAT-E, SRAT-B and LinUCB can be illustrated in Figure 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.

## 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 $\Sigma$ 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 $\nu_0$ 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 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., 2018). 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 4 for SRAT-E and SRAT-B differently, and the truncation parameter $\epsilon_n$ 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 11. 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 $O(n)$.

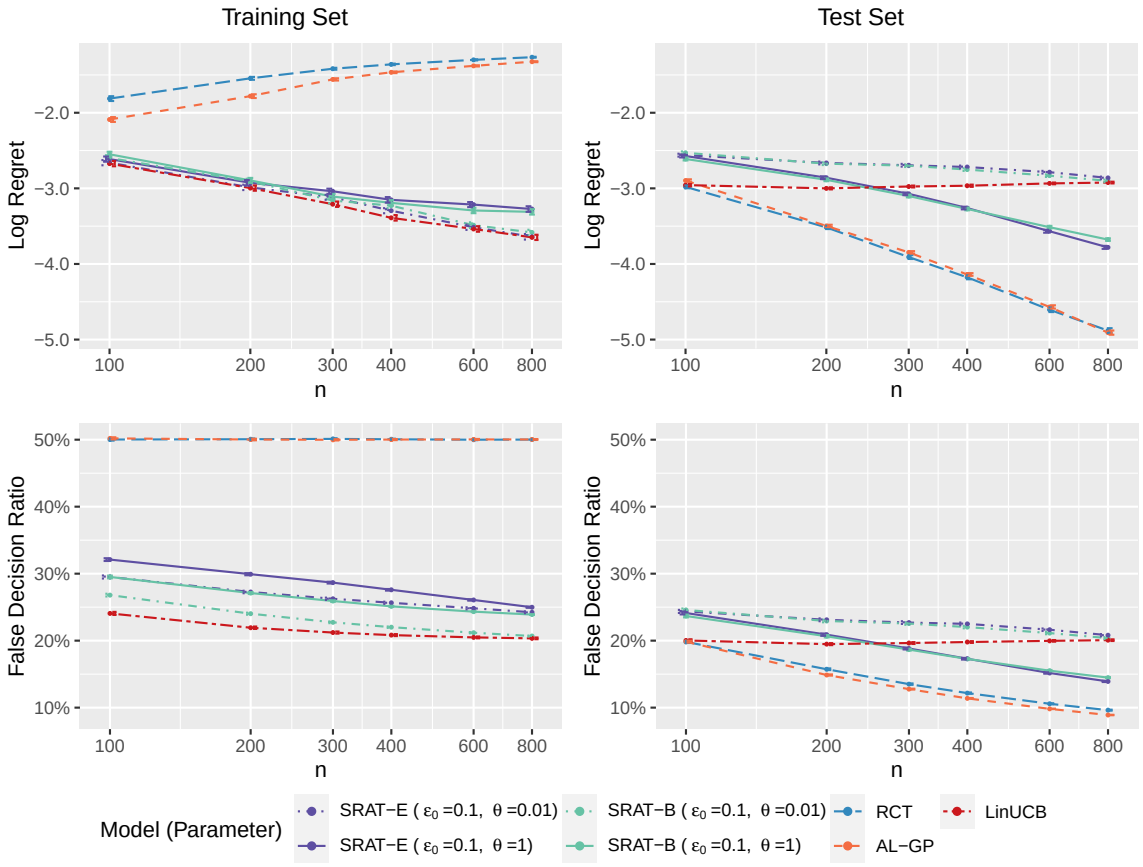


Figure 2: Scenario 1. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size $n$.

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 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 4, the training and test regrets are monotone in the parameters $\epsilon_0$ and $\theta$. Therefore, to save space, we only show two possible combinations of parameters here. The scheduling parameter $\gamma_i$ 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 $\alpha$ increases from 0.2. In our experiment settings, $\alpha$ 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 $1,000$

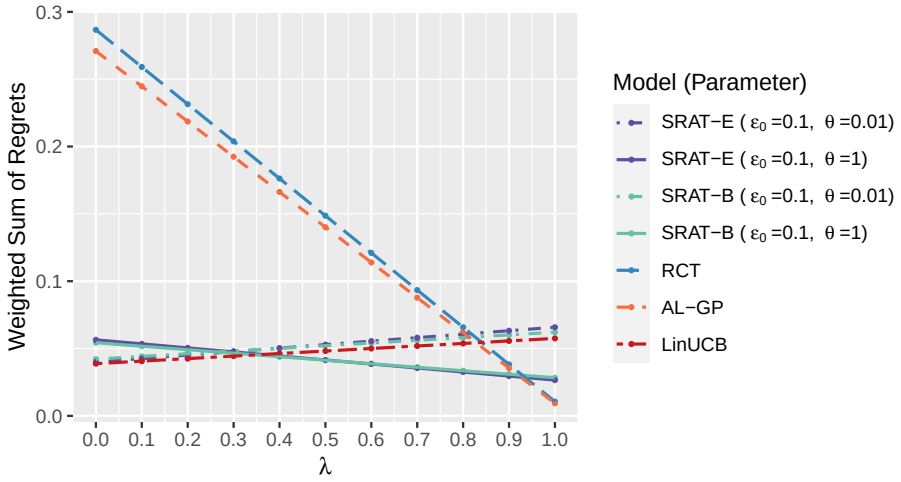


Figure 3: The weighted sum of training and test regrets in scenario 1 when $n = 800$.

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. The plot of scenario 2, Figure 7, is included in the supplementary material since it shows a similar conclusion as scenario 1.

According to Figure 2, LinUCB is the greediest on the training process, with the least regret and false decision ratio. As discussed in Section 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 $\epsilon$-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,

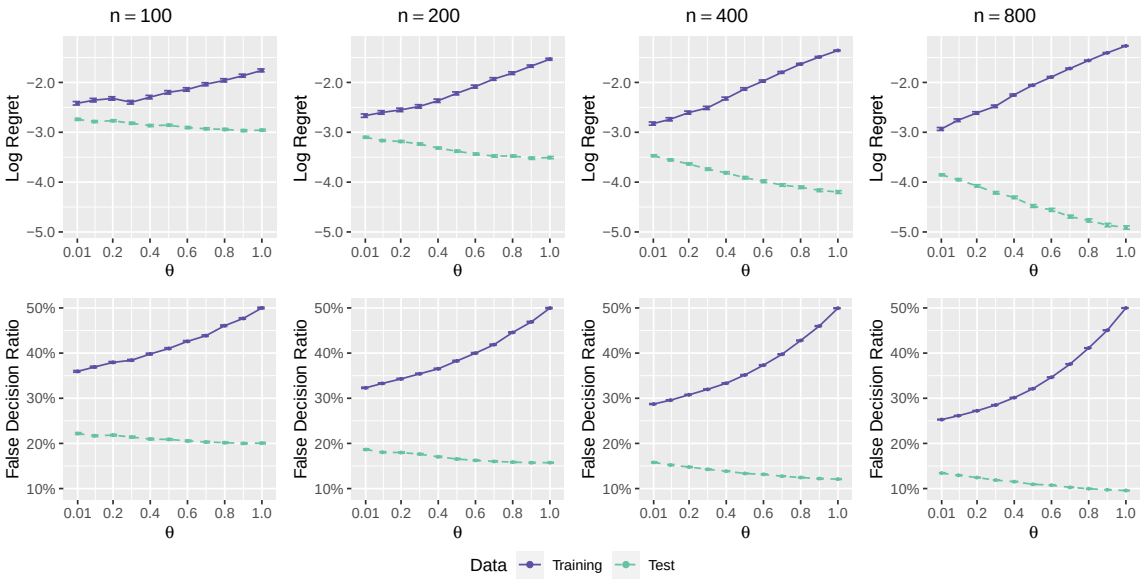


Figure 4: Scenario 1 with $\epsilon_0 = 0.5$. The regret (logarithmic scale) and the false decision ratio on the training or test set against parameter $\theta$.

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 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 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 $\epsilon_0$ equals 0.1 and the decay parameter $\theta$ 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 $\theta$ and sample size $n$ is demonstrated in Figure 4 for scenario 1. Since SRAT-B performs quite similarly to SRAT-E as shown in Figures 2 and 7, we omit it here to save space. The parameter $\theta$ 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 $\epsilon_0$'s give similar conclusion, and smaller $\epsilon_0$ means better training performance and worse test performance. The values and standard errors of the

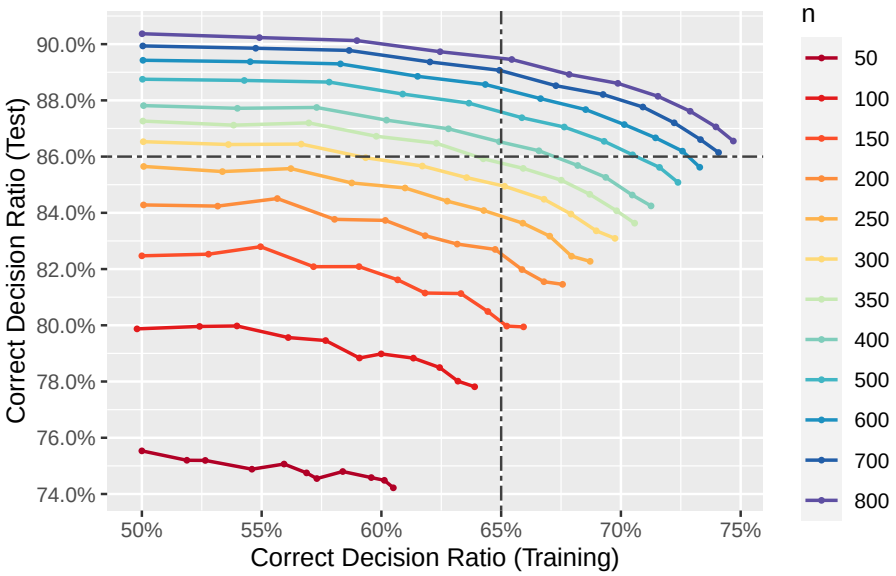


Figure 5: Sample size consideration for SRAT-E in scenario 1 with $\epsilon_0 = 0.5$. Correct decision ratios on the test set against that on the training set. Each line represents a sample size $n$ and each point on the line represents a value of $\theta$. Points to the right correspond to smaller $\theta$, and thus lead to higher correct decision ratio on the training set and lower ratio on the test set.

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 $\theta$ increases, $\epsilon_i$ 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 $\theta$. In practice, the training regret is more affected than the test regret by $\theta$. As shown in Figure 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 $\theta$ 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 5 for SRAT-E in scenario 1. Note that the variation trends of correct decision ratios against $\theta$ are opposite for the training and test data. In particular, $\theta$ should be small enough so that the decision process is 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 $n = 250$, $\theta \leq 0.4$

| 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 1: Clinical trial sample sizes needed for different requirements of correct decision ratios on the training and test sets.

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 $\theta$ for $n = 700$, and any $\theta$ 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 1. Since larger $\theta$ generates better ITR and ITR is our ultimate goal, we report the largest $\theta$ corresponding to the minimum sample size required.

## 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.

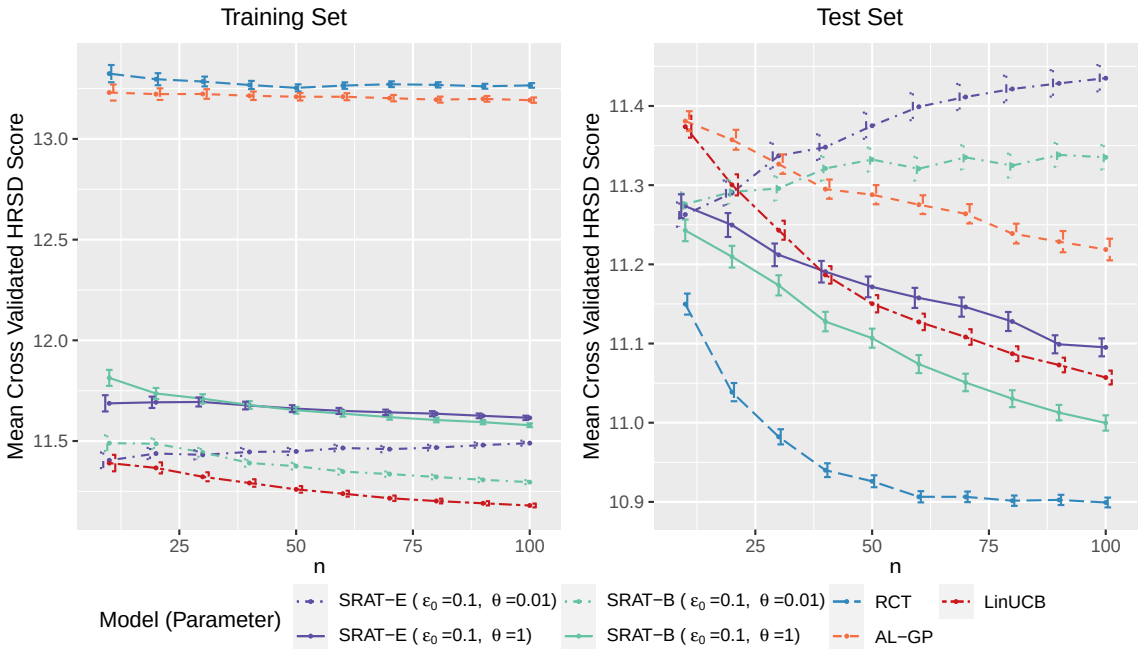


Figure 6: Mean cross-validated HRSD scores against the sample size $n$.

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 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 5, we know that the training and test values are monotone in $\epsilon_0$ and $\theta$. 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 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.

## 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 $\epsilon$-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 paper, 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, 2005; 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., 2018). 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.

## Acknowledgments

The authors would like to thank the action editor and anonymous reviewers for their insightful comments and suggestions. This research was supported in part by US NIH grants R01GM126550, GM124104 and MH117458, and NSF grant DMS 2100729.

## Appendix A. Preliminaries

We provide useful lemmas used for proving our main theorems, among which Lemmas 12 to 15 are quoted from existing literature without proof.

Talagrand's inequality (Talagrand, 1994; Bousquet, 2002) below is used to prove the convergence rate in Theorem 1 for i.i.d. data in the pilot trial. The following version is taken from Steinwart and Scovel (2007, Theorem 5.3).

**Lemma 12 (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 $\|f\|_\infty \leq U < \infty$ 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 $\sigma^2$ be a positive real number such that $\sigma^2 \geq \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathrm{Var}[f(\boldsymbol{X}_i)]$. Then for all $\delta \geq 0$,*

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

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

$$\frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{M}_n(f) \leq \mathcal{R}_n(\mathcal{F}) \leq 2\sup_{\mathbb{P}} \mathbb{E}\sup_{f\in\mathcal{F}} \mathbb{M}_n(f) + \frac{D}{2\sqrt{n}}, \tag{8}$$

where $D = \inf_{z\in\mathcal{Z}} \sup_{f,f'\in\mathcal{F}}[f(z) - f'(z)] \geq 0$. 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) $\epsilon$-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 $\left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{v}_i(\eta) - f(\boldsymbol{z}_i(\eta))|^p\right)^{1/p} \leq \epsilon$. The sequential covering number of a function class $\mathcal{F}$ on a given tree $\boldsymbol{z}$ is defined as

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

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 13 (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*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_{i-1} f(Z_i)\right| > t\right) \leq 8L\exp\left(-\frac{t^2}{c\log^3 n \mathcal{R}_n^2(\mathcal{F})}\right) \tag{9}$$

*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 $L_\psi$-Orlicz norm. Freedman's inequality is an extension of Bernstein's inequality to martingale difference sequences.

**Lemma 14 (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 $L_\psi$-Orlicz norm.

**Lemma 15 (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*

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

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

The following inequality is a key step in the proof of Lemma 6. 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 16** *Suppose $\{X_i\}_{i \ge 1}$ 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 14 in the form of a martingale process. For a $\mathcal{G}_i$-adapted sequence $\{Y_i\}_{i \ge 1}$, take $\mathbb{E}[f(Y_i)|\mathcal{G}_{i-1}] - f(Y_i)$ as $X_i$ in Lemma 14, which is a martingale difference sequence. The supremum term can be bounded as $\|\mathbb{E}[f(Y_i)|\mathcal{G}_{i-1}] - f(Y_i)\|_\infty \le 2\|f\|_\infty$. Scale both sides by a factor of $\sqrt{n}$ and we get

$$\mathbb{P}(\left|\sqrt{n} \mathbb{M}_n(f)\right| > t) \le 2 \exp \left\{ -\frac{1}{2} \frac{t^2}{\|\langle M \rangle_n\|_\infty / n + 2\|f\|_\infty t / (3\sqrt{n})} \right\}, \tag{10}$$

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 (10) to Lemma 15 and expand the $L_\psi$-Orlicz norm. ■

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

**Lemma 17** *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*

$$\mathbb{E}\left[ \frac{G(\boldsymbol{X}_i)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \right] = \mathbb{E}\left[ \frac{G(\boldsymbol{X}_i)}{\pi_i(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \middle| \mathcal{G}_{i-1} \right] = 2\mathbb{E}[G(\boldsymbol{X}_i)], \tag{11}$$

$$\mathbb{E}\left[ \frac{G(\boldsymbol{X}_i)}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \right] = \mathbb{E}\left[ \frac{G(\boldsymbol{X}_i)}{\pi_i^2(A_i; \boldsymbol{H}_{i-1}, \boldsymbol{X}_i)} \middle| \mathcal{G}_{i-1} \right] \le \left( \frac{1}{\epsilon_i} + \frac{1}{0.5} \right) \mathbb{E}[G(\boldsymbol{X}_i)]. \tag{12}$$

**Proof** For the first equation (11), notice that by tower property and the definition of $I_i$,

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

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 $\mathbb{E}[\mathbb{1}(I_i = 1) | \boldsymbol{H}_{i-1}, \boldsymbol{X}_i]$. It is also true for the second term in the bracket. Therefore, the right-hand side equals $\mathbb{E}[2G(\boldsymbol{X}_i)|\mathcal{G}_{i-1}]$. The result follows from the assumption that $\boldsymbol{X}_i$ is independent of the history. The first equality in (11) can be proved by taking expectation of both sides of the equation.

Similarly, when $G$ is divided by the square term of $\pi_i$ in (12),

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

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 $\epsilon_i$, the inequality follows. Now for the first term in (12), its upper bound can be proved by taking expectation of both sides of the inequality. ■

## Appendix B. Proof of Lemma 6

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 $\tau^2$ 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 $\tau^2$. However, we will show that $\tau^2$ 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 (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:

$$\sum_q 2^{-q}\sqrt{\log N_q} < \infty, \tag{13}$$

$$\left\| (\sup_{f,g\in\mathcal{F}_{qj}} |f-g|)^* \right\|_{\mathbb{P},2,\infty} < 2^{-q}, \quad \forall j \in \{1,\dots,N_q\}, \tag{14}$$

$$\sup_{f,g\in\mathcal{F}_{qj}} \|f-g\|_{\mathbb{P},2} < 2^{-q}, \quad \forall j \in \{1,\dots,N_q\}. \tag{15}$$

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 $N_q = \bar{N}_1 \dots \bar{N}_q$ sets. Conditions (13) - (15) continue to hold since $(\log \prod_{p=1}^q \bar{N}_p)^{1/2} \le \sum_{p=1}^q (\log \bar{N}_p)^{1/2}$.

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 $\xi$ be the function of choosing the representative. In addition, let $\Delta$ 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 $|\phi(Af_1) - \phi(Af_2)| \le |f_1 - f_2|$ since $\phi(Af)$ is Lipschitz 1 with respect to $f$. Hence we obtain that $|R_i|\,\Delta_q h^f(\boldsymbol{Z}_i) \le \Delta_q f(\boldsymbol{X}_i)/\pi_i(A_i; \hat{f}_{i-1})$. 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 $q \ge q_0$ as

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

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

$$h^f - \xi_{q_0} h^f = (h^f - \xi_{q_0} h^f) B_{q_0} f + \sum_{q=q_0+1}^{\infty} (h^f - \xi_q h^f) B_q f + \sum_{q=q_0+1}^{\infty} (\xi_q h^f - \xi_{q-1} h^f) A_{q-1} f \tag{16}$$

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 (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

$$\begin{aligned}
\mathbb{E}^* \left\|\sqrt{n}\mathbb{W}_n(f)\right\|_{\mathcal{F}} \leq & \mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n(h^f - \xi_{q_0} h^f) B_{q_0} f\right\|_{\mathcal{F}} && (17)\\
& + \mathbb{E}^* \left\| \sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n(h^f - \xi_q h^f) B_q f \right\|_{\mathcal{F}} && (18)\\
& + \mathbb{E}^* \left\| \sum_{q=q_0+1}^{\infty} \sqrt{n}\mathbb{M}_n(\xi_q h^f - \xi_{q-1} h^f) A_{q-1} f \right\|_{\mathcal{F}} && (19)\\
& + \mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n \xi_{q_0} h^f\right\|_{\mathcal{F}}. && (20)
\end{aligned}$$

To bound the first term (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 17 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

$$\mathbb{E}^* \left\|\sqrt{n}\mathbb{M}_n(h^f - \xi_{q_0} h^f) B_{q_0} f\right\|_{\mathcal{F}} \lesssim r 2^{-q_0} \sqrt{\operatorname{Log} N_{q_0}}. \tag{21}$$

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 $\left|(h^f - \xi_q h^f) B_q f\right| \leq$

$|R_i|\,\Delta_q f B_q f/\pi_i(A_i;\hat{f}_{i-1})$, the second term (18) can be bounded by

$$\begin{aligned}
&\mathbb{E}^*\left\|\sum_{q=q_0+1}^{\infty}\sqrt{n}\mathbb{M}_n(h^f-\xi_q h^f)B_q f\right\|_{\mathcal{F}}\\
=&\sum_{q=q_0+1}^{\infty}\mathbb{E}^*\left\|\sqrt{n}\mathbb{M}_n\frac{|R_i|}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q f B_q f\right\|_{\mathcal{F}} \qquad (22)\\
&+\sum_{q=q_0+1}^{\infty}2\sqrt{n}\mathbb{E}^*\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{i-1}\frac{|R_i|}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q f B_q f\right\|_{\mathcal{F}}.
\end{aligned}$$

By Corollary 16, for each $q$ in the first term in (22), the expectation can be split into two parts:

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

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

$$\left\|\frac{|R_i|}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q f B_q f\right\|_{\infty}\le\frac{r}{\epsilon_n}\sqrt{n}a_{q-1} \qquad (23)$$

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

$$\begin{aligned}
&\left\|\sum_{i=1}^{n}\mathrm{Var}\left[\frac{|R_i|}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q f B_q f\middle|\mathcal{G}_{i-1}\right]\right\|_{\infty}\\
\le&\sum_{i=1}^{n}\left\|\mathbb{E}\left[\frac{\tau^2}{\pi_i^2(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q^2 f B_q^2 f\middle|\mathcal{G}_{i-1}\right]\right\|_{\infty} \qquad (24)\\
\le&\sum_{i=1}^{n}\left\|\left(\frac{1}{\epsilon_i}+\frac{1}{0.5}\right)\mathbb{E}\left(\tau^2\Delta_q^2 f B_q^2 f\right)\right\|_{\infty}.
\end{aligned}$$

The last inequality comes from Lemma 17. For any non-negative random variable $X$, we have the inequality $\|X\|_{2,\infty}^2\le\sup_{t>0}t\mathbb{E}\left[X\mathbb{1}(X>t)\right]\le 2\|X\|_{2,\infty}^2$. Then

$$\sqrt{n}a_q\mathbb{E}\left(\Delta_q f B_q f\right)\le\sqrt{n}a_q\mathbb{E}\left(\Delta_q f\mathbb{1}(\Delta_q f>\sqrt{n}a_q)\right)\le 2\|\Delta_q f\|_{\mathbb{P},2,\infty}^2\le 2\cdot 2^{-2q}. \qquad (25)$$

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

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

Using Lemma 17 again and the inequality (25), the second term in (22) can be bounded as

$$\begin{aligned}&\mathbb{E}^*\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{i-1}\frac{|R_i|}{\pi_i(A_i;\boldsymbol{H}_{i-1},\boldsymbol{X}_i)}\Delta_q f B_q f\right\|_{\mathcal{F}}\\ \leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^*\left\|2r\mathbb{E}\left(\Delta_q f B_q f\right)\right\|_{\mathcal{F}}\leq 4r\frac{1}{\sqrt{n}a_{q-1}}2^{-2q}.\end{aligned}\tag{26}$$

Now apply the above bounds (23), (24), (26) on (18) to find

$$\begin{aligned}&\mathbb{E}^*\left\|\sum_{q=q_0+1}^{\infty}\sqrt{n}\mathbb{M}_n(h^f-\xi_q h^f)B_q f\right\|_{\mathcal{F}}\\ \lesssim&\sum_{q=q_0+1}^{\infty}\left[\frac{1}{\sqrt{n}}\frac{r}{\epsilon_n}\sqrt{n}a_{q-1}\operatorname{Log}N_q+\frac{1}{\sqrt{n}}\sqrt{2\tau^2\sum_{i=1}^{n}\left(\frac{1}{\epsilon_i}+\frac{1}{0.5}\right)\frac{a_{q-1}}{a_q}2^{-2q}}\sqrt{\operatorname{Log}N_q}\right.\\ &\left.+\sqrt{n}4r\frac{1}{\sqrt{n}a_{q-1}}2^{-2q}\right]\\ \lesssim&\sum_{q=q_0+1}^{\infty}\left[\frac{r}{\epsilon_n}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}+r\right]2^{-q}\sqrt{\operatorname{Log}N_q}.\end{aligned}\tag{27}$$

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 (19), first note that it is bounded by

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

since the partition is nested. Then we can use Corollary 16 as in the bound of the second term (18). The maximum of $L_\infty$ 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 17 and assumption (15) to find

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

Combining the two parts together, we have

$$\begin{aligned}
&\mathbb{E}^*\left\|\sum_{q=q_0+1}^{\infty}\sqrt{n}\mathbb{M}_n(\xi_q h^f-\xi_{q-1}h^f)A_{q-1}f\right\|_{\mathcal{F}}\\
\lesssim&\sum_{q=q_0+1}^{\infty}\left[\frac{1}{\sqrt{n}}\frac{r}{\epsilon_n}\sqrt{n}a_{q-1}\operatorname{Log}N_q+\frac{1}{\sqrt{n}}\sqrt{2\tau^2\sum_{i=1}^{n}\left(\frac{1}{\epsilon_i}+\frac{1}{0.5}\right)2^{-2q+2}}\sqrt{\operatorname{Log}N_q}\right] \qquad (28)\\
\lesssim&\sum_{q=q_0+1}^{\infty}\left[\frac{r}{\epsilon_n}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}\right]2^{-q}\sqrt{\operatorname{Log}N_q}.
\end{aligned}$$

For the last term (20), consider two cases on whether the envelope function $F$ is bounded by $\sqrt{n}a_{q_0}$ or not. Apply Corollary 16 in the first case. With the supremum part bounded by

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

and the conditional variance part bounded by

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

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

$$\begin{aligned}
&\mathbb{E}^*\left\|\sqrt{n}\mathbb{M}_n\xi_{q_0}h^f\right\|_{\mathcal{F}}\\
\lesssim&\frac{r}{\epsilon_n}2^{-q_0}\sqrt{\operatorname{Log}N_{q_0}}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}\|F\|_{\mathbb{P},2}\sqrt{\operatorname{Log}N_{q_0}}+\frac{r}{a_{q_0}}\|F\|_{\mathbb{P},2}^2 \qquad (29)\\
\lesssim&\left[\frac{r}{\epsilon_n}+\tau\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\epsilon_i}}+r\right]2^{-q_0}\sqrt{\operatorname{Log}N_{q_0}},
\end{aligned}$$

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

Finally, combine the four upper bounds (21), (27), (28) and (29) together to get

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

■

## Appendix C. Proof of Theorem 1

Using Lemma 13 and Lemma 6, we can give our proof of Theorem 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 $\phi$-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

$$\begin{aligned}
&\mathcal{V}(f^*) - \mathcal{V}(\hat{f}_n) \\
\leq &\frac{1}{n_0+n}\left[\sum_{i=1}^{n_0}\mathbb{E}h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) + \sum_{i=1}^{n}\mathbb{E}_{i-1}h^{\hat{f}_n}(\boldsymbol{Z}_i) - \sum_{i=1}^{n_0}h^{\hat{f}_n}(\boldsymbol{Z}_i^{(0)}) - \sum_{i=1}^{n}h^{\hat{f}_n}(\boldsymbol{Z}_i)\right] \\
\leq &\sup_{f\in\mathcal{F}}\frac{1}{n_0+n}\left[\sum_{i=1}^{n_0}\left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] + \sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right].
\end{aligned} \tag{30}$$

Now it suffices to bound the right-hand side of (30).

We will use Lemma 13 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 13 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 (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 6. 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 (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,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right| > C\frac{r^2bJ}{\epsilon_n^2}\sqrt{n\log^3 n\delta}\right) \leq e^{-\delta} \tag{31}$$

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 12, 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 12. 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 6, 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,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^{n_0}\left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] \geq 3rJ\sqrt{n_0} + \sqrt{4\delta br^2 n_0} + 4rb\delta\right) \leq e^{-\delta}. \tag{32}$$

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

$$\begin{aligned}\mathbb{P}\Bigg(\sup_{f\in\mathcal{F}}\frac{1}{n_0+n}\left[\sum_{i=1}^{n_0}\left[\mathbb{E}h^f(\boldsymbol{Z}_i^{(0)}) - h^f(\boldsymbol{Z}_i^{(0)})\right] + \sum_{i=1}^{n}\left[\mathbb{E}_{i-1}h^f(\boldsymbol{Z}_i) - h^f(\boldsymbol{Z}_i)\right]\right] \\ \geq \frac{C}{n_0+n}\left[(rJ + r\sqrt{\delta b})\sqrt{n_0} + rb\delta + \frac{r^2bJ}{\epsilon_n^2}\sqrt{n\log^3 n\delta}\right]\Bigg) \leq e^{-\delta}\end{aligned} \tag{33}$$

for some constant $C$ and any $\delta > 0$. The result on test data follows by combining inequalities (30) and (33). ■

## Appendix D. Proof of Theorem 7

**Proof** First note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{V}(\hat{f}_{i-1}) - R_i\right]\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}\left[R_i - \mathbb{E}_{i-1}R_i\right]\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{i-1}R_i - \mathcal{V}(\hat{f}_{i-1})\right|. \tag{34}$$

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 (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 14. 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,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[R_i - \mathbb{E}_{i-1}R_i\right]\right| \geq t\right) \leq 2\exp\left\{-\frac{1}{2}\frac{nt^2}{r^2 + 2rt/3}\right\}. \tag{35}$$

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

■

## Appendix E. Proof of Corollary 10

**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 (34). Now let the right-hand side of (35) be $e^{-\delta}/3$ and the right-hand side of (31) and (32) be $e^{-\delta}/3n$ by inverting the two bounds in the proof of Theorem 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 (31) does not actually exist. Hence letting $i \log^3 i = 0$ will work. ■

## Appendix F. Proof of Theorem 11

**Proof** For the test regret bound (5), note that the last term

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

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 $\epsilon$ to be non-increasing and the last term to converge to 0. Therefore, $\theta$ should be no greater than 1 and larger than 0. Similarly, for the training regret bound (7), the second term that contains $\epsilon_n$ dominates. The result follows by substituting $\epsilon_n$ into the convergence rate and letting the two rates be equal. Note that

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

■

## Appendix G. Additional Simulation Results

Figure 7 of scenario 2 demonstrates similar results as Figure 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.

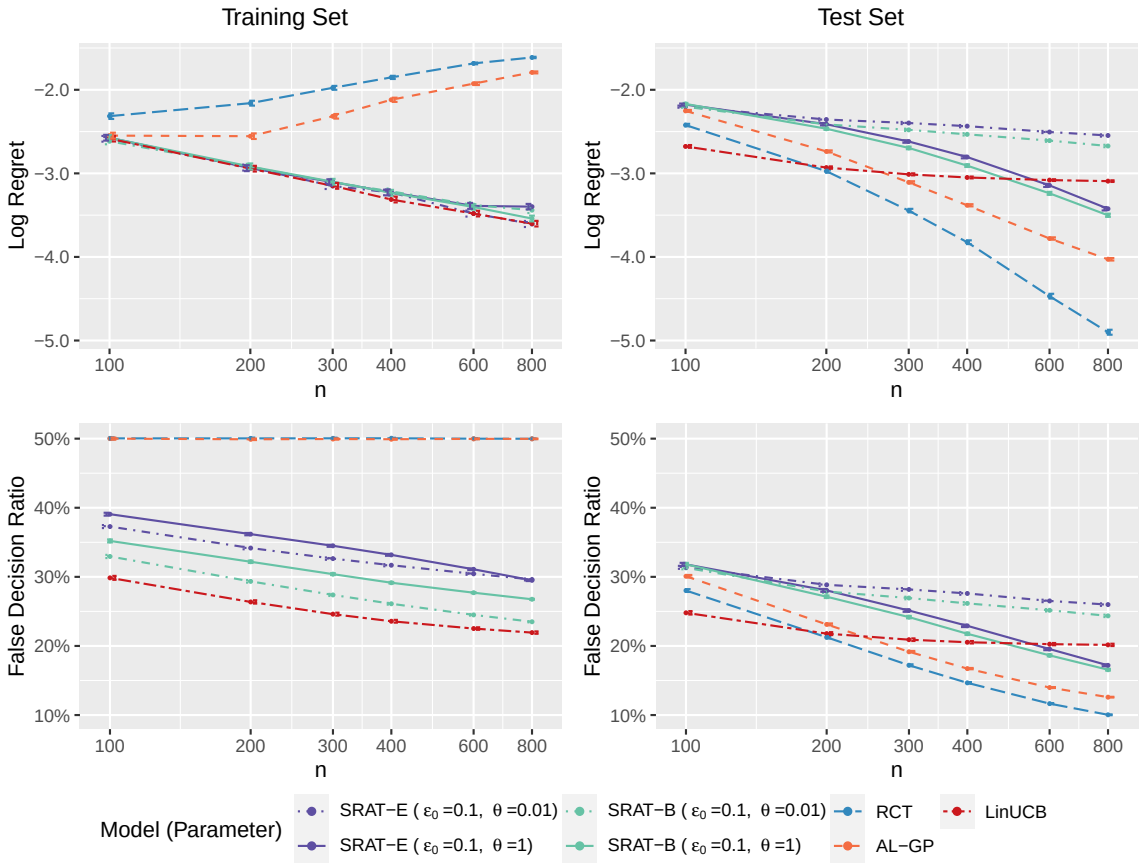


Figure 7: Scenario 2. The regret (logarithmic scale) and the false decision ratio on the training or test set against sample size $n$.

## References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Jongsig Bae and Shlomo Levental. Uniform CLT for Markov chains and its invariance principle: a martingale approach. *Journal of Theoretical Probability*, 8(3):549–570, 1995.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer, 2009.

Antoine Chambaz, Wenjing Zheng, and Mark J van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Annals of Statistics*, 45(6):2537, 2017.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

Andrew G Chapple and Peter F Thall. A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III. *Biometrics*, 75(2):371–381, 2019.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, pages 1–16, 2020.

Jingxiang Chen, Haoda Fu, Xuanyao He, Michael R Kosorok, and Yufeng Liu. Estimating individualized treatment rules for ordinal treatments. *Biometrics*, 74(3):924–933, 2018.

Yuan Chen, Ying Liu, Donglin Zeng, and Yuanjia Wang. *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*, 2019. URL `https://CRAN.R-project.org/package=DTRlearn2`. R package version 1.0.

Shein-Chung Chow. Adaptive clinical trial design. *Annual Review of Medicine*, 65:405–415, 2014.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

David A Freedman. On tail probabilities for martingales. *Annals of Probability*, pages 100–118, 1975.

Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for optimal decision making. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 149–154. Springer, 2007.

Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

Feifang Hu and William F Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons, 2006.

Jianhua Hu, Hongjian Zhu, and Feifang Hu. A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, 110(509):357–367, 2015.

Martin B Keller, James P McCullough, Daniel N Klein, Bruce Arnow, David L Dunner, Alan J Gelenberg, John C Markowitz, Charles B Nemeroff, James M Russell, Michael E Thase, et al. 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, 2000.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53, 2011.

Andreas Krause and Cheng S Ong. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.

Tze Leung Lai, Philip W Lavori, Mei-Chiung I Shih, and Branimir I Sikic. Clinical trial designs for testing biomarker-based personalized therapies. *Clinical Trials*, 9(2):141–154, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Philip W Lavori and Ree Dawson. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society. Series A*, 163(1): 29–38, 2000.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, pages 661–670, 2010.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. 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, 2020.

Ying Liu, Yuanjia Wang, Michael R Kosorok, Yingqi Zhao, and Donglin Zeng. Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, 37(26):3776–3788, 2018.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020.

Stanislav Minsker, Ying-Qi Zhao, and Guang Cheng. Active clinical trials for personalized medicine. *Journal of the American Statistical Association*, 111(514):875–887, 2016.

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

Susan A Murphy, David W Oslin, A John Rush, and Ji Zhu. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*, 32(2):257–262, 2007.

Yoichi Nishiyama. Some central limit theorems for $\ell_{infty}$-valued semimartingales and their applications. *Probability Theory and Related Fields*, 108(4):459–494, 1997.

Yoichi Nishiyama et al. Weak convergence of some classes of martingales with jumps. *Annals of Probability*, 28(2):685–712, 2000.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2):693–721, 2013.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180, 2011.

Alexander Rakhlin and Karthik Sridharan. *Statistical learning and sequential prediction*. Book Draft, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161 (1-2):111–153, 2015.

Lindsay A Renfro, Himel Mallick, Ming-Wen An, Daniel J Sargent, and Sumithra J Mandrekar. Clinical trial designs incorporating predictive biomarkers. *Cancer Treatment Reviews*, 43:74–82, 2016.

Marie-Karelle Riviere, Ying Yuan, Jacques-Henri Jourdan, Frédéric Dubois, and Sarah Zohar. Phase I/II dose-finding design for molecularly targeted agent: plateau determination using adaptive randomization. *Statistical Methods in Medical Research*, 27(2):466–479, 2018.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Michel Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, pages 28–76, 1994.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

Peter F Thall. Ethical issues in oncology biostatistics. *Statistical methods in medical research*, 11(5):429–448, 2002.

Peter F Thall, Hoang Q Nguyen, Thomas M Braun, and Muzaffar H Qazilbash. Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes. *Biometrics*, 69(3):673–682, 2013.

Sara Van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Annals of Statistics*, pages 1779–1801, 1995.

Aad W Van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, 30(1):100–121, 2002.

Li-Xin Zhang, Feifang Hu, Siu Hung Cheung, and Wai Sum Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *Annals of Statistics*, 35(3):1166–1182, 2007.

Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

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

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.