Optimal sequential treatment allocation

Anders Bredahl Kock and Martin Thyrsgaard*
University of Oxford, Aarhus University, Northwestern University, and CREATES

First version: May 2017
This version: June 2020

Abstract

In treatment allocation problems the individuals to be treated often arrive sequentially. We study a problem in which the policy maker is not only interested in the expected cumulative welfare but is also concerned about the uncertainty/risk of the treatment outcomes. A sequential treatment policy, which attains near minimax optimal regret, is studied with and without covariates. We also demonstrate that the expected number of suboptimal treatments only grows slowly in the number of treatments. Finally, we study a setting where outcomes are observed only with delay. Simulations illustrate the theoretical results.

Keywords: Sequential treatment allocation, General welfare function, Outcomes observed with delay, Multi-armed bandits, Covariates, Ethical guarantees
JEL classifications: C18, C22, J68

1 Introduction

A policy maker must often assign treatments gradually as the individuals to be treated do not arrive simultaneously. For example, people become unemployed gradually throughout the year and assignment to one of several unemployment programs is often made shortly thereafter. Similarly, patients with too high blood pressure arrive gradually to a medical clinic and the doctor assigns one of several treatments to each of them. The policy maker or doctor gradually accrues information by observing the outcome of previous treatments prior to the next assignment. Throughout the paper we shall use these two examples as illustrations of our results and be particularly concerned with

*Address for correspondence: anders.kock@economics.ox.ac.uk. This work was supported by CREATES which is funded by the Danish National Research Foundation (DNRF78). Furthermore, Martin Thyrsgaard's work is supported by grant 9033-00003B from the Danish National Research Foundation.

how treatments should be assigned in order to maximize welfare. In doing so, one faces a tradeoff between exploring which treatment works best and exploiting the information gathered so far from previous assignments in order to assign the best treatment to as many individuals as possible.

The above setup is in stark contrast to typical estimation of treatment effects where one presupposes the existence of a data set of a certain size N (perhaps obtained from a randomized control trial). Thus, in the typical setting, the size and composition of the data set are determined prior to estimation. Based on this given data set, treatment effects would be estimated and assignments made. We consider the case where the observed data that the treatment assignments must be based on is a part of the policy in the sense that it depends on the assignments made by the policy maker. Thus, the policy maker enters already in the design phase of the treatment program and can adjust the experiment as data is accumulated. Furthermore, the sample size itself may be a random variable unknown to the policy maker as one may not know a priori how many individuals will become unemployed in the course of the year that the program is scheduled to run. Thus, the exact amount of experimentation of a good treatment rule will depend on the expected number of individuals to be treated.

We consider a setting where the desirability of a treatment cannot be measured only by its expected outcome. A sensible welfare function must take into account the risk of a treatment. For example, it may well be that drug A is expected to lower the blood pressure slightly more than drug B but A might still not be preferred if it is much more risky than B. In this paper we shall measure the *risk* of a treatment by its variance and take into account that mean as well as variance may be relevant in determining the most desirable treatment. We also indicate how one may incorporate more than two moments into the welfare function thus allowing welfare functions that allow for policy makers to be, say, skewness averse. We study a treatment policy, which we call the *sequential treatment policy*, and show that it achieves near minimax optimal regret compared to the infeasible policy that knows in advance which treatment is best for each individual and assigns this. As individuals with different characteristics may react differently to the same treatment, we also study how covariate information can be incorporated. Thus, we allow for heterogeneous treatment distributions and optimality of the sequential treatment policy is established. Furthermore, an upper bound on the expected number of times that the sequential treatment policy assigns any suboptimal treatment is provided as well. This is an important ethical guarantee since it ensures that the near minimax optimal regret is not obtained at the cost of wild experimentation or maltreatment of many individuals in order to achieve a greater cumulative welfare in the long run.

In addition, we contribute by studying the properties of the sequential treatment policy when the outcomes of previous treatments are observed only with delay. In a medical trial, for example, one may choose to delay the measurement of the outcome of the treatment in order to obtain more precise information of the effect of a certain drug as it takes time for the effect of a drug to set in. The price of this delay is that less information is available when treating other patients prior to

the measurement being made. Thus, there is a tradeoff between obtaining imprecise information quickly (by making the measurement shortly after the treatment) and obtaining more precise information later (by postponing the measurement). We quantify this tradeoff and indicate the optimal delay (when this is a choice variable) and establish that our policy is guaranteed to deliver high welfare even in this setting.

Our approach easily accommodates practical policy concerns restricting the type of treatment rules that are feasible. For instance, the policy maker may want rules that depend on the individual's characteristics in a simple way due to political or ethical reasons.

It should be noted that the goal of this paper is not to test whether one treatment is better than the other ones at the end of the treatment period. This would amount to a pure exploration problem where the purpose of the sampling is to maximize the amount of information at the end of the sample without regard to the welfare of the treated individuals.

1.1 Related literature

Our paper is related to two strands of literature: the literature on statistical treatment rules in econometrics and the one on multi-armed bandit problems. In the former Manski (2004) proposed *conditional empirical success* (CES) rules which take a finite partition of the covariate space and on each set of this partition dictate to assign the treatment with the highest sample average. When implementing CES rules, one must decide on how fine to choose the partition of the covariate space and thus faces a tradeoff between using highly individualized rules and having enough data to accurately estimate the treatment effects for each group in the partition. Among other things, Manski (2004) provides sufficient conditions for full individualization to be optimal. The tradeoff between full individualization of treatments and having sufficient data to estimate the treatment effects accurately is also found in our sequential treatment setting.

Stoye (2009) showed that if one does not restrict how outcomes vary with covariates then full individualization is alway minimax optimal. Thus, if age is a covariate, information on treatment effects for 30 year olds should not be used when making treatment decisions for 31 year olds. This result relies on the fact that without any restrictions on how the outcome distribution varies with covariates, this relationship could be arbitrarily non-smooth such that even similar individuals may carry no information about how treatments affect the other person. Our assumptions rule out such non-smoothness as no practical policy can be expected to work well in such a setting.

Furthermore, our work is related to the recent paper by Kitagawa and Tetenov (2018) who consider treatment allocation through an *empirical welfare maximization* lens. The authors take the view that realistic policies are often constrained to be simple due to ethical, legislative, or political reasons. Our approach is related to theirs in that we also allow the policy maker to focus on simple rules in the sequential framework. Furthermore, Athey and Wager (2017) have used concepts from semiparametric efficiency theory to establish regret bounds that scale with the semiparametrically

efficient variance. The importance of considering other characteristics than the mean treatment outcome has also been emphasized by Qi et al. (2019).

Other papers on statistical treatment rules in econometrics focusing on the case where the sample is given include Chamberlain (2000), Dehejia (2005), Hirano and Porter (2009), Bhattacharya and Dupas (2012), Stoye (2012), Tetenov (2012) and Kasy (2014).

The most important distinguishing feature of our work compared to the classic literature on statistical treatment rules is that we are working in a sequential setting where the individuals to be treated arrive gradually. Thus, we do not have a data set of size N at our disposal from the outset based on which the best treatment must be found. Rather the goal of the policy maker is to construct a way of sampling that strikes the optimal balance between exploration and exploitation. The sequential setting under study poses new challenges such as not maltreating too many individuals in the search for the best treatment and how to handle treatment outcomes that are only observed with delay. Consequently, our paper is also related to the vast literature on bandit problems. In the classic bandit problems one seeks to maximize the expected cumulative reward from pulling arms with unknown means one by one. In a seminal paper Robbins (1952) introduced a class of bandit problems and proposed some initial solutions guaranteeing that the average reward will converge to the mean of the best arm.

Broadly speaking, bandit problems can be classified into three categories based on the nature of the reward process: i) stochastic bandits where the arms are iid across time, ii) the markovian setting where the state of the arms changes according to a Markov process, iii) the adversarial setting in which nature chooses (an adversarial) sequence of rewards at the same time as the experimenter pulls an arm. In this work we focus on the stochastic setting as patients to be treated or unemployed individuals to be assigned to job training programs do not generally coordinate their effort against the doctor or policy maker in an adversarial manner. In the medical example in particular, the interests of the doctor and patient are often well-aligned. Furthermore, the markovian setting is concerned with infinite time horizons amounting to infinitely many treatments being made. In this work we are interested in the case where we have to make a finite, albeit often unknown, number of treatments. In our context of treatment assignment, the adversarial setting has the questionable feature that two individuals with similar covariates may have entirely unrelated treatment outcomes. This is due to the fact that the treatment outcomes are simply chosen by "the opponent/nature" in a way maximizing losses of the policy maker. That being said, we certainly believe that also the adversarial setting or the markovian setting can be of interest to study in the context of sequential treatment allocation problems. In the latter setting, the Gittins index, Gittins (1979), is the most famous procedure. We refer to Lykouris et al. (2017) for an example of a paper focusing on the adversarial setting and to Bubeck et al. (2012) as well as Lattimore and Szepesvári (2018) for general overviews on multi-armed bandits.

While most of the literature on multi-armed bandits has focused on the setting in which one

only targets the arm with the highest mean, there has been a realization in recent years that also the variance of assignments is of importance. For example, the papers Sani et al. (2012) and Vakili and Zhao (2016) studied a setting in which the decision maker targets the arm with the lowest value of the empirical variance minus a multiple of the mean. However, in addition to only studying one specific function of the mean and the variance, the notion of regret employed in these papers allowed for misassignments of some individuals to be offset by gains to others. This may not be viable in many economic applications or in clinical trials where every individual matters. Thus, the notion of welfare employed in this paper takes into account the welfare of every individual.

The paper closest in spirit to ours is the one of Zimin et al. (2014). Like us, they also allow the policy maker to be interested in a function of the mean and the variance. However, there are several differences to that work. First, they study a different policy build around the UCB-policy. Second, the paper by Zimin et al. (2014) does not allow the treatment outcome distributions to depend on covariates. However, it is clear that individuals with different characterisitcs may react differently to the same treatment. It is thus important to allow for individual-specific treatment outcomes and we show that our way of incorporating covariates results in minimax optimal expected regret; cf. Corollary 3.1.1 and Theorem 3.2. Also in the absence of covariates we derive worst-case performance guarantees of our sequential treatment policy in addition to pointwise upper bounds established in Zimin et al. (2014) (who do not study uniform upper bounds on the regret of their procedure). Third, we provide upper bound on the expected number of suboptimal assignments made by our sequential treatment policy, which ensures that only few individuals are not assigned to the best treatment. Fourth, we consider the case where the outcome of treatments is only observed with delay. As explained, delay creates a tradeoff between obtaining imprecise information quickly and obtaining precise information later. This makes the sequential treatment problem more challenging as the policy maker must now also choose when to make a measurement in addition to which treatment to assign.

The first paper which considered bandit problems where one observes a covariate prior to making an allocation decision was Woodroofe (1979) who made a parametric assumption on how covariates affect outcomes. The first work allowing covariates to affect the distribution of outcomes in a nonparametric way was Yang et al. (2002). These works considered a setting where the decision maker is interested solely in the mean outcome. From an algorithmic point of view our work is related to Perchet and Rigollet (2013) who introduced a successive elimination (SE) policy of suboptimal arms. Their policy is in turn related to the work of Even-Dar et al. (2003). Finally, Shin et al. (2019) have studied the estimation error after adaptive allocation.

The term *optimal sequential treatment allocation* in the bandit framework as discussed in this paper should not be confused with similar terms in the medical statistics literature. In that literature *adaptive treatment strategies/adaptive interventions* and *dynamic treatment regimes* refer to a setting where the same individual is observed repeatedly over time and the level as well as the type

of the treatment is adjusted according to the individual's needs. References to this setting include Robins (1997), Lavori et al. (2000), Murphy et al. (2001), Murphy (2003) and Murphy (2005).

The remainder of the paper is organized as follows. Section 2 considers a setting where the treatment outcomes do not depend on observable individual specific characteristics. Next, Section 3 introduces covariates and establishes regret bounds for the sequential treatment policy. When grouping individuals in a specific way, these bounds are near minimax optimal. It is also shown that the expected number of sub-optimal assignments increases slowly and we investigate how to handle discrete covariates. Section 4 investigates the effect of outcomes being observed with delay. In Section 5, we evaluate the finite sample performance of the sequential treatment policy via simulations. Finally, 6 concludes while 7 contains all proofs.

2 The treatment problem without covariates

We begin by considering the sequential treatment problem where the distributions of treatment outcomes do not depend on observable individual specific characteristics. While this setting may often be too restrictive, the regret bounds established in this section will be used as ingredients in establishing the properties of our treatment rules in the setting where covariates are observed on each individual prior to the treatment assignment.

Consider a setting with $K+1$ different treatments and N assignments[1]. K is fixed throughout the paper. N is a random variable whose value need not be known to the policy maker at the beginning of the treatment assignment problem. For example, at the beginning of the year, he does not know how many will become unemployed during the year. Let $Y_t^{(i)} \in [0,1]$ denote the outcome from assigning treatment $i,\ i = 1, ..., K+1$ to individual $t,\ t = 1, ..., N$ where the subscript t indicates the order in which individuals are treated. It is merely for technical reasons that we assume the treatment outcomes to take values in $[0,1]$ and this interval can be generalized to any interval $[I_1, I_2]$ for some $I_1, I_2 \in \mathbb{R},\ I_1 \leq I_2$ or $Y_t^{(i)}$ being sub-gaussian without qualitatively changing our results. The framework accommodates treatments with different costs since, whenever it makes sense, $Y_t^{(i)}$ can be defined net of costs.

We allow for the data to arrive in M batches of sizes $m_b,\ b = 1, ..., M$, such that the total number of assignments is $N = \sum_{b=1}^{M} m_b$. If an unemployment program is run for twelve months and new programs start every month then $M = 12$ and the m_b indicate how many individuals become unemployed in the bth month. The m_b are allowed to be random variables as the policy maker does not a priori know how many will become unemployed each month. This is in contrast to typical treatment allocation problems where the size as well as the composition of the data set are taken as given. Every individual t belongs to exactly one of the batches. For each batch

[1]We consider a setting with $K+1$ treatments for purely notational reasons since it is the number of suboptimal treatments, K, which will enter our regret bounds as well as many of the arguments in the appendix.

the outcomes of the assignments are only observed at the end of the batch. Thus, the treatment assignments for individuals belonging to batch $\tilde{b}$ can only depend on the outcomes observed from previous batches $b = 1, ..., \tilde{b}-1$. Let $B(b) = \sum_{k=1}^{b} m_k$ be the total number of assignments made in the course of the first b batches with the convention $B(0) = 0$. In the setting studied in this paper, the policy maker has no control of the number or size of the batches. For works investigating how to optimally choose the number or size of batches (when this is a choice variable) in the setting where the welfare function depends on the mean only, we refer to Perchet et al. (2016) as well as Gao et al. (2019) who generalize that work.

For each $t = 1, ..., N$ the treatment outcomes can be arbitrarily correlated in the sense that we put no restrictions on the dependence structure of the entries of the vector $Y_t = (Y_t^{(1)}, ..., Y_t^{(K+1)})$, i.e. the joint distribution of the entries of Y_t is left unspecified. This in accordance with real applications where an unemployed individual's response to two types of job training programs may be highly correlated. As individuals arrive independently, we assume the Y_t are i.i.d, which is a standard assumption also in the literature on statistical treatment rules.

The setting described above bears similarity to the one of the classic bandit problem reviewed in Bubeck et al. (2012). However, we consider general *welfare functions* $f : \mathbb{R}^2 \to \mathbb{R}$ of the mean $\mu^{(i)} = \mathbb{E}Y_t^{(i)}$ and variance $(\sigma^2)^{(i)} = \mathbb{E}(Y_t^{(i)} - \mu^{(i)})^2$ of the treatment outcome $Y_t^{(i)}$ [2]. Except for Zimin et al. (2014) this is in contrast to most other work which only considers the expected *effect* of a treatment which amounts to only considering welfare functions depending on the mean. However, it is often very important to also take into account the risk of a decision.

Defining $f^{(i)} = f(\mu^{(i)}, (\sigma^2)^{(i)})$, the *welfare maximizing* (best) treatment is denoted by $*$ and satisfies $f^{(*)} = \max_{1 \leq i \leq K+1} f^{(i)}$ [3]. The welfare maximizing treatment strikes the optimal balance between expected treatment outcome and the riskiness of the treatment. Let $\Delta_i = f^{(*)} - f^{(i)} \geq 0$ be the difference between the best and the ith treatment and assume that $\Delta_1 \geq ... \geq \Delta_K > \Delta_* = 0$. The ranking of the Δ_i is without loss of generality .

A treatment allocation rule is a sequence of functions $\pi = \{\pi_t\}$ assigning a treatment from the set $\{1, ..., K+1\}$ to every individual $t = 1, ..., N$. π_t can take as argument only the treatment outcomes from previous batches. If individual t belongs to batch b, we therefore have that π_t is a mapping from $[0,1]^{B(b-1)}$ into the set of treatments $\{1, ..., K+1\}$. With a slight abuse of notation, we shall often let π_t denote the actual assignment rather than a function of previous treatment outcomes.

Our goal is to provide a rule π that maximizes expected cumulated welfare over the N treatments. This is equivalent to minimizing the expected difference to the infeasible welfare that would have been obtained by always assigning the best treatment $*$, i.e. minimizing the expected value

[2] Since μ and σ^2 both take their values in $[0,1]$, it actually suffices that f is well-defined on $[0,1]^2$.

[3] We assume without loss of generality that the best treatment is unique. If several best treatments are available, the problem would only become easier in the sense that more treatments would yield a zero contribution to regret in (2.1).

of the *regret*

$$R_N(\pi) = \sum_{t=1}^{N} \left(f^{(*)} - f^{(\pi_t)} \right) = \sum_{b=1}^{M} \sum_{i=1}^{m_b} \left(f^{(*)} - f^{(\pi_{b,i})} \right). \tag{2.1}$$

where the second equality is due to the fact that each individual t can be uniquely identified with an assignment i made in a batch b; the assignment rule can also be written as $\pi_{b,i}$ for $b \in \{1, ..., M\}$ and $i \in \{1, ..., m_b\}$. Note that the definition of regret takes into account the welfare of *all* treated individuals. If one is only interested in using the N treated individuals to gain as much information as possible such that one can treat individuals $N + 1$ an onwards as well as possible, then the treatment problem would be of a different nature. In that case the treatment problem would be purely exploratory and other policies than the ones studied in this paper may be optimal. We refer to Bubeck et al. (2009) for a discussion of pure exploration problems in the setting where the policy maker targets the mean only.

2.1 Examples

Throughout this paper we assume that f is Lipschitz continuous from $[0, 1]^2$, equipped with the ℓ_1-norm, to $\mathbb{R}$ with Lipschitz constant $\mathcal{K} > 0$, i.e.

$$|f(u_1, u_2) - f(v_1, v_2)| \leq \mathcal{K} \left(|u_1 - v_1| + |u_2 - v_2| \right),$$

for (u_1, u_2) and (v_1, v_2) in $[0, 1]^2$. By making concrete choices for f our framework contains the following instances as special cases.

1. $f(\mu, \sigma^2) = \mu$ (implying $\mathcal{K} = 1$) amounts to the classic bandit problem where one only targets the mean.

2. $f(\mu, \sigma^2) = \frac{\mu}{\sigma}$ amounts to Sharpe ratios which are frequently used in financial applications to measure risk-return tradeoffs. If $(\sigma^2)^{(i)} \geq c$ for some $c > 0$ for all $i = 1, ..., K + 1$ then one has by the mean value theorem that $\mathcal{K} = \max(\frac{1}{\sqrt{c}}, \frac{1}{2c^{3/2}})$ is a Lipschitz constant for f. Note that f is nonlinear in σ^2 for all μ.

3. $f(\mu, \sigma^2) = -\sigma/\mu$ is the negative of the coefficient of variation in the literature on the measurement of inequality, see Atkinson (1970). The coefficient of variation is a measure of inequality. Minimizing this is encompassed by our framework. Here $\mathcal{K} = \max(\frac{1}{c^2}, \frac{1}{2c^{3/2}})$ is a Lipschitz constant for f if $\mu, \sigma^2 \geq c$ for some $c > 0$.

4. $f(\mu, \sigma^2) = \mu - \alpha\sigma^2$ for a risk aversion parameter $\alpha > 0$ is another typical way of measuring the tradeoff between expected outcomes and their variance. Here $\mathcal{K} = \max(1, \alpha)$ is a Lipschitz constant for f.

5. $f(\mu, \sigma^2) = \mu - 2\alpha\sqrt{\sigma^2}$ for a risk aversion parameter $\alpha > 0$. Here $\mathcal{K} = \max(1, \frac{\alpha}{c^{0.5}})$ is a Lipschitz constant for f if $\sigma^2 \geq c$ for some $c > 0$.

6. $f(\mu, \sigma^2) = -\sigma^2$ (implying $\mathcal{K} = 1$) amounts to the case where one is interested only in minimizing the variance.

7. The theory developed in this paper can be extended to the case where one is interested in maximizing cumulative welfare with a welfare functions depending on any finite number of moments, i.e. $f(\mu_1^{(i)}, ..., \mu_d^{(i)})$ for some $d \geq 1$, where $\mu_k^{(i)} = \mathbb{E}\left[(Y_t^{(i)})^k\right]$ is the $k'th$ moment of $Y_t^{(i)}$. Higher moments than the second one may be relevant if the policy maker has, say, skewness aversion. This is relevant in dynamic portfolio allocation problems and finance as in Harvey and Siddique (2000).

2.2 The sequential treatment policy

Heuristically, the sequential treatment policy works by eliminating treatments that are deemed to be inferior based on the outcomes observed so far. One then take turns assigning each of the remaining treatments in the next batch. This is the exploration step. After this step, elimination can take place again. The policy is inspired by the successive elimination policy which was also studied in Perchet and Rigollet (2013) in a setting where only the expected treatment outcome is targeted.

To describe the policy more precisely, let $m_{i,b} = \sum_{t=B(b-1)+1}^{B(b)} 1_{\{\pi_t = i\}}$ be the number of times treatment i is assigned in batch b. Thus, $m_b = \sum_{i=1}^{K+1} m_{i,b}$ and we define $B_i(b) = \sum_{k=1}^{b} m_{i,k}$ as the number of times treatment i has been assigned up to and including batch b, $b = 1, ..., M$. Whenever $\iota_s := \inf\{t \in \mathbb{N} : \sum_{r=1}^{t} 1_{\{\pi_r = i\}} = s\}$ is finite (meaning that treatment i is eventually assigned at least s times), it is the s-th time treatment i is been assigned by π. We then define $\hat{\mu}_s^{(i)} = \frac{1}{s}\sum_{r=1}^{s} Y_{\iota_r}^{(i)}$ and $(\hat{\sigma}_s^2)^{(i)} = \frac{1}{s}\sum_{r=1}^{s}(Y_{\iota_r}^{(i)} - \hat{\mu}_s^{(i)})^2$.

Sequential treatment policy: Denote by $\hat{\pi}$ the sequential treatment policy. Let $\mathcal{I}_b \subseteq \{1, ..., K+1\}$ be the set of remaining treatments before batch b and let $\underline{B}(b) = \min_{i \in \mathcal{I}_b} B_i(b)$ be the number of times that each remaining treatment at least has been assigned up to and including batch b.

1. In each batch $b = 1, ..., M$ we take turns assigning each remaining treatment. We first assign any treatments that have been assigned fewer times than any of the other remaining treatment(s). Thus, the difference between the number of times that any pair of remaining treatments has been assigned at the end of a batch is at most one.

2. At the end of batch b eliminate treatment $\tilde{i} \in \mathcal{I}_b$ if

$$\max_{i \in \mathcal{I}_b} f(\hat{\mu}_{\underline{B}(b)}^{(i)}, (\hat{\sigma}_{\underline{B}(b)}^2)^{(i)}) - f(\hat{\mu}_{\underline{B}(b)}^{(\tilde{i})}, (\hat{\sigma}_{\underline{B}(b)}^2)^{(\tilde{i})}) \geq 8\gamma\sqrt{\frac{2}{\underline{B}(b)}\overline{\log}\left(\frac{T}{\underline{B}(b)}\right)}$$

where $\gamma > 0,\ T \in \mathbb{N}$ and $\overline{\log}(x) = \log(x) \vee 1$.

The sequential treatment policy uses the sample counterparts of $\mu^{(i)}$ and $(\sigma^2)^{(i)}$ to evaluate whether treatment i is inferior to the best of the remaining treatments. Concrete choices of γ and T guaranteeing a low rate of regret are given in Theorem 2.1 and we provide some intuition here. The parameter γ controls how aggressively treatments are eliminated. Small values of γ make it easier to eliminate inferior treatments but also induce a risk of potentially eliminating the best treatment. The exact form of the elimination threshold comes from the fact the sample moments concentrate at rate $1/\sqrt{\underline{B}(b)}$ around their population counterparts. In the case of two available treatments, the sequential treatment policy is thus reminiscent of A/B-testing with a data-dependent stopping rule. When two treatments are easy to distinguish, the exploration is stopped early while one explores for longer if the treatments are very similar. The parameter T, which will often be set equal to the expected sample size $n = \mathbb{E}(N)$, is needed exactly to ensure that we are cautious eliminating treatments after the first couple of batches where $\hat{\mu}^{(i)}_{\underline{B}(b)}$ and $(\hat{\sigma}^2_{\underline{B}(b)})^{(i)}$ could be based on few observations and thus need not be precise estimates of $\mu^{(i)}$ and $(\sigma^2)^{(i)}$, respectively [4]. From a technical point of view, this ensures that we can uniformly (over treatments) control the probability of eliminating the best treatment. Note that eliminating the best treatment is very costly as regret will accumulate linearly after such a mistake[5].

Remark 1 We note that the sequential treatment policy is different in spirit from the workhorse UCB bandit algorithm in that it eliminates a treatments once one is sufficiently certain that this is not the optimal treatment. The UCB algorithm, on the other hand, does never eliminate any treatment but gradually assigns less promising treatments less frequently. Lai and Robbins (1985), Auer et al. (2002) and Bubeck et al. (2012) provide more information on the UCB policy. As mentioned in the introduction, we refer to Zimin et al. (2014) for a paper that studies a UCB-type policy in our setup. In the discussion of Theorem 2.1 we highlight how our regret guarantee relates to that paper and how the conceptual difference between the sequential treatment policy and UCB manifests itself in different guarantees on maximal expected regret.

2.3 Optimal treatment assignment without covariates

Without an upper bound on the size of the batches it is clear that no non-trivial upper bound on regret can be established. For example, the data could arrive in one batch of size N implying that feedback is never received prior to any assignment. Thus, we shall assume that no batch is larger

[4] We are slightly more cautious than $1/\sqrt{\underline{B}(b)}$. On the other hand, one does not want to be too cautious either since this results in slow elimination of suboptimal treatments.

[5] If the best treatment is eliminated then the regret from each subsequent treatment is $f^{(*)} - f^{(\hat{\pi}_t)} \geq \Delta_K > 0$.

than $\overline{m}$ where $\overline{m}$ is non-random, i.e. $m_b \leq \overline{m}$ for $b = 1, ..., M$. Our first result provides an upper bound on the regret incurred by the sequential treatment policy.

Theorem 2.1 *Consider a treatment problem with $(K+1)$ treatments and an unknown number of assignments N with expectation n that is independent of the treatment outcomes. By implementing the sequential treatment policy with parameters $\gamma = \mathcal{K}$ and $T = n$, one obtains the following bound on the expected regret*

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C \min \left(\underbrace{\overline{m}\mathcal{K}^2 \sum_{i=1}^{K} \frac{1}{\Delta_i} \overline{\log}\left(\frac{n\Delta_i^2}{\mathcal{K}^2}\right)}_{\text{A: Distribution dependent}}, \underbrace{\sqrt{n\mathcal{K}^2\overline{m}K\log(\overline{m}K)}}_{\text{B: Uniform}} \right) \tag{2.2}$$

for a positive universal constant C.

The upper bound in Theorem 2.1 is the minimum of two terms, A and B. The first term, A, depends on the unknown distributional characterisitcs Δ_i and is therefore of a pointwise/adaptive (non-uniform) nature. Note that A only increases logarithmically in the the expected number of treatments n. This logarithmic rate is unimprovable in general since it is known to be optimal even in the case where one only targets the mean (which in our setting corresponds to $f(x, y) = x$) such that $\mathcal{K} = 1$) and the treated individuals arrive one-by-one (such that $\overline{m} = 1$), see e.g. Theorem 2.2 in Bubeck et al. (2012). In the absence of batches ($\overline{m} = 1$), Zimin et al. (2014) also established a pointwise upper bound on the regret of their UCB-type policy of the same order as the one in A. Note, however, that A can be made arbitrarily large by letting, e.g., $\Delta_1 \to 0$. Thus, it is not useful for "small" Δ_i.

The second part (B) of the minimum in (2.2) is uniform over *all* $(K+1)$-tuples of distributions on the Borel sets of $[0, 1]$ as it does not depend on any distributional characteristics. In fact, it yields the minimax optimal rate of expected regret up to a factor of $\sqrt{\log(K)}$ even in the case where only the welfare function $f(x, y) = x$ is considered and $\overline{m} = 1$. In this "mean-only" special case, the factor $\sqrt{\log(K)}$ would be replaced by $\sqrt{\log(n)}$ if one uses UCB, see Section 2.4.3 in Bubeck et al. (2012), rather than the sequential treatment policy (which in this case is akin to the successive elimination policy studied in Perchet and Rigollet (2013)). Thus, as n is much larger than K in most treatment problems, the conceptual difference between the sequential treatment policy and the UCB policy explained in Remark 1 manifests itself in better performance guarantees for the former than for the latter.

It is reasonable that both parts of the upper bound in (2.2) are increasing in $\overline{m}$ since as the maximum batch size increases, the time between potential elimination of suboptimal treatments increases implying that these are assigned more often. Note also that as part of the proof of Theorem 2.1 we show that the probability that the best treatment is ever eliminated is sufficiently low.

This is reminiscent of the work of Howard et al. (2018) who construct certain sequences of confidence sets and show that the probability of the parameter of interest lying outside any of these sets is low. It should be mentioned that we do not know the optimal multiplicative constants in (2.2). However, we stress that the rates in n are minimax optimal. Furthermore, it is sensible that larger $\mathcal{K}$ lead to larger upper bounds on expected regret as larger $\mathcal{K}$ amount to less smooth f and, therefore, more difficult treatment problems.

Remark 2 One may argue that demanding the uniform part of the upper bound on expected regret in Theorem 2.1 to be uniformly valid over all $(K+1)$-tuples of distributions on the Borel sets of $[0,1]$ is too much to ask. In other words, it may be too pessimistic to consider the worst-case performance over all $(K+1)$-tuples of distributions. Hence, it is useful to also know if and how the uniform upper bound on expected regret changes if one restricts the class of distributions. In particular, if it is known that there exists a $\underline{\Delta} > 0$ such that $\Delta_i \geq \underline{\Delta}$ for all $1 \leq i \leq K$ then the expected regret of the sequential treatment policy only increases at a rate of $\log(n)$ uniformly over all $(K+1)$-tuples of distributions on the Borel sets of $[0,1]$ satisfying this constraint. To see this, note that by the $\mathcal{K}$-Lipschitz continuity of f one has that $\Delta_i \leq 2\mathcal{K}$ for $1 \leq i \leq K$. Using this, together with $\Delta_i \geq \underline{\Delta}$ for $1 \leq i \leq K$, in the distribution dependent/pointwise part of the upper of Theorem 2.1 yields that

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C\overline{m}\mathcal{K}^2 \sum_{i=1}^{K} \frac{1}{\underline{\Delta}} \overline{\log}\,(4n) = \frac{C\overline{m}\mathcal{K}^2 K}{\underline{\Delta}} \overline{\log}\,(4n)\,.$$

It follows from Theorem 2.2 in Bubeck et al. (2012) that this is the minimax optimal rate in n over this restricted class since even the pointwise expected regret increases at rate $\log(n)$ (for any consistent policy targeting the mean only). Therefore, irrespectively of whether we consider the maximal expected regret over all $(K+1)$-tuples of distributions on the Borel sets of $[0,1]$ or whether we restrict attention to those distributions satisfying $\Delta_i \geq \underline{\Delta}$, the dependence of the sequential treatment policy in n can not generally be improved.

Remark 3 Note that the implementation of the sequential treatment algorithm requires knowledge of the expected number of individuals that are going to be treated. In medical experiments the total number of individuals participating is often determined a priori making N known and deterministic (and equal to n). On the other hand, when allocating unemployed to treatments, the total number of individuals becoming unemployed in the course of the year is unknown. However, one often has a good estimate of the expected value n which is what matters for the treatment policy. For example, one may use averages of the number of individuals who have become unemployed in previous years to estimate n. Alternatively, one can use the doubling trick, which resets the treatment policy at prespecified times, in order to avoid any assumptions on the size of N or n. Usage of the doubling trick would imply that eliminated treatments reappear and get another chance every time the policy

is reset thus allowing for the efficiency of treatments to vary over time. For further details on the doubling trick and its implementation we refer to Shalev-Shwartz et al. (2012).

2.4 Suboptimal treatments

Theorem 2.1 showed that the expected cumulated welfare of the sequential treatment policy will not be much smaller than the one of the infeasible policy that always assigns the best treatment. However, for an assignment rule to be ethically and politically viable it is important that it does not yield high welfare at the cost of maltreating certain individuals by wild experimentation. For example, it may not be ethically defendable for a doctor to assign a suboptimal treatment to a patient in order to gain more certainty for future treatments. The following theorem shows that the sequential treatment policy does not suffer from such a problem in the sense that the expected number of times *any* suboptimal treatment is assigned only increases logarithmically in the sample size.

Theorem 2.2 *Suppose that the sequential treatment policy is implemented with parameters $T = n$ and $\gamma = \mathcal{K}$. Let $T_i(t)$ denote the number of times treatment i is assigned by the sequential treatment policy up to and including observation t. Then*

$$\mathbb{E}\left[T_i(N)\right] \leq C\left(\mathcal{K}^2 K \frac{\overline{\log}\left(\frac{n}{\mathcal{K}^2}\right)}{\Delta_i^2} + K\overline{m} + \mathcal{K}^2\right),$$

for any suboptimal treatment $i \in \{1, ..., K\}$ and a positive universal constant C.

The important ethical guarantee on the treatment rule is that it only assigns very few persons to a suboptimal treatment (logarithmic growth rate in the sample size). It is in line with intuition that the closer any suboptimal treatment is to being optimal (Δ_i closer to zero) the more difficult it is to guarantee that this treatment is rarely assigned. The reason is that this treatment must be assigned more often before it confidently can be concluded that it is suboptimal and thus eliminated. On the other hand, the regret incurred by assigning such a treatment is low exactly because Δ_i is small.

3 Treatment outcomes depending on covariates

So far we have considered the case where the outcome of a treatment does not depend on the characteristics of the individual it is assigned to. In reality, however, different persons react differently to the same type of treatment: while a certain medicine may work well for one person it may be outright dangerous to assign it to another person if this person is allergic to some of its substances. Similarly, the effect of further education on the probability of an unemployed individual finding a job may also depend on, e.g., the age of the individual: individuals close to the retirement age may

benefit more from short courses updating their skill set while young individuals may benefit more from going back to school for an extended period of time.

Prior to assigning individual t to a treatment we observe a vector $X_t \in [0,1]^d$ of covariates with distribution $\mathbb{P}_X$. In the case of assigning unemployed persons to various unemployment programs, X_t could include age, length of education, and years of experience. $\mathbb{P}_X$ is assumed to be absolutely continuous with respect to the Lebesgue measure λ_d on the Borel sets of $[0,1]^d$ with density bounded from above by $\bar{c} > 0$. This rules out discrete covariates which may be relevant in practice. In Section 3.4 we shall show how policies with low regret in the presence of discrete covariates can be constructed.

As we now observe covariates on each individual prior to the treatment assignment we condition on these. Thus, in close analogy to the setting without covariates, we now define the conditional means and variances $\mu^{(i)}(X_t) = \mathbb{E}(Y_t^{(i)}|X_t)$ and $(\sigma^2)^{(i)}(X_t) = \mathbb{E}\left[(Y_t^{(i)} - \mu^{(i)}(X_t))^2|X_t\right]$ as well as $f^{(i)}(X_t) = f(\mu^{(i)}(X_t), (\sigma^2)^{(i)}(X_t))$. As $\mu^{(i)}(\cdot)$ and $(\sigma^2)^{(i)}(\cdot)$ are unknown to the policy maker, they must gradually be learned by experimentation. In the presence of covariates a policy is a sequence of functions taking as arguments only treatment outcomes and covariates from previous batches as well as the covariates of the current individual to be treated. Thus, if individual t belongs to batch b, then π_t is mapping from $[0,1]^{(d+1)B(b-1)} \times [0,1]^d$ to $\{1, ..., K+1\}$. For any $x \in [0,1]^d$, a social planner (oracle) who knows the conditional mean and variance functions and wishes to maximize welfare assigns [6]

$$\pi^\star(x) := \min \underset{i=1,...,K+1}{\arg\max} f^{(i)}(x)$$

and receives $f^{\left(\pi^\star(x)\right)}(x) = \max_{i=1,...,K+1} f^{(i)}(x) =: f^{(\star)}(x)$. Thus, $f^{(\star)}(x)$ is the pointwise maximum of the $f^{(i)}(x)$, $i = 1, ..., K+1$. The welfare loss (regret) of a policy π compared to the oracle is[7]

$$R_N(\pi) = \sum_{t=1}^{N} \left(f^{\left(\pi^\star(X_t)\right)}(X_t) - f^{(\pi_t)}(X_t)\right) = \sum_{t=1}^{N} \left(f^{(\star)}(X_t) - f^{(\pi_t)}(X_t)\right) \tag{3.1}$$

It is important to note the difference between equation (2.1) and (3.1). While (2.1) considers the difference between unconditional moments, (3.1) considers the difference between conditional moments. The latter is more ambitious as we consider each individual separately through their covariates and seek to minimize the distance to the treatment that would have been optimal for this specific person (with covariates X_t).

In order to prove upper bounds on the regret we restrict the $\mu^{(i)}(\cdot)$ and $(\sigma^2)^{(i)}(\cdot)$ to be reasonably smooth. This is a sensible property to impose since individuals with similar characteristics can

[6] The minimum is here used as an arbitrary tie-breaker in case there are several treatments maximizing $f^{(i)}(x)$.

[7] As π^* only depends on X_t, we keep this explicit while we suppress the arguments of π_t as it depends on all previously observed treatments and the covariates of individual t.

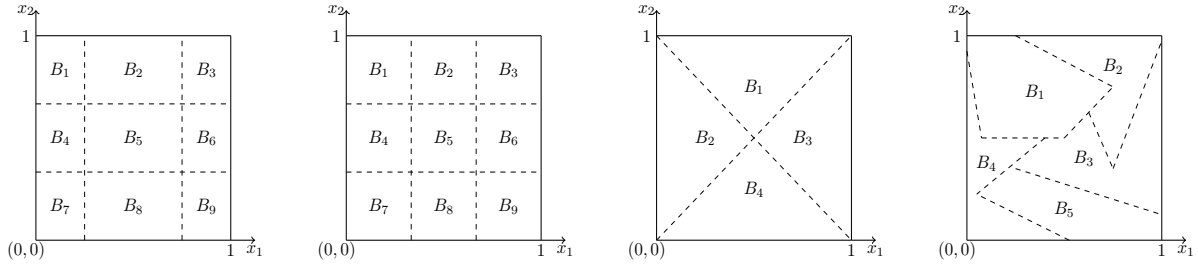

Figure 1: Four examples of partitioning $[0,1]^d$ for $d = 2$. The two leftmost ways of grouping individuals correspond to simple rules where group membership is detmermined by checking whether x_1 and x_2 are above or below certain values. The third rule corresponds to the intersection of two linear eligibility scores $a_i + b_i'x \geq c_i,\ i = 1, 2$. The fourth grouping, though not very practically applicable, serves to illustrate that in principle our theory allows for very general ways of grouping individuals.

be expected to react similarly to the same treatment. In particular, we assume that $\mu^{(i)}(\cdot)$ and $\sigma^{(i)}(\cdot)$ are $(\beta, L)-$Hölder continuous. To be precise, letting $\|\cdot\|$ denote the Euclidean norm on $[0,1]^d$, we assume that $\mu^{(i)},\ (\sigma^2)^{(i)} \in \mathcal{H}(\beta, L)$ for all $i = 1, ..., K+1$, where $\mathcal{H}(\beta, L)$ are those $g : [0,1]^d \to [0,1]$ such that there exist $\beta \in (0,1]$ and $L > 0$ such that

$$\left|g(x) - g(y)\right| \leq L \left\|x - y\right\|^{\beta} \qquad \text{for all } x, y \in [0,1]^d.$$

3.1 Grouping individuals

In the presence of covariates the idea of the sequential treatment policy is to group individuals into groups according to the values of the covariates. Thus, we define a partition of $[0,1]^d$ which consists of Borel measurable sets $B_1, ..., B_F$, called groups/bins, such that $\mathbb{P}_X(B_j) > 0$, $\cup_{j=1}^F B_j = [0,1]^d$, and $B_j \cap B_k = \emptyset$ for $j \neq k$. However, the policy maker may be constrained by political or ethical considerations in his choice of grouping individuals. For example, a realistic unemployment policy cannot group individuals into overly many groups and the rules determining which group an individual belongs to cannot be too complicated. Most realistic policies would choose the groups in such a way that individuals with similar characteristics belong to the same group as it can be expected that the same policy is best for similar individuals. Figure 1 illustrates various ways of grouping individuals. Note that grouping/binning was also considered in Rigollet and Zeevi (2010) as a way to handle continuous covariates in a setting where the policy maker is only interested in the mean of the treatment outcomes. These authors focused on partitions consisting of hypercubes such as the ones in the second display in Figure 1, cf. also equation (3.3) below. We allow general partitions covering also practically relevant eligibility scores such as the ones in the third display of Figure 1. Finally, we also study how to handle discrete covariates in Section 3.4.

For any group B_j define

$$\bar{\mu}_j^{(i)} = \mathbb{E}(Y_t^{(i)}|X_t \in B_j) = \frac{1}{\mathbb{P}_X(B_j)} \int_{B_j} \mu^{(i)}(x) d\mathbb{P}_X(x)$$

and

$$(\bar{\sigma}^2)_j^{(i)} = Var(Y_t^{(i)}|X_t \in B_j) = \mathbb{E}({Y_t^{(i)}}^2|X_t \in B_j) - [\mathbb{E}(Y_t^{(i)}|X_t \in B_j)]^2$$

as the mean and variance of $Y_t^{(i)}$ given that X_t falls in B_j. We apply the sequential treatment policy without covariates separately to each group. To do so, let $f_j^{(i)} = f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ be the welfare of treatment i for the average individual belonging to group j. We use the sequential treatment policy without covariates of Section 2 to target $\max_{1\leq i\leq K+1} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ for each group $j = 1, ..., F$. By the smoothness assumptions on $f, \mu^{(i)}$ and $(\sigma^2)^{(i)}$, $\max_{1\leq i\leq K+1} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ will not be far from the "fully individualized" target $f^{(\star)}(x) = \max_{1\leq i\leq K+1} f(\mu^{(i)}(x), (\sigma^2)^{(i)}(x))$ for any $x \in B_j$ as formalized in the appendix.

Let $N_{B_j}(t) = \sum_{s=1}^{t} 1_{\{X_s \in B_j\}}$ denote the number of individuals who have been assigned to group B_j when t individuals have been treated. Furthermore, $\bar{B}_j = \lambda_d(B_j)$ denotes the Lebesgue measure of group B_j. Let $\hat{\pi}_{B_j, N_{B_j}(t)}$ be the assignment made by the sequential treatment policy without covariates applied only to individuals who belong to group B_j. This policy is implemented with parameters $\gamma = \mathcal{K}L$ and $T = n\bar{B}_j$. The sequential treatment treatment policy $\bar{\pi}$ with covariates is then assigns

$$\bar{\pi}_t(x) = \hat{\pi}_{B_j, N_{B_j}(t)}, \qquad x \in B_j.$$

Thus, when $X_t \in B_j$, individual t is the $N_{B_j}(t)$-th individual falling in group B_j. The sequential treatment policy with covariates then makes the assignment that the policy without covariates applied only to those individuals belonging to group B_j would have made to the $N_{B_j}(t)$-th individual.

3.2 Upper and lower bounds on regret

Denote by $\mathcal{S} = \mathcal{S}(\beta, L, \mathcal{K}, d, \bar{c}, \overline{m})$ a treatment problem where f is Lipschitz continuous with constant $\mathcal{K}$, $X_t \in [0,1]^d$ has distribution $\mathbb{P}_X$ which is absolutely continuous with respect to the Lebesgue measure λ_d with density bounded from above by $\bar{c} > 0$, maximal batch size $\overline{m}$ and $\mu^{(i)}$, $(\sigma^2)^{(i)} \in \mathcal{H}(\beta, L)$ for all $i = 1, ..., K+1$. Unless stated otherwise, we will consider problems in $\mathcal{S}$ in the sequel.

The performance of our policy depends critically on the way the policy maker chooses to group individuals. To characterize this grouping, define $V_j = \sup_{x,y\in B_j} \|x - y\|$ as the maximal possible difference in the characteristics of any two individuals assigned to group j. The next result provides an upper bound on the regret compared to the infeasible oracle which knows $\mu^{(i)}(\cdot)$ and $(\sigma^2)^{(i)}(\cdot)$ and thus whose treatment is optimal for an individual with characteristics $x \in [0,1]^d$.

Theorem 3.1 *Consider a treatment problem in $\mathcal{S}$. Then, for a grouping characterized by $\{V_1, ..., V_F\}$ and $\{\bar{B}_1, ..., \bar{B}_F\}$, expected regret is bounded by*

$$\mathbb{E}\left[R_N(\bar{\pi})\right] \leq C \sum_{j=1}^{F} \left[\sqrt{\overline{m}K \log(\overline{m}K) n \bar{B}_j} + n \bar{B}_j V_j^{\beta}\right] \tag{3.2}$$

for a positive universal constant C. In particular, (3.2) is valid uniformly over $\mathcal{S}$.

Theorem 3.1 provides an upper bound on the regret of the sequential treatment policy for *any* type of grouping of individuals that the policy maker may choose. Allowing for groups with arbitrary characteristics is useful since the policy maker may be constrained in a way such that choosing the groups such that the right hand side of (3.2) is minimized over groups is not possible. The size of the upper bound on expected regret depends on the characteristics $\bar{B}_j$ and V_j of the grouping. Note that the upper bound on the regret is increasing in these two quantities. However, choosing the groups such that $\bar{B}_j$ and V_j are small implies that the number of groups, F, must be large. In general the upper bound in (3.2) cannot be improved by much in terms of their dependence on n since by choosing the groups as in Corollary 3.1.1 below one nearly achieves the minimax rate of regret. We elaborate further on this below.

The first part of the upper bound in (3.2) is the regret accumulated from implementing the sequential treatment policy without covariates on each group separately targeting $\max_{1 \leq i \leq K+1} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ for group $j = 1, ..., F$. The second part of the bound in (3.2) is the approximation error resulting from targeting $\max_{1 \leq i \leq K+1} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ instead of $\max_{1 \leq i \leq K+1} f(\mu^{(i)}(x), (\sigma^2)^{(i)}(x))$.

A particular type of grouping, studied already in Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) in a setting where one targets the mean, is the one consisting of squares constructed from using hard thresholds for each entry of X_t to create hypercubes that partition $[0, 1]^d$. These are particularly relevant in practice due to their simplicity and an example of these bins is given in the second display of Figure 1. More precisely, fix $P \in \mathbb{N}$ and define

$$B_k = \left\{x \in \mathcal{X} : \frac{k_l - 1}{P} \leq x_l \lesssim \frac{k_l}{P},\ l = 1, ..., d\right\} \tag{3.3}$$

for $k = (k_1, ..., k_d) \in \{1, ..., P\}^d$ where $\lesssim$ is to be interpreted as $<$ for $l = 1, ..., d-1$ and $=$ for $l = d$. Thus, P is the number of splits along each dimension of X_t. This creats a partition of P^d smaller hypercubes $B_1, ..., B_{P^d}$ with side lengths $1/P$.

Corollary 3.1.1 *Consider a treatment problem in $\mathcal{S}$. Set $P = \lfloor \left(\frac{n}{\overline{m}K \log(\overline{m}K)}\right)^{1/(2\beta+d)} \rfloor$. Then, expected regret is bounded by*

$$\mathbb{E}\left[R_N(\bar{\pi})\right] \leq Cn \left(\frac{\overline{m}K \log(\overline{m}K)}{n}\right)^{\frac{\beta}{2\beta+d}}. \tag{3.4}$$

for a positive universal constant C. In particular, (3.4) is valid uniformly over $\mathcal{S}$.

Note that the larger the number of covariates d, the smaller will the number of splits P in each dimension be as it must be ensured that enough observations fall in each group. The larger the number of potential treatments $K+1$ is, the more experimentation will take place and hence the regret compared to the infeasible oracle policy increases.

In the case of the policy maker only targeting the mean, such that $\mathcal{K}=1$, and in the absence of batches, such that $\overline{m}=1$, the upper bound on expected regret in Corollary 3.1.1 almost reduces to the one in Theorem 4.1 of Perchet and Rigollet (2013). The only difference is that the exponent $\frac{\beta}{2\beta+d}$ can be replaced by something slightly smaller due to an extra assumption, the *margin condition*, made in Perchet and Rigollet (2013). We shall discuss some consequences of the margin condition in Section 3.3.

Note that the groups $B_1, ..., B_F$ must be chosen a priori by the policy maker. It would also be interesting to study a policy that adaptively chooses the groups as a function of the treatment outcomes observed so far. However, Theorem 3.2 below shows that the current policy is already nearly minimax optimal in terms of its dependence on n. Thus, under this notion of optimality, there is little scope for improvement by altering the sequential treatment policy.

The bound in (3.4) is, as a function of n, not generally improvable in a minimax sense. To see this, it suffices to show that even when the policy maker only targets the mean, no policy can have a much smaller maximal expected regret than the one of the sequential treatment policy. Consider the the case of $\overline{m}=1$ and $K=1$ (two treatments are available) such that (3.4) reduces to $\mathbb{E}\left[R_N(\bar{\pi})\right] \leq Cn^{1-\frac{\beta}{2\beta+d}}$.

Theorem 3.2 *Let $f(\mu,\sigma^2)=\mu$, $\overline{m}=1$ and $K=1$. Then for any $\varepsilon>0$ there exists a constant $C(\varepsilon)$ such that for any policy π*

$$\sup_{\mathcal{S}} \mathbb{E}\left[R_N(\pi)\right] \geq C(\varepsilon) n^{1-\frac{\beta}{2\beta+d}} \cdot n^{-\varepsilon}$$

Theorem 3.2 shows that the upper bound on maximal expected regret of the sequential treatment policy in Corollary 3.1.1 can not generally be improved by much. In particular, the order of such an improvement must be $o(n^{\varepsilon})$ for any $\varepsilon>0$, e.g. logarithmic.

3.3 Ethical considerations

We next show that even in the presence of covariates the sequential treatment policy does not make many suboptimal assignments. Our first result is a consequence of Theorem 2.2. On any bin $1 \leq j \leq F$ the result bounds the number of times that a treatment $1 \leq i \leq K+1$ which does not maximize $f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ is assigned. Let $T_{i,j}(N)$ be the number of times treatment i is assigned on bin j in the course of a total of N assignments. Calling treatment i *suboptimal on bin* B_j if $\Delta_{i,j} := f_j^{(*)} - f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)}) > 0$ we have the following result.

Theorem 3.3 *Consider a treatment problem in $\mathcal{S}$. Then, for group B_j characterized by V_j and $\bar{B}_j$,*

$$\mathbb{E}\left[T_{i,j}(N)\right] \leq C\left(\mathcal{K}^2 K \frac{\log\left(\frac{n\bar{B}_j}{\mathcal{K}^2}\right)}{\Delta_{i,j}^2} + K\overline{m} + \mathcal{K}^2\right),$$

for any treatment i that is suboptimal on bin B_j and a positive universal constant C.

Theorem 3.3 guarantees that any treatment whose combination of mean and variance over B_j does not maximize f will only rarely be assigned. In fact, the number of times a treatment that is suboptimal on bin B_j is assigned only grows logarithmically in the expected number of individuals belonging to bin B_j. Notice the similarity to Theorem 2.2 where n has now been replaced by $n\bar{B}_j$ which up to the constant $\bar{c}$ is an upper bound on the expected number of individuals falling in group j.

A potential shortcoming of Theorem 3.3 is that for each group B_j, the maximizer of $f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ depends on the way the policy maker has chosen B_j. A different way of assessing the number of suboptimal treatments assigned is to consider each person individually and check whether the optimal treatment was assigned to this person. We say that treatment i is *suboptimal for individual t* if $f^{(\star)}(X_t) > f^{(i)}(X_t)$. Therefore, another way of declaring the fairness of a policy π is to provide an upper bound on the number of individuals

$$S_N(\pi) = \sum_{t=1}^{N} 1_{\{f^{(\star)}(X_t) \neq f^{(\pi_t)}(X_t)\}},$$

to whom a suboptimal treatment was assigned. It is sensible that a nontrivial upper bound on $\mathbb{E}(S_N(\pi))$ (a bound less than n) can only be established if the best treatment is sufficiently much better than the second best — otherwise these cannot be distinguished from each other. To formalize this notion let

$$f^{(\sharp)}(x) = \begin{cases} \max_{i=1,\dots,K+1}\{f^{(i)}(x) : f^{(i)}(x) < f^{(\star)}(x)\} & \text{if } \min_{i=1,\dots,K+1} f^{(i)}(x) < f^{(\star)}(x) \\ f^{(\star)}(x) & \text{otherwise} \end{cases}$$

denote the value of the second best treatment for an individual with characteristics $x \in [0,1]^d$.

Assumption 1 (Margin condition) *We say that the margin condition is satisfied with parameter $\alpha > 0$ if there exists a constant $C > 0$ and a $\delta_0 \in (0,1)$ such that*

$$\mathbb{P}\left(0 < f^{(\star)}(X_t) - f^{(\sharp)}(X_t) < \delta\right) \leq C\delta^{\alpha} \qquad \forall \delta \in (0, \delta_0]$$

The margin condition limits the probability that the best and the second best treatment are very close to each other. Larger values of α mean that it is easier to distinguish the best and second best treatment from each other. The margin condition has been used in the literature on statistical

treatment rules by Kitagawa and Tetenov (2018) to improve the rates of their empirical welfare maximization classifier. Before this, similar assumptions had been used in the literature on classification, Mammen et al. (1999), Tsybakov (2004b). Perchet and Rigollet (2013) and Rigollet and Zeevi (2010) have used the margin condition in the context of bandits. The margin condition is satisfied if, for example, $f^{(\star)}(X_t) - f^{(\sharp)}(X_t)$ has a density with respect to the Lebesgue measure which is bounded from above by a constant $a > 0$. In that case we may set $C = a$ and $\alpha = 1$.

Theorem 3.4 *Fix $\beta \in (0, 1]$, $\mathcal{K}, L > 0$, $d \geq 2$ and consider a treatment problem in $\mathcal{S}$ which also satisfies the margin condition. Then for any policy π,*

$$\mathbb{E}(S_N(\pi)) \leq Cn^{\frac{1}{1+\alpha}} \mathbb{E}\left[R_N(\pi)\right]^{\frac{\alpha}{1+\alpha}} \tag{3.5}$$

for a positive universal constant C. Using the sequential treatment policy $\bar{\pi}$ and grouping individuals as in (3.3) yields

$$\mathbb{E}(S_N(\pi)) \leq Cn \left[\frac{\overline{m}K \log(\overline{m}K)}{n}\right]^{\frac{\alpha\beta}{(1+\alpha)(2\beta+d)}}. \tag{3.6}$$

(3.5) provides an upper bound on the expected number of times a policy π assigns a treatment which is suboptimal for individual t. This is done in terms of the regret incurred by the policy and is an extension of Lemma 3.1 in Rigollet and Zeevi (2010) to the setting allowing the policy maker to target Lipschitz continuous functions of the mean and variance. (3.6) considers the case of the sequential treatment policy with a particular group structure and follows by using the upper bound on expected regret from Corollary 3.1.1 in (3.5). Note that $\mathbb{E}(S_N(\pi))$ is guaranteed to grow only sublinearily in n. It is likely that (3.6) can be slightly improved by also using the margin condition to sharpen the upper bound of Corollary 3.1.1. However, this is beyond the scope of this paper as we mostly consider a setting without the margin condition.

3.4 Discrete covariates

Until now we have assumed $\mathbb{P}_X$ to be absolutely continuous with respect to the Lebesgue measure λ_d on the Borel sets of $[0, 1]^d$. However, many covariates that may influence the identity of the optimal treatment are discrete. For example, gender may affect the outcome of an allocation in an unemployment program. Furthermore, we may not always observe a continuous variable perfectly as data might only be informative about which of finitely many wealth groups an individual belongs to without providing the exact, continuously scaled, wealth. Since a combination of continuous and discrete covariates has to our knowledge not been studied even in the case where one is only interested in the mean, we also contribute to that setting.

In order to accommodate discrete covariates, partition $X_t = (X'_{t,D}, X'_{t,C})'$ where $X_{t,D} \in A = A_1 \times ... \times A_{d_D}$ contains the measurements of the d_D discrete covariates. Each $A_l \subseteq \mathbb{N}$, $l = 1, ..., d_D$

is finite with cardinality $|A_l|$. For the continuous covariates we assume $X_{t,C} \in [0,1]^{d_C}$ such that X_t is $(d_D + d_C)$-dimensional. As in (3.1) the regret of our treatment policy is measured against the infeasible target $f^{(\star)}(X_t) = \max_{1 \leq i \leq K+1} f(\mu^{(i)}(X_t), (\sigma^2)^{(i)}(X_t))$. On the other hand, it does not make sense to assume $\mu^{(i)}(x) = \mu^{(i)}(x_D, x_C)$ or $(\sigma^2)^{(i)}(x) = (\sigma^2)^{(i)}(x_D, x_C)$ to be $(\beta, L)-$Hölder continuous in x_D. Thus, discrete covariates must be handled differently from continuous ones. Instead we shall now assume that for each *fixed* $a \in A$ one has that $\mu_a^{(i)}(x_C) := \mu^{(i)}(a, x_C)$ and $(\sigma^2)_a^{(i)}(x_C) := (\sigma^2)^{(i)}(a, x_C)$ belong to $\mathcal{H}(\beta, L)$. Since a can only take $F_D = |A| = |A_1| \cdot ... \cdot |A_{d_D}|$ possible values it is without loss of generality to assume β and L not to depend on a.

Our treatment policy now works by fully individualizing treatments across the discrete covariates. In other words, for any of the F_D possible values of the vector of discrete covariates, we implement the sequential treatment policy $\bar{\pi}$ by constructing groups only based on the continuous variables just as in Section 3.1. For each value of the discrete covariate we allow for different ways of grouping based on the continuous covariates. For example, one may want to construct different wealth groups for men and women in order to obtain, e.g., groups with equally many individuals. For each $a \in A$ let $B_{a,j},\ j = 1, ..., F_a$ be the partition of $[0,1]^{d_C}$ used.

Formally, for each $a \in A$, let $\bar{\pi}_{t,a}$ be the sequential treatment policy with continuous covariates applied to the grouping $B_{a,j},\ j = 1, ..., F_a$. Thus, if individual t belongs to bin b, the sequential treatment policy in the presence of discrete covariates, $\tilde{\pi}$, is a mapping $\tilde{\pi}_t : (A_1 \times ... \times A_{d_D} \times [0,1]^{(d_C+1)})^{B(b-1)} \times A_1 \times ... \times A_{d_D} \times [0,1]^{d_C} \to \{1, ..., K+1\}$ where

$$\tilde{\pi}_t(x) = \bar{\pi}_{t,a}(x_C) = \hat{\pi}_{(\{a\} \times B_{a,j}), N_{a,j}(t)}, \qquad x_D = a \text{ and } x_C \in B_{a,j}$$

with $N_{a,j}(t) = \sum_{s=1}^{t} 1_{\left(X_{s,D} = a, X_{s,C} \in B_{a,j}\right)}$. Denote by $\tilde{\mathcal{S}} = \tilde{\mathcal{S}}(\beta, L, \mathcal{K}, d_C, \bar{c}, \overline{m})$ a treatment problem where f is Lipschitz continuous with constant $\mathcal{K}$, $X_{t,D} \in A$ is discrete, $X_{C,t} \in [0,1]^d$ has distribution $\mathbb{P}_X$ which is absolutely continuous with respect to the Lebesgue measure with density bounded from above by $\bar{c}$, maximal batch size $\overline{m}$ and $\mu_a^{(i)},\ (\sigma^2)_a^{(i)} \in \mathcal{H}(\beta, L)$ for all $i = 1, ..., K+1$ and $a \in A$. Letting $V_{a,j} = \sup_{x,y \in B_{a,j}} ||x - y||$ we have that $\tilde{\pi}$ enjoys the following upper bound on regret.

Theorem 3.5 *Consider a treatment problem in $\tilde{\mathcal{S}}$. Then, if for for each $a \in A$ individuals are grouped as $\{B_{a,1}, ..., B_{a,F_a}\}$, expected regret is bounded by*

$$\begin{aligned}\mathbb{E}\left[R_N(\tilde{\pi})\right] \leq & C \sum_{a \in A} \sum_{j=1}^{F_a} \Bigg(\sqrt{\overline{m} K \log(\overline{m} K) n \mathbb{P}(X_{t,D} = a, X_{t,C} \in B_{a,j})} \\ & + n \mathbb{P}(X_{t,D} = a, X_{t,C} \in B_{a,j}) V_{a,j}^{\beta} \Bigg). \end{aligned} \tag{3.7}$$

for a positive universal constant C. In particular, (3.7) is valid uniformly over $\tilde{\mathcal{S}}$.

The upper bound on regret in (3.7) generalizes the upper bounds in Theorems 2.1 (no covariates) and 3.1 (continuous covariates only). For example, the latter follows from (3.7) by letting $|A| = 1$

and using that $X_{t,C}$ is absolutely continuous with respect to the Lebesgue measure with density bounded from above by $\bar{c}$. Also, the case of purely discrete covariates is covered as a special case of (3.7). In that case the approximation error vanishes as $V_{a,j} = 0$.

4 Treatment outcomes observed with delay

Oftentimes the outcome of a treatment is only observed with delay. For example, a medical doctor may choose not to measure the effect of a treatment immediately after it has been assigned as it takes time for the treatment to work. However, delaying the measurement for an extended period of time also implies that many new patients will arrive before the outcome of the previous treatment is known. Thus, the type of treatment assigned to these patients must be decided based on less information. Put differently, there is a tradeoff between getting imprecise information now and obtaining precise information later. A similar tradeoff exists when assigning unemployed to job training programs as it takes time to find a job. Therefore, it may not be advisable to measure the effect of a job training program very shortly after its termination.

We note that the role of delayed observation has also been studied Joulani et al. (2013). However, the focus there is solely on the negative effect of delay through postponed information accrual. The potential for obtaining more precise information through delayed measurements as described in the previous paragraph is not considered. Thus, no tradeoff in when to make the measurement is present. In addition, the focus is on the mean-only setting without covariates. For a recent paper studying adversarial bandits with delay, we refer to Cesa-Bianchi et al. (2019).

In this section we formalize the above intuition by proposing the following model for treatments being observed with delay. For simplicity, we focus first on the setting without covariates. We can decompose $Y_t^{(i)}$ as

$$Y_t^{(i)} = \mu^{(i)} + \eta_t^{(i)}$$

where $\mathbb{E}(\eta_t^{(i)}) = 0$. Since $Y_t^{(i)}, \mu^{(i)} \in [0,1]$ it follows that $\eta_t^{(i)} = Y_t^{(i)} - \mu^{(i)} \in [-1,1]$. Thus, without further assumptions, the deviations of $Y_t^{(i)}$ around its mean are in $[-1,1]$. We shall model the idea of measurements becoming more precise if they are delayed by restricting this interval. To be precise, we assume that

$$\eta_t^{(i)} = Y_t^{(i)} - \mu^{(i)} \in [-\bar{a}_l, \bar{a}_u] \tag{4.1}$$

where $\bar{a}_l, \bar{a}_u \in [0,1]$. In this section we let $\bar{a}(D) = \bar{a}_u(D) - \bar{a}_l(D)$ be a function of the number of batches D the measurements are delayed by. Thus, if $\bar{a}(D)$ is a decreasing function, increasing the delay results in $Y_t^{(i)}$ being a less noisy measure of $\mu^{(i)}$. Restricting the support of $\eta_t^{(i)}$ is not the only way of modelling that measurements become more precise if they are delayed. One could also let the variance of the $\eta_t^{(i)}$ be a decreasing function of D. In fact, any assumption which implies

stronger concentration of sample averages around the population means will suffice. As the welfare function f also depends on the second moment $\mu_2^{(i)} = \mathbb{E}\left[{Y_t^{(i)}}^2\right]$ and since ${Y_t^{(i)}}^2, \mu_2^{(i)} \in [0,1]$ we will model increased measurement precision of second moments due to delay as[8]

$$ {Y_t^{(i)}}^2 - \mu_2^{(i)} \in [-\bar{a}_l, \bar{a}_u] \tag{4.2} $$

First, we establish an upper bound on expected regret of the sequential treatment policy when treatment outcomes are observed with delay in the absence of covariates.

Sequential treatment policy Denote by $\hat{\pi}$ the sequential treatment policy. Let $\mathcal{I}_b \subseteq \{1, ..., K+1\}$ be the set of remaining treatments before batch b and let $\underline{B}(b) = \min_{i \in \mathcal{I}_b} B_i(b)$ be the number of outcomes that have been observed for each of the remaining treatments after batch b.

1. In each batch $b = 1, ..., D-1$ we take turns assigning the treatments $\{1, ..., K+1\}$. No elimination takes place as no outcomes are observed.

2. In each batch $b = D, ..., M$ we take turns assigning each remaining treatment (treatments in $\mathcal{I}_b$).

3. At the end of batch $b = D, ..., M$ eliminate treatment $\tilde{i} \in \mathcal{I}_b$ if

$$ \max_{i \in \mathcal{I}_b} f(\hat{\mu}^{(i)}_{\underline{B}(b)}, (\hat{\sigma}^2_{\underline{B}(b)})^{(i)}) - f(\hat{\mu}^{(\tilde{i})}_{\underline{B}(b)}, (\hat{\sigma}^2_{\underline{B}(b)})^{(\tilde{i})}) \geq 8\gamma \sqrt{\frac{2\bar{a}^2}{\underline{B}(b)} \overline{\log}\left(\frac{T}{\underline{B}(b)}\right)} $$

   where $\gamma > 0,\ T \in \mathbb{N}$ and $\overline{\log}(x) = \log(x) \vee 1$.

Notice how the sequential treatment policy in the presence of delay differs from the one without delay. First, no elimination takes place after the first $D-1$ batches as no treatment outcomes are observed after these. Second, the elimination rule has been slightly modified as we can now eliminate more aggressively if $\bar{a}$ is small, i.e. the treatment outcomes are less noisy measurements of the population parameters.

Theorem 4.1 (No covariates) *Consider a treatment problem with $(K+1)$ treatments and an unknown number of assignments N with expectation n that is independent of the treatment outcomes. The treatment outcomes are observed with a delay of D batches as outlined above. By implementing the sequential treatment policy with parameters $\gamma = \mathcal{K}$ and $T = n$ one obtains the following bound on the expected regret*

$$ \mathbb{E}\left[R_N(\hat{\pi})\right] \leq $$

[8] Assuming the same lower and upper bounds in (4.1) and (4.2) is without loss of generality as one can simply take the smallest of the lower bounds and the largest of the upper bounds as the common lower and upper bound.

$$C \min \left(\underbrace{\mathcal{K}^2 \bar{a}^2 \sum_{i=1}^{K} \frac{1}{\Delta_i} \overline{\log} \left(\frac{n\Delta_i^2}{\bar{a}^2} \right) + \overline{m}(K+D)}_{\textit{A: Distribution dependent}}, \underbrace{\sqrt{\mathcal{K}^2 \bar{a}^2 \overline{m} K \overline{\log}\left(\overline{m}K\right) n} + \overline{m}(K+D)}_{\textit{B: Uniform}} \right), \tag{4.3}$$

for a positive universal constant C.

Assume that $\bar{a} = \bar{a}(D)$ is a decreasing function of D. Then Theorem 4.1 illustrates the tradeoff between getting imprecise information now and precise information later. This tradeoff is found in the distribution dependent part (A) as well as the uniform part (B) of the upper bound on regret of the sequential treatment policy. Increasing D directly increases the upper bound on regret since information is obtained later but indirectly decreases the regret via a reduced $\bar{a}$. By making a concrete choice for $\bar{a}(D)$ one can determine the optimal delay by minimizing the upper bound on regret. It can also be shown that the bound in Theorem 4.1 reduces to the one in Theorem 2.1 when $D = 0$ and $\bar{a} = 1$.

Joulani et al. (2013) (Theorem 7) provide adaptive/pointwise upper bounds on the expected regret of UCB which like our bounds have the property that the downside of delay enters additively. However, only targeting the mean is considered and there is no potential for obtaining more precise information by delaying measurements. This corresponds to the special case of $\overline{a}_u$ and $\overline{a}_l$, and hence $\overline{a}$, not depending on D in our setting. No uniformly valid upper bound on expected regret is provided.

We turn next to the setting with continuous covariates and treatment outcomes being observed with delay. To be precise, we assume that

$$Y_t^{(i)} - \mu_1^{(i)}(X_t), {Y_t^{(i)}}^2 - \mu_2^{(i)}(X_t) \in [-\bar{a}_l, \bar{a}_u],$$

where $\mu_1^{(i)}(X_t) = \mathbb{E}\left[Y_t^{(i)} | X_t\right]$ and $\mu_2^{(i)}(X_t) = \mathbb{E}\left[{Y_t^{(i)}}^2 | X_t\right]$. As in the setting without delay, we implement the sequential treatment policy separately for each group $B_1, ..., B_F$ with parameters $\gamma = \mathcal{K}L$ and $T = n\bar{B}_j,\ j = 1, ..., F$.

Theorem 4.2 *Fix $\beta \in (0, 1]$, $\mathcal{K}, L > 0$, $d \geq 2$ and consider a treatment problem in $\mathcal{S}$ where the outcomes are observed with a delay of D batches. Then, for a grouping characterized by $\{V_1, ..., V_F\}$ and $\{\bar{B}_1, ..., \bar{B}_F\}$, expected regret is bounded by*

$$\mathbb{E}\left[R_N(\bar{\pi})\right] \leq C \left(\sum_{j=1}^{F} \left[\sqrt{\overline{m}K\bar{a}^2 \log(\overline{m}K) n\bar{B}_j} + n\bar{B}_j V_j^{\beta} + K\overline{m} \right] + \overline{m}D \right). \tag{4.4}$$

for a positive universal constant C. In particular, (4.4) is valid uniformly over $\mathcal{S}$.

The first part of the upper bound on expected regret in (4.4) (the sum over the F groups) is identical to the upper bound in Theorem 3.1 except for the presence of $\bar{a}$. The smaller $\bar{a}$ is the smaller will this part be as the observed outcomes of the treatments will be very close to the population counterparts and the treatment that is best for each group is quickly found. As $\bar{a}$ is usually a decreasing function in D, the upper bound in (4.4) clearly illustrates the tradeoff between postponing the measurement to get precise information later and getting (imprecise) information quickly. The term under the square root holds the key to the benefit from delaying as it corresponds to the regret of a treatment problem which starts only after D batches but where measurements are observed more precisely. The term $\overline{m}D$ is an upper bound on the regret incurred from assigned individuals blindly for D batches each of which contains no more than $\overline{m}$ individuals.

5 Simulations

The theoretical results derived in the previous sections provide strong guarantees regarding the expected regret incurred by the sequential treatment policy. In this section, we illustrate some of these results by a simulation study. We compare the performance of the the sequential treatment policy to that of a traditional two-step policy which i) first explores the available treatments and for a predetermined period ii) then commits to the one that performed empirically best during the exploration phase. That is, a policy which uses an empirical success type commitment rule after the exploration phase.

The data is generated by drawing samples of size n from the beta distribution with parameters α and β taking values in $(0, \infty)$. The beta family of distributions is rather flexible, and by varying the parameters used it is possible to generate many different shapes of densities. To evaluate the worst-case (uniform) performance guarantee provided in 2.1, we calculate the expected regret of over a large class of beta distributions and then consider the maximum of these expected regrets. To be precise, we consider a setup with $K = 2$ treatments. The outcome of these treatments are beta distributed with $\alpha_i = 1$ and β_i taking a value in $\mathcal{G} = \{0.5, 0.75, 0.85, 0.95, 0.975, 1, 1.025, 1.05, 1.15, 1.25, 2.5\}$ for $i = 1, 2$. As the regret is zero when the two treatments have the same beta distribution, we only consider outcome distributions of the form $Beta(1, \beta_1) \otimes Beta(1, \beta_2)$ for distinct β_1 and β_2 in $\mathcal{G}$. This leaves us with 55 different joint treatment outcome distributions over which we study the worst case performance. The second ingredient that is required for the proposed methodology is the welfare function f. We consider two different welfare functions: The mean, $f(\mu, \sigma^2) = \mu$, and a mean-variance utility function, $f(\mu, \sigma^2) = \mu - \frac{\gamma}{2}\sigma^2$, with risk aversion parameter γ equal to 1. Note that the mean is already covered by classic bandit theory and we include it here merely for reference.

Note that intuitively elimination can take place earlier when the two outcome distributions of the two treatments are far apart while one may want to explore for longer before eliminating if

the distributions are close to each other. The sequential treatment policy does not require any prior knowledge regarding the similarity of the outcome distributions but instead monitors one the difference between the empirical welfare functions becomes significant such that elimination can take place. This is in contrast to the rule that decides up front on the length of the exploration period.

We consider 5 sample sizes, n, namely $20,000, 40,000, 60,0000, 80,000,$ and $100,000$. To gauge the impact of batching, we consider three different scenarios, namely one in which individuals arrive one-by-one, and another two where they arrive in batches of size $1,000$ or $5,000$, respectively. Note that with a batch size of $5,000$, the first time it is possible to eliminate a treatment is at time $t = 5,000$ (no matter how far the outcome distributions are apart). For each joint outcome distribution $Beta(1,\beta_1) \otimes Beta(1,\beta_2)$, sample size and batch size we calculate the regret 100 times in order to approximate the expected regret at time n. We then take the maximum over all 55 potential joint outcome distributions in order to estimate the maximal expected regret.

In Table 1 we report the maximal (across joint outcome distributions) expected regret for the two policies. This is done separately for the two welfare functions and batching schemes. For the traditional two-step policy, we consider length of exploration phases, i.e. decision points equaling, $n/4, n/3$ and $n/2$. This way, the batch size is always no larger than the exploration phase.

	$f(\mu,\sigma^2)=\mu$					$f(\mu,\sigma^2)=\mu-\frac{1}{2}\sigma^2$				
$n:$	20,000	40,000	60,000	80,000	100,000	20,000	40,000	60,000	80,000	100,000
Sequential elimination policy										
No batching	255.53	366.51	437.28	493.59	565.34	1068.43	1560.43	1899.48	2193.92	2455.15
Batch size: 1,000	265.79	370.44	451.81	510.62	564.90	1088.21	1569.10	1898.77	2192.72	2471.34
Batch size: 5,000	952.38	952.38	952.38	952.38	952.38	1319.32	1777.96	1945.63	2223.32	2473.60
Traditional policy										
Decision point $n/4$	952.38	1904.76	2857.14	3809.52	4761.90	897.96	1795.92	2693.88	3591.84	4489.80
Decision point $n/3$	1269.71	2539.43	3809.52	5079.24	6348.95	1197.16	2394.32	3591.84	4789.00	5986.16
Decision point $n/2$	1904.76	3809.52	5714.29	7619.05	9523.81	1795.92	3591.84	5387.76	7183.67	8979.59

Table 1: The table reports the maximal average regret of the sequential treatment policy and of a traditional policy that picks the treatment that maximizes the empirical counterpart of the objective function after an exploraton period of a predetermined length.

A couple of findings are worth highlighting. First, the sublinear increase in maximal expected regret as function of the sample size predicted by the uniform part of the upper bound in Theorem 2.1 is confirmed. In fact, for both welfare functions the maximal expected regret only increases by a factor of a little over 2 even when the sample size is increased five fold from 20,000 to 100,000.

Secondly, batching the observation of outcomes into groups of 1,000 individuals only has a limited impact on the performance of the policy. This is due to the fact that for those joint distributions

attaining the maximal expected regret, no treatment had been eliminated after 1,000 assignments even in the case without batching. Thus, batching constraint is not binding.

Thirdly, since the decision to eliminate a treatment can only be made after the arrival of a new batch of observed outcomes, the maximal expected regret is increasing in the batch size. The cost of increased batch sizes becomes more pronounced the larger the difference in expected outcomes is. The clearest illustration hereof can be found in row 3 of Table 1. Prior to the arrival of the first batch of information with sufficient evidence to eliminate a treatment, regret increases linearly (as every second assignment is suboptimal), and while a decision is made when the outcomes become available, numerous suboptimal treatments with a large associated regret have already been made. In fact, this initial accumulation of regret outweighs the total regret accumulated by the policy over any of the other possible distributions of the outcomes even when considering as many as five times the number of treated individuals. Note, however, that this is not an issue that is unique to the sequential treatment policy. In fact, the traditional RCT setup involves an initial exploration phase, and would thus also accumulate a large amount of regret due to not making a decision earlier. This is captured in rows 4-6 of the table. Maximal expected regret increases linearly in the sample size, and all of this regret is essentially accumulated during the initial exploration phase.

Relative to the sequential elimination policy, the traditional empirical success type policy results in more than 3.5 times as much regret when considering the mean as the outcome of interest. The gap is smaller when the mean-variance framework is considered. In fact, when the sample size only is 20,000 individuals, the sequential elimination policy accumulates a slightly higher maximal expected regret. This is because the elimination rule has to be conservative enough to ensure that the probability of eliminating the optimal treatment is uniformly controlled across all possible joint distributions of the outcomes.

6 Conclusions

This paper considers a treatment allocation problem where the individuals to be treated arrive gradually and potentially in batches. The goal of the policy maker is to maximize the welfare over the treatment assignments made. As the policy maker does not know a priori about the virtues of the available treatments, he faces an exploration-exploitation tradeoff. Prior to each assignment she observes covariates on the individual to be treated thus allowing for the optimal treatment to vary across individuals. Our setup allows the welfare function not only to depend on the expected treatment outcome but also on the risk of the treatment. We show that the sequential treatment policy can not generally be improved in the sense that it attains the minimax optimal expected rate of regret in n in the special case of targeting the mean. This strong welfare guarantee does not come at the price of overly wild experimentation as we show that the number of suboptimal treatments only grows quite slowly in the total number of assignments made. We also establish

upper bounds on the regret of the sequential treatment policy when the outcome of the treatments are observed with delay.

7 Appendix

Throughout the appendix we let $C > 0$ be a constant that may change from line to line. However, it does not depend on any distributional characteristics of the treatment problem.

7.1 Proof of Theorems 2.1 and 2.2

The following lemma will lead to Theorem 2.1. Its proof is structured as the one of Theorem 2.1 in Perchet and Rigollet (2013).

Lemma 7.1 *Consider a treatment problem with $(K+1)$ treatments and unknown number of assignments N with expectation n that is independent of the treatment outcomes. Suppose that f is Lipschitz continuous with known constant $\mathcal{K}$. For any $\Delta > 0$, $T > 0$ and $\gamma \geq \mathcal{K}$ the expected regret from running the sequential treatment policy can then be bounded as*

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C\left(\frac{\gamma^2 K}{\Delta}\left(1+\frac{n}{T}\right)\overline{\log}\left(\frac{T\Delta^2}{288\gamma^2}\right) + n\Delta^- + \frac{n\overline{m}K}{T}\right), \tag{7.1}$$

where Δ^- is the largest Δ_j such that $\Delta_j < \Delta$ if such a Δ_j exists, and $\Delta^- = 0$ otherwise.

Proof. Define $\epsilon_s = u(s,T) = 8\sqrt{\frac{2}{s}\overline{\log}\left(\frac{T}{s}\right)}$ and $\hat{\Delta}_i(s) = f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\hat{\mu}_s^{(i)}, (\hat{\sigma}_s^2)^{(i)})$. Recall that if the optimal treatment as well as some treatment i have not been eliminated before batch b (i.e., $i, * \in \mathcal{I}_b$), then the optimal treatment will eliminate treatment i if $\hat{\Delta}_i(\underline{B}(b)) \geq \gamma\epsilon_{\underline{B}(b)}$, and treatment i will eliminate the optimal treatment if $\hat{\Delta}_i(\underline{B}(b)) \leq -\gamma\epsilon_{\underline{B}(b)}$.

To say something about when either of these two events occurs we introduce the quantity τ_i^* which is defined through the relation

$$\Delta_i = 12\gamma\sqrt{\frac{2}{\tau_i^*}\overline{\log}\left(\frac{T}{\tau_i^*}\right)}, \qquad i = 1, ..., K.$$

Since τ_i^* in general will not be an integer, we also define $\tau_i = \lceil \tau_i^* \rceil$. Next introduce the hypothetical batch $b_i = \min\{l : \underline{B}(l) \geq \tau_i^*\}$. It is the first batch after which we have more than τ_i^* observations on all remaining treatment. Notice that

$$\tau_i^* \leq \underline{B}(b_i) \leq \tau_i^* + \overline{m} \leq C\frac{\gamma^2}{\Delta_i^2}\overline{\log}\left(\frac{T\Delta_i^2}{288\gamma^2}\right) + \overline{m}, \tag{7.2}$$

$$\tau_i \leq \underline{B}(b_i), \tag{7.3}$$

$$\underline{B}(b_i) \leq \tau_i + \overline{m}, \tag{7.4}$$

Notice that $1 \leq \tau_1 \leq ... \leq \tau_K$ and $1 \leq b_1 \leq ... \leq b_K$. Define the following events:

$$\begin{aligned}
\mathcal{A}_i &= \{\text{The optimal treatment has not been eliminated before batch } b_i\}, \\
\mathcal{B}_i &= \{\text{Every treatment } j \in \{1, ..., i\} \text{ has been eliminated after batch } b_j\}.
\end{aligned}$$

Furthermore, let $\mathcal{C}_i = \mathcal{A}_i \cap \mathcal{B}_i$, and observe that $\mathcal{C}_1 \supseteq ... \supseteq \mathcal{C}_K$. For any $i = 1, ..., K$, the contribution to regret incurred after batch b_i is at most $\Delta_{i+1} N$ on $\mathcal{C}_i$. In what follows we fix a treatment, K_0, which we will have more to say about later. Using this we get the following decomposition of expected regret:

$$\begin{aligned}
\mathbb{E}\left[R_N\left(\hat{\pi}\right)\right] &= \mathbb{E}\left[R_N\left(\hat{\pi}\right)\left(\sum_{i=1}^{K_0} 1_{\mathcal{C}_{i-1}\backslash\mathcal{C}_i} + 1_{\mathcal{C}_{K_0}}\right)\right] \\
&\leq n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1}\backslash\mathcal{C}_i\right) + \sum_{i=1}^{K_0} B_i(b_i)\Delta_i + n\Delta_{K_0+1}.
\end{aligned} \tag{7.5}$$

where $\mathcal{C}_0$ denotes the underlying sample space. For every $i = 1, ..., K$ the event $\mathcal{C}_{i-1}\backslash\mathcal{C}_i$ can be decomposed as $\mathcal{C}_{i-1}\backslash\mathcal{C}_i = (\mathcal{C}_{i-1} \cap \mathcal{A}_i^c) \cup (\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1})$. Therefore, the first term on the right-hand side of (7.5) can be written as

$$n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1}\backslash\mathcal{C}_i\right) = n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) + n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right). \tag{7.6}$$

Notice that $\mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) = 0$ if $b_{i-1} = b_i$. On the event $\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}$ the optimal treatment has not eliminated treatment i after batch b_i. Therefore, for the last term on the right hand side of equation (7.6), we find that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right) &\leq \mathbb{P}\left(\hat{\Delta}_i(\underline{B}(b_i)) \leq \gamma\epsilon_{\underline{B}(b_i)}\right) \\
&\leq \mathbb{P}\left(\hat{\Delta}_i(\underline{B}(b_i)) - \Delta_i \leq \gamma\epsilon_{\tau_i} - \Delta_i\right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(|\hat{\Delta}_i(\underline{B}(b_i)) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i} | \underline{B}(b_i)\right)\right]
\end{aligned}$$

For any $s \geq \tau_i$ we have that

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\Delta}_i(s) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\hat{\mu}_s^{(i)}, (\hat{\sigma}_s^2)^{(i)}) + f(\mu^{(i)}, (\sigma^2)^{(i)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(|f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right) + \mathbb{P}\left(|f(\hat{\mu}_s^{(i)}, (\hat{\sigma}_s^2)^{(i)}) - f(\mu^{(i)}, (\sigma^2)^{(i)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right). \tag{7.7}$$

Furthermore, for any $j \in \{i, *\}$, the mean value theorem yields that

$$\begin{aligned}
&\mathbb{P}\left(|f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |(\hat{\sigma}_s^2)^{(j)} - (\sigma^2)^{(j)}| \geq \frac{1}{4\mathcal{K}}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(3|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |\hat{\mu}_{2,s}^{(j)} - \mu_2^{(j)}| \geq \frac{1}{4\mathcal{K}}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|\hat{\mu}_s^{(j)} - \mu^{(j)}| \geq \frac{1}{16\mathcal{K}}\gamma\epsilon_{\tau_i}\right) + \mathbb{P}\left(|(\hat{\mu}_{2,s})^{(j)} - \mu_2^{(j)}| \geq \frac{1}{16\mathcal{K}}\gamma\epsilon_{\tau_i}\right)
\end{aligned} \tag{7.8}$$

where $\mu_2^{(j)} = \mathbb{E}\left[(Y_1^{(j)})^2\right]$ and $\hat{\mu}_{2,s}^{(j)} = \frac{1}{s}\sum_{r=1}^{s}(Y_{\iota_r}^{(j)})^2$ with ι_r as defined in Section 2.2. By combining equations (7.7) and (7.8), and applying Hoeffding's inequality along with Doob's optional skipping theorem as well as the fact that $\gamma \geq \mathcal{K}$, we arrive at the following bound,

$$\begin{aligned}
\mathbb{P}\left(|\hat{\Delta}_i(s) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right) &\leq C\exp\left(-\frac{2}{256}\epsilon_{\tau_i}^2 s\right) \\
&\leq C\exp\left(-\frac{1}{128}\epsilon_{\tau_i}^2\tau_i\right) \\
&= C\exp\left(-\overline{\log}\left(\frac{T}{\tau_i}\right)\right) \\
&\leq C\frac{\tau_i}{T}.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right) \leq C\frac{\tau_i}{T} \tag{7.9}$$

On the event $\mathcal{C}_{i-1} \cap \mathcal{A}_i^c$ the optimal treatment is eliminated between batch $b_{i-1}+1$ and b_i. Furthermore, every suboptimal treatment $j \leq i-1$ has also been eliminated. Therefore the probability of this event can be bounded as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) &\leq \mathbb{P}\left(\exists(j,s), i \leq j \leq K, b_{i-1}+1 \leq s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma\epsilon_{\underline{B}(s)}\right) \\
&\leq \sum_{j=i}^{K}\mathbb{P}\left(\exists s, b_{i-1}+1 \leq s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma\epsilon_{\underline{B}(s)}\right) \\
&= \sum_{j=i}^{K}\left[\Phi_j(b_i) - \Phi_j(b_{i-1})\right],
\end{aligned}$$

where $\Phi_j(b) = \mathbb{P}\left(\exists s \leq b; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma\epsilon_{\underline{B}(s)}\right)$. We now proceed to bound terms of the form $\Phi_j(b_i)$ for $j \geq i$.

$$\begin{aligned}
\mathbb{P}\left(\exists s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma\epsilon_{\underline{B}(s)}\right) &\leq \mathbb{P}\left(\exists s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) - \Delta_j \leq -\gamma\epsilon_{\underline{B}(s)}\right) \\
&\leq \mathbb{P}\left(\exists s \leq \underline{B}(b_i); \hat{\Delta}_j(s) - \Delta_j \leq -\gamma\epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; \hat{\Delta}_j(s) - \Delta_j \leq -\gamma\epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; |f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{2}\gamma\epsilon_s\right) \\
&+ \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; |f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{2}\gamma\epsilon_s\right).
\end{aligned}$$

For any $j \in \{i, ..., K, *\}$ we find that, by the mean value theorem,

$$\begin{aligned}
&\mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{2}\gamma\epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |\hat{\mu}_s^{(j)} - \mu^{(j)}| + |(\hat{\sigma}_s^2)^{(j)}) - (\sigma^2)^{(j)})| \geq \frac{1}{2\mathcal{K}}\gamma\epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : 3|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |\hat{\mu}_{2,s}^{(j)} - \mu_2^{(j)}| \geq \frac{1}{2\mathcal{K}}\gamma\epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |\hat{\mu}_s^{(j)} - \mu^{(j)}| \geq \frac{1}{8\mathcal{K}}\gamma\epsilon_s\right) + \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |(\hat{\mu}_{2,s})^{(j)} - \mu_2^{(j)}| \geq \frac{1}{8\mathcal{K}}\gamma\epsilon_s\right) \\
&\leq C\frac{\tau_i + \overline{m}}{T}
\end{aligned}$$

where we have used equation (7.4), Doob's optional skipping theorem and Lemma A.1 in Perchet and Rigollet (2013). It follows that

$$\begin{aligned}
\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) &\leq \sum_{i=1}^{K_0} \Delta_i \sum_{j=i}^{K} \left[\Phi_j(b_i) - \Phi_j(b_{i-1})\right] \\
&\leq \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \Phi_j(b_i)\left(\Delta_i - \Delta_{i+1}\right) + \sum_{j=K_0}^{K} \Delta_{K_0}\Phi_j(b_{K_0}) + \sum_{j=1}^{K_0-1} \Delta_j \Phi_j(b_j) \\
&\leq \frac{C}{T}\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \left(\tau_i + \overline{m}\right)\left(\Delta_i - \Delta_{i+1}\right) + \frac{C}{T}\sum_{j=1}^{K} \Delta_{j\wedge K_0}\left(\tau_{j\wedge K_0} + \overline{m}\right).
\end{aligned} \tag{7.10}$$

Using equation (7.3) we obtain

$$\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \tau_i\left(\Delta_i - \Delta_{i+1}\right) \leq C\gamma^2 \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \frac{(\Delta_i - \Delta_{i+1})}{\Delta_i^2}\overline{\log}\left(\frac{T\Delta_i^2}{288\gamma^2}\right)$$

$$\leq C\gamma^2 \sum_{j=1}^{K} \int_{\Delta_{j\wedge K_0}}^{\Delta_1} \frac{1}{x^2} \overline{\log}\left(\frac{Tx^2}{288\gamma^2}\right) dx$$
$$\leq C\gamma^2 \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}} \overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\gamma^2}\right).$$

The part involving $\overline{m}$ in equation (7.10)can be bounded by

$$\overline{m} \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} (\Delta_i - \Delta_{i+1}) + \sum_{j=1}^{K} \Delta_{j\wedge K_0} \overline{m} \leq \overline{m} K.$$

Bringing things together we have

$$\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) \leq C\left(\frac{\gamma^2}{T} \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}} \overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\gamma^2}\right) + \frac{K\overline{m}}{T}\right) \tag{7.11}$$

Combining this with equation (7.6) and (7.5) we obtain

$$\begin{aligned}
\mathbb{E}\left[R_N\left(\hat{\pi}\right)\right] &\leq C\Bigg(\frac{\gamma^2 n}{T} \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}} \overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\gamma^2}\right) + \frac{n\gamma^2}{T} \sum_{j=1}^{K_0} \frac{1}{\Delta_j} \overline{\log}\left(\frac{T\Delta_j^2}{288\gamma^2}\right) \\
&\quad + \sum_{i=1}^{K_0} B_i(b_i)\Delta_i + n\Delta_{K_0+1} + \frac{n\overline{m}K}{T}\Bigg) \\
&\leq C\Bigg(\left(1 + \frac{n}{T}\right)\gamma^2 \sum_{j=1}^{K_0} \frac{1}{\Delta_j} \overline{\log}\left(\frac{T\Delta_j^2}{288\gamma^2}\right) + \frac{\gamma^2 n}{T} \frac{K - K_0}{\Delta_{K_0}} \overline{\log}\left(\frac{T\Delta_{K_0}^2}{288\gamma^2}\right) \\
&\quad + n\Delta_{K_0+1} + \frac{n\overline{m}K}{T}\Bigg).
\end{aligned} \tag{7.12}$$

Fix $\Delta > 0$ and let K_0 be such that $\Delta_{K_0+1} = \Delta^-$. Define the function

$$\phi(x) = \frac{1}{x} \overline{\log}\left(\frac{Tx^2}{288\gamma^2}\right),$$

and notice that $\phi(x) \leq 2e^{-1/2}\phi(x')$ for any $x \geq x' \geq 0$. Using this with $x' = \Delta$ and $x = \Delta_i$ for $i \leq K_0$ we obtain

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C\left(\frac{\gamma^2 K}{\Delta}\left(1 + \frac{n}{T}\right) \overline{\log}\left(\frac{T\Delta^2}{288\gamma^2}\right) + n\Delta^- + \frac{n\overline{m}K}{T}\right). \tag{7.13}$$

■

Proof of Theorem 2.1. Consider the sequential treatment policy with $\gamma = \mathcal{K}$ and $T = n$. From equation (7.12) it follows that for any $K_0 \leq K$

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C\Bigg(\mathcal{K}^2 \sum_{j=1}^{K_0} \frac{1}{\Delta_j} \overline{\log}\left(\frac{n\Delta_j^2}{288\mathcal{K}^2}\right) + \mathcal{K}^2 \frac{K - K_0}{\Delta_{K_0}} \overline{\log}\left(\frac{n\Delta_{K_0}^2}{288\mathcal{K}^2}\right) + n\Delta_{K_0+1} + \frac{n\overline{m}K}{T}\Bigg). \tag{7.14}$$

This can be used to show

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C \min\left(\mathcal{K}^2 \sum_{j=1}^{K} \frac{1}{\Delta_j} \overline{\log}\left(\frac{n\Delta_j^2}{288\mathcal{K}^2}\right) + \overline{m}K, \sqrt{n\mathcal{K}^2\overline{m}K \log(\overline{m}K/\mathcal{K})}\right). \tag{7.15}$$

where the first part of the upper bound in (7.15) follows by using (7.14) with $K = K_0$. The second part follows from Lemma 7.1 by choosing $\Delta = \frac{\mathcal{K}\sqrt{K}\overline{m}}{\sqrt{n}}$.

Proof of Theorem 2.2. The idea of the proof is similar to the one used in the proof of Theorem 1 in Auer et al. (2002). For now we will keep N fixed.[9] First, note that for any positive integer l

$$T_i(N) = 1 + \sum_{t=K+1}^{N} \mathbf{1}_{\{\hat{\pi}_t = i\}} \leq l + \sum_{t=K+1}^{N} \mathbf{1}_{\{\hat{\pi}_t = i,\ T_i(t-1) \geq l\}} \leq l + \sum_{t=K+1}^{N} \mathbf{1}_{\{T_i(t-1) \geq l\}} \leq l + N\mathbf{1}_{\{T_i(N-1) \geq l\}}$$

It remains to bound the probability of the event $\{T_i(N-1) \geq l\}$. This is the probability that treatment i has not been eliminated before having been assigned at least l times. Define $\tilde{b}_i = \max\{b : \sum_{j=1}^{b} m_{i,j} < l\}$ and note that if treatment i is assigned l times then it cannot have been eliminated after $\tilde{b}_i$ batches. In particular, it cannot have been eliminated by the optimal treatment. Let

$$\mathcal{A}_i = \{\text{the optimal treatment has not been eliminated after batch } \tilde{b}_i\}$$

For any t we have that $\{T_i(N-1) \geq l\} \subseteq (\{T_i(N-1) \geq l\} \cap \mathcal{A}_i) \cup \mathcal{A}_i^c$. Thus, $(\{T_i(N-1) \geq l\} \cap \mathcal{A}_i) \subseteq \left\{\hat{\Delta}_i(\underline{B}(\tilde{b}_i)) - \Delta_i \leq \gamma\epsilon_{\underline{B}(\tilde{b}_i)} - \Delta_i\right\}$ which implies

$$\begin{aligned}\mathbb{E}\left[T_i(N)\right] &\leq l + N\mathbb{P}\left(\{T_i(N-1) \geq l\} \cap \mathcal{A}_i\right) + N\mathbb{P}\left(\mathcal{A}_i^c\right) \\ &\leq l + N\mathbb{P}\left(\hat{\Delta}_i(\underline{B}(\tilde{b}_i)) - \Delta_i \leq \gamma\epsilon_{\underline{B}(\tilde{b}_i)} - \Delta_i\right) + N\mathbb{P}\left(\mathcal{A}_i^c\right)\end{aligned}$$

From equations (7.2) and (7.3) of Lemma 7.1 we have that (where τ_i is defined in the said lemma)

$$\tau_i \leq \frac{C\mathcal{K}^2}{\Delta_i^2} \overline{\log}\left(\frac{n}{288\mathcal{K}^2}\right)$$

[9] In other words all calculations are done conditional on N.

Thus, by letting $l = \bar{m} + \lceil \frac{288\mathcal{K}^2}{\Delta_i^2} \overline{\log}\left(\frac{n}{288\mathcal{K}^2}\right) \rceil$ it follows that $\tau_i \leq l - \bar{m} \leq \underline{B}(\tilde{b}_i) < l$. In particular, we have that $\gamma\epsilon_{\underline{B}(\tilde{b}_i)} \leq \gamma\epsilon_{\tau_i} \leq \frac{2}{3}\Delta_i$. Hence,

$$\begin{aligned}
\mathbb{P}\left(\hat{\Delta}_i(\underline{B}(\tilde{b}_i)) - \Delta_i \leq \gamma\epsilon_{\underline{B}(\tilde{b}_i)} - \Delta_i\right) &\leq \mathbb{P}\left(|\hat{\Delta}_i(\underline{B}(\tilde{b}_i)) - \Delta_i| \geq \frac{1}{3}\Delta_i\right) \\
&\leq C\mathbb{E}\left[\exp\left(-\frac{2\underline{B}(\tilde{b}_i)\Delta_i^2}{576}\right)\right] \\
&\leq C\exp\left(-\frac{(l-\bar{m})\Delta_i^2}{288}\right) \\
&\leq C\frac{\mathcal{K}^2}{n}.
\end{aligned}$$

Next, we bound the term involving $\mathcal{A}_i^c$. To this end we start by noting that if the optimal treatment does not survive until batch $\tilde{b}_i$, then it must have been eliminated in one of the batches before $\tilde{b}_i$.

$$\mathbb{P}\left(\mathcal{A}_i^c\right) \leq \sum_{j=1}^{K} \mathbb{P}\left(\exists s \leq \underline{B}(\tilde{b}_i) : \hat{\Delta}_j(s) \leq -\gamma\epsilon_s\right) \tag{7.16}$$

$$\leq \sum_{j=1}^{K} \mathbb{P}\left(\exists s \leq l : \hat{\Delta}_j(s) \leq -\gamma\epsilon_s\right) \tag{7.17}$$

$$\leq CK\frac{l}{n}, \tag{7.18}$$

where the last inequality follows from an application of Lemma A.1 in Perchet and Rigollet (2013). Bringing things together, taken expectations with respect to N and using Jensen's inequality in order to replace N with its expectation yields the desired result.

7.2 Proof of Theorems in Section 3

Proof of Theorem 3.1. It is convenient to define the constant $c_1 = 6L\mathcal{K} + 1$, which will enter several of the bounds derived below. Furthermore, we let c denote a positive constant which may change from line to line. By the construction of the treatment policy it follows that the regret can be written as $R_N(\bar{\pi}) = \sum_{j=1}^{F} R_j(\bar{\pi})$, where

$$R_j(\bar{\pi}) = \sum_{t=1}^{N}\left(f^{(\star)}(X_t) - f^{(\hat{\pi}_{B_j}, N_{B_j}(t))}(X_t)\right)1_{\left(X_t \in B_j\right)}.$$

We start by providing an upper bound on the welfare lost for each group B_j due to the policy targeting $f_j^{(*)} = \max_{1\leq i\leq K} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)})$ instead of $f^{(\star)}(x)$. To this end note that

$$f^{(\star)}(x) = \max_{1\leq i\leq K+1} f(\mu^{(i)}(x), (\sigma^2)^{(i)}(x))$$

$$\leq \max_{1\leq i\leq K+1} f(\bar{\mu}_j^{(i)}, (\bar{\sigma}^2)_j^{(i)}) + \mathcal{K} \max_{1\leq i\leq K+1} |\mu^{(i)}(x) - \bar{\mu}_j^{(i)}| + \mathcal{K} \max_{1\leq i\leq K+1} |(\sigma^2)^{(i)}(x) - (\bar{\sigma}^2)_j^{(i)}|.$$

Fix $x \in B_j$ and $i \in \{1, ..., K+1\}$. Then, for all $y \in B_j$, one has by the (L, β)-Hölder continuity of $\mu^{(i)}(x)$

$$\mu^{(i)}(x) \leq \mu^{(i)}(y) + |\mu^{(i)}(x) - \mu^{(i)}(y)| \leq \mu^{(i)}(y) + LV_j^{\beta},$$

which upon integrating over y yields $\mu^{(i)}(x) \leq \bar{\mu}_j^{(i)} + LV_j^{\beta}$. Similarly, it holds that $\mu^{(i)}(x) \geq \bar{\mu}_j^{(i)} - LV_j^{\beta}$ such that for all $x \in B_j$ we have $|\mu^{(i)}(x) - \bar{\mu}_j^{(i)}| \leq LV_j^{\beta}$. Next, note that the map $[0,1] \ni z \mapsto z^2$ is Lipschitz continuous with constant 2 which implies that $(\mu^{(i)}(x))^2$ is $(2L, \beta)$-Hölder. This, together with the (L, β)-Hölder continuity of $(\sigma^2)^{(i)}(x) = \mathbb{E}(Y_t^{(i)^2}|X_t = x) - (\mu^{(i)}(x))^2$ implies that $\mathbb{E}(Y_t^{(i)^2}|X_t = x)$ is $(3L, \beta)$-Hölder continuous. Thus, by similar arguments as above $|\mathbb{E}(Y_t^{(i)^2}|X_t = x) - \mathbb{E}(Y_t^{(i)^2}|X_t \in B_j)| \leq 3LV^{\beta}$ for all $x \in B_j$. The mean value theorem also yields that $|(\mu^{(i)}(x))^2 - \bar{\mu}_j^{(i)^2}| \leq 2LV_j^{\beta}$ for all $x \in B_j$. Therefore,

$$\begin{aligned}\left|(\sigma^2)^{(i)}(x) - (\bar{\sigma}^2)_j^{(i)}\right| &= \left|\mathbb{E}(Y_t^{(i)^2}|X_t = x) - (\mu^{(i)}(x))^2 - [\mathbb{E}(Y_t^{(i)^2}|X_t \in B_j) - \bar{\mu}_j^{(i)^2}]\right| \\ &\leq 5LV_j^{\beta}\end{aligned}$$

Thus, for $x \in B_j$,

$$f^{(\star)}(x) \leq f_j^{(*)} + c_1 V_j^{\beta}.$$

A similar argument to the above yields that for all $x \in B_j$

$$f^{(\bar{\pi}_t)}(x) \geq \bar{f}_j^{(\bar{\pi}_t)} - c_1 V_j^{\beta}.$$

Next we define $\tilde{R}_j(\bar{\pi}) = \sum_{t=1}^{N_{B_j}(N)} \left(f_j^{(*)} - \bar{f}_j^{(\hat{\pi}_{B_j,t})}\right)$. This corresponds to the regret associated with a treatment problem without covariates where treatment i yields reward $\bar{f}_j^{(i)}$, and the best treatment yields $f_j^{(*)} = \max_i \bar{f}_j^{(i)} \leq \bar{f}_j^{\star}$. Therefore, we can write

$$\begin{aligned}R_j(\bar{\pi}) &= \sum_{t=1}^{N} \left(f^{(\star)}(X_t) - f^{(\hat{\pi}_{B_j,N_{B_j}(t)})}(X_t)\right) 1_{(X_t \in B_j)} \leq \sum_{t=1}^{N} \left(f_j^{(*)} - \bar{f}_j^{(\hat{\pi}_{B_j,N_{B_j}(t)})} + 2c_1 V_j^{\beta}\right) 1_{(X_t \in B_j)} \\ &= \tilde{R}_j(\bar{\pi}) + 2c_1 V_j^{\beta} N_{B_j}(N),\end{aligned}$$

where $N_{B_j}(N)$ is the number of observations falling in bin j given that there are N observations in total. Taking expectations, and using that the density of X_t is bounded from above implies that $\mathbb{E}\left[N_j(N)\right] \leq \bar{c} n \bar{B}_j$ gives

$$\mathbb{E}\left[R_j(\bar{\pi})\right] \leq \mathbb{E}\left[\tilde{R}_j(\bar{\pi})\right] + n\bar{c}\bar{B}_j c_1 V_j^{\beta}. \tag{7.19}$$

Since $\mathbb{E}\left[\tilde{R}_j(\bar{\pi})\right]$ is the expected regret of a treatment problem without covariates we can apply Theorem 7.1 with the following values

$$\Delta = \sqrt{\frac{\overline{m}K\log(\overline{m}K)}{n\bar{B}_j}}, \quad \gamma = \mathcal{K}L, \quad T = n\bar{B}_j,$$

for each bin $j = 1, ..., F$ to obtain the following bound on the regret accumulated across any group j:

$$\mathbb{E}\left[R_j(\bar{\pi})\right] \leq C\left[\sqrt{\overline{m}K\log(\overline{m}K)n\bar{B}_j} + n\bar{B}_j V_j^{\beta}\right].$$

Thus, adding up the expected regret over all F groups yields

$$\mathbb{E}\left[R_N(\bar{\pi})\right] \leq C\sum_{j=1}^{F}\left[\sqrt{\overline{m}K\log(\overline{m}K)n\bar{B}_j} + n\bar{B}_j V_j^{\beta}\right].$$

Proof of Corollary 3.1.1. The result follows from Theorem 3.1 upon noting that $\bar{B}_j = P^{-d}$ and $V_j = \sqrt{d}P^{-1}$ for $j = 1, ..., P$ (and ignoring the constant $\sqrt{d}$) with P as in the theorem completes the proof.

Proof of Theorem 3.2. Define $\underline{S} = \mathcal{S}(\alpha, \beta, L, \mathcal{K}, d, \bar{c}, \overline{m})$ to be the subset of $\mathcal{S}(\beta, L, \mathcal{K}, d, \bar{c}, \overline{m}) =: \bar{S}$ which also satisfies the margin condition. Set $\overline{m} = 1, K = 2$ and fix a policy π. Then, by Theorem 4.1 in Rigollet and Zeevi (2010), there exists a constant $C(\alpha)$ such that for all $\alpha > 0$

$$\sup_{\underline{S}} \mathbb{E}\left[R_N(\pi)\right] \geq C(\alpha)n^{1-\frac{\beta+\beta\alpha}{2\beta+d}}.$$

Thus, choosing an $\alpha = \alpha(\varepsilon)$ such that $\frac{\beta\alpha}{2\beta+d} \leq \varepsilon$, we conclude

$$\sup_{\bar{S}} \mathbb{E}\left[R_N(\pi)\right] \geq \sup_{\underline{S}} \mathbb{E}\left[R_N(\pi)\right] \geq C(\alpha)n^{1-\frac{\beta+\beta\alpha}{2\beta+d}} \geq C(\alpha)n^{1-\frac{\beta}{2\beta+d}} \cdot n^{-\varepsilon} = C(\varepsilon)n^{1-\frac{\beta}{2\beta+d}} \cdot n^{-\varepsilon},$$

where the last inequality used that α is a function of ε.

Proof of Theorem 3.3. The proof is identical to the proof of Theorem 2.2 but with with $n = \mathbb{E}(N)$ replaced by $\bar{c}n\bar{B}_j \geq \mathbb{E}\left(N_{B_j}(N)\right)$. Thus, the expected number of assignments is replaced by an upper bound on the expected number of individuals falling in group j and the result of Theorem 2.2 is applied on each group separately.

Proof of Theorem 3.4. The proof is similar to that found in Tsybakov (2004a) and Rigollet and Zeevi (2010). Fix $\delta < \delta_0$. Then, for any policy π

$$R_N(\pi) \geq \delta\sum_{t=1}^{N} 1_{\{f^{(\star)}(X_t) - f^{(\pi_t)}(X_t) > \delta\}}$$

$$\geq \delta \left(S_N(\pi) - \sum_{t=1}^{N} 1_{\{0<|f^{(\star)}(X_t)-f^{(\pi_t)}(X_t)|\leq\delta\}} \right)$$
$$\geq \delta \left(S_N(\pi) - \sum_{t=1}^{N} 1_{\{0<|f^{(\star)}(X_t)-f^{(\sharp)}(X_t)|\leq\delta\}} \right)$$

Since $S_n(\pi) \leq N$ there exists a $c > 0$ not depending on N such that $\left(\frac{S_n(\pi)}{cn}\right)^{\frac{1}{\alpha}} < \delta_0$. Thus, we can set $\delta = \left(\frac{S_n(\pi)}{cn}\right)^{\frac{1}{\alpha}}$ and use the margin condition upon integration on both sides of the above display to get (3.5). To obtain (3.6) insert (3.4) into (3.5).

Proof of Theorem 3.5. The proof is similar to the proof of Theorem 3.1 once we fix a value of the discrete covariates. Let c denote a positive constant which may change from line to line. By the construction of the treatment policy it follows that the regret can be written as $R_N(\tilde{\pi}) = \sum_{a\in A}\sum_{j=1}^{F_a} R_{a,j}(\hat{\pi})$, where

$$R_{a,j}(\hat{\pi}) = \sum_{t=1}^{N} \left(f^{(\star)}(X_t) - f^{(\hat{\pi}_{(\{a\}\times B_{a,j}),N_{a,j}(t)})}(X_t) \right) 1_{\left(X_{t,D}=a, X_{t,C}\in B_{a,j}\right)}.$$

For any bin $B_{a,j}$ define

$$\bar{\mu}_{a,j}^{(i)} = \mathbb{E}(Y_t^{(i)}|X_{t,D}=a, X_{t,C}\in B_{a,j}) = \frac{1}{\mathbb{P}_X(X_{t,D}=a, X_{t,C}\in B_{a,j})}\int_{a\times B_{a,j}} \mu^{(i)}(x)d\mathbb{P}_X(x)$$

and

$$\begin{aligned}(\bar{\sigma}^2)_{a,j}^{(i)} &= Var(Y_t^{(i)}|X_{t,D}=a, X_{t,C}\in B_{a,j}) \\ &= \mathbb{E}({Y_t^{(i)}}^2|X_{t,D}=a, X_{t,C}\in B_{a,j}) - [\mathbb{E}(Y_t^{(i)}|X_{t,D}=a, X_{t,C}\in B_{a,j})]^2\end{aligned}$$

Furthermore, let $\bar{f}_{a,j}^{(i)} = f(\bar{\mu}_{a,j}^{(i)}, (\bar{\sigma}^2)_{a,j}^{(i)})$ with $f_{a,j}^{(*)} = \max_{1\leq i\leq K+1} \bar{f}_{a,j}^{(i)}$. By exactly the same arguments as in the proof of Theorem 3.1 we now get that for $x \in \{a\}\times B_{a,j}$,

$$f^{(\star)}(x) \leq f_{a,j}^{(*)} + cV_{a,j}^{\beta},$$

as well as,

$$f^{(\hat{\pi}_{(a\times B_{a,j}),N_{a,j}(t)})}(x) \geq \bar{f}_{a,j}^{(\hat{\pi}_{(\{a\}\times B_{a,j}),N_{a,j}(t)})} - cV_{a,j}^{\beta}.$$

Next we define $\tilde{R}_{a,j}(\tilde{\pi}) = \sum_{t=1}^{N_{a,j}} \left(f_{a,j}^{(*)} - \bar{f}_{a,j}^{(\hat{\pi}_{(\{a\}\times B_{a,j}),t)}} \right)$. This corresponds to the regret associated with a treatment problem without covariates where treatment i yields reward $\bar{f}_{a,j}^{(i)}$, and the best treatment yields $f_{a,j}^{(*)} = \max_i \bar{f}_{a,j}^{(i)} \leq \bar{f}_{a,j}^{\star}$. Therefore, we can write

$$R_{a,j}(\hat{\pi}) = \sum_{t=1}^{N} \left(f^{(\star)}(X_t) - f^{(\hat{\pi}_{(a\times B_{a,j}),N_{a,j}(t)})}(X_t) \right) 1_{\left(X_{t,D}=a, X_{t,C}\in B_{a,j}\right)}$$

$$\leq \sum_{t=1}^{N} \left(f_{a,j}^{(*)} - \bar{f}_{a,j}^{(\hat{\pi}(\{a\}\times B_{a,j}),N_{a,j}(t))} + 2cV_{a,j}^{\beta} \right) 1_{\left(X_{t,D}=a, X_{t,C}\in B_{a,j}\right)}$$
$$= \tilde{R}_{a,j}(\hat{\pi}) + 2cV_{a,j}^{\beta} N_{a,j}(N),$$

where $N_{a,j}(N)$ is the number of observations for which $x \in a \times B_{a,j}$ given that there are N observations in total. Taking expectations, and using that N is independent of all other random variables implies $\mathbb{E}\left[N_{a,j}(N)\right] \leq n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})$ gives

$$\mathbb{E}\left[R_{a,j}(\hat{\pi})\right] \leq \mathbb{E}\left[\tilde{R}_{a,j}(\hat{\pi})\right] + n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})cV_{a,j}^{\beta}.$$

Since $\mathbb{E}\left[\tilde{R}_{a,j}(\hat{\pi})\right]$ is the expected regret of a treatment problem without covariates we can apply Theorem 7.1 with the following values

$$\Delta = \sqrt{\frac{\overline{m}K\log(\overline{m}K)}{n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})}}, \quad \gamma = \mathcal{K}L, \quad T = n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j}),$$

for each $a \in A$ and $B_{a,j},\ j = 1, ..., F_a$ to obtain the following bound on the regret accumulated across any group:

$$\mathbb{E}\left[R_{a,j}(\bar{\pi})\right] \leq c\left[\sqrt{\overline{m}K\log(\overline{m}K)n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})} + n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})_j V_j^{\beta}\right].$$

Thus, adding up the expected regret over all groups yields

$$\mathbb{E}\left[R_N(\tilde{\pi})\right] \leq c\sum_{a\in A}\sum_{j=1}^{F_a}\left[\sqrt{\overline{m}K\log(\overline{m}K)n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})} + n\mathbb{P}(X_{t,D}=a, X_{t,C}\in B_{a,j})V_{a,j}^{\beta}\right].$$

7.3 Proof of Theorems in Section 4

Proof of Theorem 4.1. Define $\epsilon_s = u(s,T) = 8\sqrt{\frac{2\bar{a}^2}{s}\overline{\log}\left(\frac{T}{s}\right)}$. In the following we will distinguish between two types of batches, namely batches of individuals that have to be assigned a treatment, and batches of information on the outcome of previously assigned treatments. The latter type of batches will be the key object of interest when determining whether or not to eliminate a given treatment, whereas the former will be relevant when counting the total regret from running the treatment policy. In this proof we let $\underline{B}(s)$ denote the minimal number of observed outcomes per treatment based on s batches of information. Consider a batch b of information. Recall that if the optimal treatment as well as some treatment i have not been eliminated, then the optimal treatment will eliminate treatment i if $\hat{\Delta}_i(\underline{B}(b)) \geq \gamma\epsilon_{\underline{B}(b)}$, and treatment i will eliminate the optimal treatment if $\hat{\Delta}_i(\underline{B}(b)) \leq -\gamma\epsilon_{\underline{B}(b)}$.

To be able to say something about when either of these two events occurs we introduce the (unknown) quantity, τ_i^*, which is defined through the relation

$$\Delta_i = 12\gamma\sqrt{\frac{2\bar{a}^2}{\tau_i^*}\overline{\log}\left(\frac{T}{\tau_i^*}\right)}, \qquad i = 1, ..., K.$$

Since τ_i^* in general will not be an integer, we also define $\tau_i = \lceil \tau_i^* \rceil$. Next introduce the hypothetical batch (of information) $b_i = \min\{l : \underline{B}(l) \geq \tau_i^*\}$. It is the first batch of information after which we have more than τ_i^* observations of the outcome of treatment i. Notice that

$$\tau_i^* \leq \underline{B}(b_i) \leq C \left(\frac{\bar{a}^2\gamma^2}{\Delta_i^2} \overline{\log} \left(\frac{T\Delta_i^2}{288\bar{a}^2\gamma^2} \right) + \overline{m} \right), \tag{7.20}$$

$$\tau_i \leq \underline{B}(b_i), \tag{7.21}$$

$$\underline{B}(b_i) \leq \tau_i + \overline{m}, \tag{7.22}$$

Notice that $1 \leq \tau_1 \leq ... \leq \tau_K$ and $1 \leq b_1 \leq ... \leq b_K$. Define the following events:

$\mathcal{A}_i = \{\text{The optimal treatment has not been eliminated after batch } b_i \text{ has been observed}\}$,

$\mathcal{B}_i = \{\text{Every treatment } j \in \{1, ..., i\} \text{ has been eliminated after batch } b_j \text{ has been observed}\}$.

Furthermore, let $\mathcal{C}_i = \mathcal{A}_i \cap \mathcal{B}_i$, and observe that $\mathcal{C}_1 \supseteq ... \supseteq \mathcal{C}_K$. For any $i = 1, ..., K$, the contribution to regret incurred after batch b_i of information is at most $\Delta_{i+1}N$ on $\mathcal{C}_i$. In what follows we fix a treatment, K_0, which we will be specific about later. Using this and letting m denote the expected number of observations in a batch we get the following decomposition of expected regret:

$$\begin{aligned}
\mathbb{E}\left[R_N\left(\hat{\pi}\right)\right] &= \mathbb{E}\left[R_N\left(\hat{\pi}\right)\left(\sum_{i=1}^{K_0} 1_{\mathcal{C}_{i-1}\setminus\mathcal{C}_i} + 1_{\mathcal{C}_{K_0}}\right)\right] \\
&\leq n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1}\setminus\mathcal{C}_i\right) + \sum_{i=1}^{K_0} B_i(b_i)\Delta_i + n\Delta_{K_0+1} + Dm,
\end{aligned} \tag{7.23}$$

where the last term is due to the fact that the delay means that the all treatment allocations during the first $D+1$ batches have to be made without any information about the treatment outcomes. For every $i = 1, ..., K$ the event $\mathcal{C}_{i-1}\setminus\mathcal{C}_i$ can be decomposed as follows

$$\mathcal{C}_{i-1}\setminus\mathcal{C}_i = \left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) \cup \left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right).$$

Therefore, the first term on the right-hand side of (7.23) can be written as

$$n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1}\setminus\mathcal{C}_i\right) = n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) + n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right). \tag{7.24}$$

Notice that the first term on the right-hand side will be zero if $b_{i-1} = b_i$. On the event $\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}$ the optimal treatment has not eliminated treatment i at batch b_i. Therefore, for the last term on the right hand side of equation (7.24) we find that

$$\mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right) \leq \mathbb{P}\left(\hat{\Delta}_i(\underline{B}(b_i)) \leq \gamma\epsilon_{\underline{B}(b_i)}\right)$$

$$\leq \mathbb{P}\left(\hat{\Delta}_i(\underline{B}(b_i)) - \Delta_i \leq \gamma\epsilon_{\tau_i} - \Delta_i\right)$$
$$\leq \mathbb{E}\left[\mathbb{P}\left(|\hat{\Delta}_i(\underline{B}(b_i)) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}|\underline{B}(b_i)\right)\right]. \tag{7.25}$$

For any $s \geq \tau_i$ we have that

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\Delta}_i(s) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\hat{\mu}_s^{(i)}, (\hat{\sigma}_s^2)^{(i)}) + f(\mu^{(i)}, (\sigma^2)^{(i)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right) + \mathbb{P}\left(|f(\hat{\mu}_s^{(i)}, (\hat{\sigma}_s^2)^{(i)}) - f(\mu^{(i)}, (\sigma^2)^{(i)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right).
\end{aligned} \tag{7.26}$$

Furthermore, for any $j \in \{i, *\}$, the mean value theorem yields that

$$\begin{aligned}
&\mathbb{P}\left(|f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{4}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |(\hat{\sigma}_s^2)^{(j)} - (\sigma^2)^{(j)}| \geq \frac{1}{4\mathcal{K}}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(3|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |\hat{\mu}_{2,s}^{(j)} - \mu_2^{(j)}| \geq \frac{1}{4\mathcal{K}}\gamma\epsilon_{\tau_i}\right) \\
&\leq \mathbb{P}\left(|\hat{\mu}_s^{(j)} - \mu^{(j)}| \geq \frac{1}{16\mathcal{K}}\gamma\epsilon_{\tau_i}\right) + \mathbb{P}\left(|(\hat{\mu}_{2,s})^{(j)} - \mu_2^{(j)}| \geq \frac{1}{16\mathcal{K}}\gamma\epsilon_{\tau_i}\right)
\end{aligned} \tag{7.27}$$

where $\mu_2^{(j)} = \mathbb{E}\left[(Y_1^{(j)})^2\right]$ and $\hat{\mu}_{2,s}^{(j)} = \frac{1}{s}\sum_{r=1}^{s}(Y_{\iota_r}^{(j)})^2$ with ι_r as defined in Section 2.2. By combining equations (7.26) and (7.27), and applying Hoeffding's inequality along with Doob's optional skipping theorem as well as the fact that $\gamma \geq \mathcal{K}$, we arrive at the following bound,

$$\begin{aligned}
\mathbb{P}\left(|\hat{\Delta}_i(s) - \Delta_i| \geq \frac{1}{2}\gamma\epsilon_{\tau_i}\right) &\leq C\exp\left(-\frac{2}{256\bar{a}^2}\epsilon_{\tau_i}^2 s\right) \\
&\leq C\exp\left(-\frac{1}{128\bar{a}^2}\epsilon_{\tau_i}^2\tau_i\right) \\
&= C\exp\left(-\overline{\log}\left(\frac{T}{\tau_i}\right)\right) \\
&\leq C\frac{\tau_i}{T}.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\right) \leq C\frac{\tau_i}{T}$$

On the event $\mathcal{C}_{i-1} \cap \mathcal{A}_i^c$ the optimal treatment is eliminated between the time batch $b_{i-1}+1$ and b_i of information arrives. Furthermore, every suboptimal treatment $j \leq i-1$ has also been eliminated. Therefore, the probability of this event can be bounded as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) &\leq \mathbb{P}\left(\exists (j,s), i \leq j \leq K, b_{i-1}+1 \leq s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma \epsilon_{\underline{B}(s)}\right) \\
&\leq \sum_{j=i}^{K} \mathbb{P}\left(\exists s, b_{i-1}+1 \leq s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma \epsilon_{\underline{B}(s)}\right) \\
&= \sum_{j=i}^{K} \left[\Phi_j(b_i) - \Phi_j(b_{i-1})\right],
\end{aligned}$$

where $\Phi_j(b) = \mathbb{P}\left(\exists s \leq b; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma \epsilon_{\underline{B}(s)}\right)$. We proceed to bounding terms of the form $\Phi_j(b_i)$ for $j \geq i$.

$$\begin{aligned}
\mathbb{P}\left(\exists s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) \leq -\gamma \epsilon_{\underline{B}(s)}\right) &\leq \mathbb{P}\left(\exists s \leq b_i; \hat{\Delta}_j(\underline{B}(s)) - \Delta_j \leq -\gamma \epsilon_{\underline{B}(s)}\right) \\
&\leq \mathbb{P}\left(\exists s \leq B_j(b_i); \hat{\Delta}_j(s) - \Delta_j \leq -\gamma \epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; \hat{\Delta}_j(s) - \Delta_j \leq -\gamma \epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; |f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{2}\gamma \epsilon_s\right) \\
&+ \mathbb{P}\left(\exists s \leq \tau_i + \overline{m}; |f(\hat{\mu}_s^{(*)}, (\hat{\sigma}_s^2)^{(*)}) - f(\mu^{(*)}, (\sigma^2)^{(*)})| \geq \frac{1}{2}\gamma \epsilon_s\right).
\end{aligned}$$

For any $j \in \{i, ..., K, *\}$ we find that, by the mean value theorem,

$$\begin{aligned}
&\mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |f(\hat{\mu}_s^{(j)}, (\hat{\sigma}_s^2)^{(j)}) - f(\mu^{(j)}, (\sigma^2)^{(j)})| \geq \frac{1}{2}\gamma \epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |\hat{\mu}_s^{(j)} - \mu^{(j)}| + |(\hat{\sigma}_s^2)^{(j)}) - (\sigma^2)^{(j)})| \geq \frac{1}{2\mathcal{K}}\gamma \epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : 3|\hat{\mu}_s^{(j)} - \mu^{(j)}| + |\hat{\mu}_{2,s}^{(j)} - \mu_2^{(j)}| \geq \frac{1}{2\mathcal{K}}\gamma \epsilon_s\right) \\
&\leq \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |\hat{\mu}_s^{(j)} - \mu^{(j)}| \geq \frac{1}{8\mathcal{K}}\gamma \epsilon_s\right) + \mathbb{P}\left(\exists s \leq \tau_i + \overline{m} : |(\hat{\mu}_{2,s})^{(j)} - \mu_2^{(j)}| \geq \frac{1}{8\mathcal{K}}\gamma \epsilon_s\right) \\
&\leq C\frac{\tau_i + \overline{m}}{T}
\end{aligned}$$

where we once more have used equation (7.22) and Lemma A.1 in Perchet and Rigollet (2013). Using this we find that

$$\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) \leq \sum_{i=1}^{K_0} \Delta_i \sum_{j=i}^{K} \left[\Phi_j(b_i) - \Phi_j(b_{i-1})\right]$$

$$\leq \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0-1} \Phi_j(b_i)\left(\Delta_i - \Delta_{i+1}\right) + \sum_{j=K_0}^{K} \Delta_{K_0}\Phi_j(b_{K_0}) + \sum_{j=1}^{K_0-1} \Delta_j \Phi_j(b_j)$$
$$\leq C\left(\frac{1}{T}\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0-1} \left(\tau_i + \overline{m}\right)\left(\Delta_i - \Delta_{i+1}\right) + \frac{1}{T}\sum_{j=1}^{K} \Delta_{j\wedge K_0}\left(\tau_{j\wedge K_0} + \overline{m}\right)\right). \tag{7.28}$$

Observe that, by (7.21),

$$\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0-1} \tau_i\left(\Delta_i - \Delta_{i+1}\right) \leq C\gamma^2\bar{a}^2 \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0-1} \frac{(\Delta_i - \Delta_{i+1})}{\Delta_i^2} \overline{\log}\left(\frac{T\Delta_i^2}{288\bar{a}^2\gamma^2}\right)$$
$$\leq C\bar{a}^2\gamma^2 \sum_{j=1}^{K} \int_{\Delta_{j\wedge K_0}}^{\Delta_1} \frac{1}{x^2}\overline{\log}\left(\frac{Tx^2}{288\bar{a}^2\gamma^2}\right) dx$$
$$\leq C\bar{a}^2\gamma^2 \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}} \overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\bar{a}^2\gamma^2}\right). \tag{7.29}$$

The parts involving $\overline{m}$ in equation (7.28) can be bounded by

$$\overline{m}\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0-1} \left(\Delta_i - \Delta_{i+1}\right) + \sum_{j=1}^{K} \Delta_{j\wedge K_0}\overline{m} \leq \overline{m}K. \tag{7.30}$$

Bringing together equations (7.28), (7.29) and (7.30) we see that

$$\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\left(\mathcal{C}_{i-1} \cap \mathcal{A}_i^c\right) \leq C\left(\frac{\bar{a}^2\gamma^2}{T}\sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}}\overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\bar{a}^2\gamma^2}\right) + \frac{K\overline{m}}{T}\right). \tag{7.31}$$

Combining this with equation (7.24) and (7.23) we obtain

$$\mathbb{E}\left[R_N\left(\hat{\pi}\right)\right] \leq C\Bigg(\frac{\bar{a}^2\gamma^2 n}{T}\sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}}\overline{\log}\left(\frac{T\Delta_{j\wedge K_0}^2}{288\bar{a}^2\gamma^2}\right) + \frac{n\bar{a}^2\gamma^2}{T}\sum_{j=1}^{K_0} \frac{1}{\Delta_j}\overline{\log}\left(\frac{T\Delta_j^2}{288\bar{a}^2\gamma^2}\right)$$
$$+ \sum_{i=1}^{K_0} B_i(b_i)\Delta_i + n\Delta_{K_0+1} + \frac{n\overline{m}K}{T} + mD\Bigg)$$
$$\leq C\Bigg(\left(1 + \frac{n}{T}\right)\bar{a}^2\gamma^2 \sum_{j=1}^{K_0} \frac{1}{\Delta_j}\overline{\log}\left(\frac{T\Delta_j^2}{288\bar{a}^2\gamma^2}\right) + \frac{\bar{a}^2\gamma^2 n}{T}\frac{K - K_0}{\Delta_{K_0}}\overline{\log}\left(\frac{T\Delta_{K_0}^2}{288\bar{a}^2\gamma^2}\right)$$
$$+ n\Delta_{K_0+1} + \frac{n\overline{m}K}{T} + \overline{m}D\Bigg). \tag{7.32}$$

Fix $\Delta > 0$ and let K_0 be such that $\Delta_{K_0+1} = \Delta^-$. Define the function $\phi(\cdot)$ by

$$\phi(x) = \frac{1}{x}\overline{\log}\left(\frac{Tx^2}{288\bar{a}^2\gamma^2}\right),$$

and notice that $\phi(x) \leq 2e^{-1/2}\phi(x')$ for any $x \geq x' \geq 0$. Using this with $x' = \Delta$ and $x = \Delta_i$ for $i \leq K_0$ we obtain the following bound on the expected regret.

$$\mathbb{E}\left[R_N(\hat{\pi})\right] \leq C\left(\frac{\bar{a}^2\gamma^2 K}{\Delta}\left(1+\frac{n}{T}\right)\overline{\log}\left(\frac{T\Delta^2}{288\bar{a}^2\gamma^2}\right) + n\Delta^- + \frac{n\overline{m}K}{T} + mD\right). \tag{7.33}$$

Note that we by definition we have that $m \leq \overline{m}$. The theorem then follows by arguments similar to those in the proof of theorem 2.1.

■

Proof of Theorem 4.2. Recall equation (7.19). Applying Theorem 4.1 with the following values

$$\Delta = \sqrt{\frac{\overline{m}K\bar{a}^2\log(\overline{m}K)}{n\bar{B}_j}}, \quad \gamma = \mathcal{K}L, \quad T = n\bar{B}_j,$$

for each bin $j = 1, ..., F$, we obtain the following bound on the regret accumulated across the any bin j:

$$\mathbb{E}\left[R_j(\bar{\pi})\right] \leq c\left[\sqrt{\overline{m}\bar{a}^2 K\log(\overline{m}K)n\bar{B}_j} + n\bar{B}_j V_j^{\beta} + \overline{m}K + m^{(j)}D\right],$$

where $m^{(j)}$ is the expected batch size associated with bin j. Note that $m^{(j)} \leq \overline{cm}\bar{B}_j$. Thus,

$$\mathbb{E}\left[R_N(\bar{\pi})\right] \leq c\left(\sum_{j=1}^{F}\left[\sqrt{\overline{m}K\bar{a}^2\log(\overline{m}K)n\bar{B}_j} + n\bar{B}_j V_j^{\beta} + \overline{m}K\right] + \overline{m}D\right).$$

References

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Anthony B Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

Debopam Bhattacharya and Pascaline Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

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.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Gary Chamberlain. Econometrics and decision theory. *Journal of Econometrics*, 95(2):255–283, 2000.

Rajeev H Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125(1): 141–173, 2005.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for reinforcement learning. In *ICML*, pages 162–169, 2003.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*, 2019.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

Campbell R Harvey and Akhtar Siddique. Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295, 2000.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77 (5):1683–1701, 2009.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

Maximilian Kasy. Using data to inform policy. Technical report, 2014.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Philip W Lavori, Ree Dawson, and A John Rush. Flexible treatment strategies in chronic disease: clinical and research implications. *Biological psychiatry*, 48(6):605–614, 2000.

Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. *arXiv preprint arXiv:1711.03639*, 2017.

Enno Mammen, Alexandre B Tsybakov, et al. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Charles F. Manski. Statistical treatment rules for heterogenous populations. *Econometrica*, 2004.

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

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

Susan A Murphy, Mark J van der Laan, and James M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Annals of Statistics*, 2013.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

Zhengling Qi, Ying Cui, Yufeng Liu, and Jong-Shi Pang. Estimation of individualized decision rules based on an optimized covariate-dependent equivalent of random outcomes. *arXiv preprint arXiv:1908.10742*, 2019.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 1952.

James M Robins. Causal inference from complex longitudinal data. pages 69–117, 1997.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*, 2019.

J. Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151: 70–81, 2009.

Jörg Stoye. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012.

Aleksey Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 2004a.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004b.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

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

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.