Functional Sequential Treatment Allocation

Anders Bredahl Kock\textsuperscript{a,b}, David Preinerstorfer\textsuperscript{c}, and Bezirgen Veliyev\textsuperscript{b}

\textsuperscript{a}Department of Economics, University of Oxford, Oxford, UK; \textsuperscript{b}Department of Economics and Business Economics, CREATES, Aarhus University, Aarhus, Denmark; \textsuperscript{c}ECARES, SBS-EM, Université libre de Bruxelles, Bruxelles, Belgium

ABSTRACT

Consider a setting in which a policy maker assigns subjects to treatments, observing each outcome before the next subject arrives. Initially, it is unknown which treatment is best, but the sequential nature of the problem permits learning about the effectiveness of the treatments. While the multi-armed-bandit literature has shed much light on the situation when the policy maker compares the effectiveness of the treatments through their mean, much less is known about other targets. This is restrictive, because a cautious decision maker may prefer to target a robust location measure such as a quantile or a trimmed mean. Furthermore, socio-economic decision making often requires targeting purpose specific characteristics of the outcome distribution, such as its inherent degree of inequality, welfare or poverty. In the present article, we introduce and study sequential learning algorithms when the distributional characteristic of interest is a general functional of the outcome distribution. Minimax expected regret optimality results are obtained within the subclass of explore-then-commit policies, and for the unrestricted class of all policies. Supplementary materials for this article are available online.

1. Introduction

A fundamental question in statistical decision theory is how to optimally assign subjects to treatments. Important recent contributions include Chamberlain (2000), Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009), Bhattacharya and Dupas (2012), Stoye (2012), Tetenov (2012), Manski and Tetenov (2016), Athey and Wager (2017), Kitagawa and Tetenov (2018), and Manski (2019b); cf. also the overview in Hirano and Porter (2018). In the present article, we focus on assignment problems where the subjects to be treated arrive sequentially. Thus, in contrast to the above mentioned articles, the dataset is gradually constructed during the learning process. In this setting, a policy maker who seeks to assign subjects to the treatment with the highest expected outcome (but who initially does not know which treatment is best), can draw on a rich literature on “multi-armed bandits.” Important contributions include Thompson (1933), Robbins (1952), Gittins (1979), Lai and Robbins (1985), Agrawal (1995), Auer et al. (1995), Audibert and Bubeck (2009); cf. Bubeck and Cesa-Bianchi (2012) and Lattimore and Szepesvári (2020) for introductions to the subject and further references. In many applications, however, the quality of treatments cannot successfully be compared according to the expectation of the outcome distribution: A cautious policy maker may prefer to use another (more robust) measure of location, for example, a quantile or a trimmed mean; or may actually want to make assignments targeting a different distributional characteristic than its location. Examples falling into the latter category are encountered in many socio-economic decision problems, where one wants to target, for example, a welfare measure that incorporates inequality or poverty implications of a treatment. Inference for such “distributional policy effects” has received a great deal of attention in nonsequential settings (e.g., Gastwirth 1974; Manski 1988; Thistle 1990; Mills and Zandvakili 1997; Davidson and Duclos 2000; Abadie 2002; Abadie, Angrist, and Imbens 2002; Chernozhukov and Hansen 2005; Davidson and Flachaire 2007; Barrett and Donald 2009; Hirano and Porter 2009; Schluter and van Garderen 2009; Rostek 2010; Rothe 2010, 2012; Chernozhukov, Fernández-Val, and Melly 2013; Kitagawa and Tetenov 2019; Manski 2019a).\footnote{In contrast to much of the existing theoretical results concerning inference on inequality, welfare, or poverty measures, we do not investigate (first or higher-order) asymptotic approximations, but we establish exact finite sample results with explicit constants. To this end we cannot rely on classical asymptotic techniques, for example, distributional approximations based on linearization arguments.}

Motivated by robustness considerations and the general interest in distributional policy effects, we consider a decision maker who seeks to minimize regret compared to always assigning the unknown best treatment according to a functional of interest. To achieve a low regret, the policy maker must sequentially learn the distributional characteristic of interest for all available treatments, yet treat as many subjects as well as possible.

While most of the multi-armed bandit literature focuses on targeting the treatment with the highest expectation, there are articles going beyond the first moment. This previous work has focused on risk functionals: Maillard (2013) considers problems where one targets a coherent risk measure. Sani, Lazaric, and Munos (2012), Vakili and Zhao (2016), and Vakili, Boukouvalas,
and Zhao (2018) studied a problem targeting the mean-variance functional, that is, the variance minus a multiple of the expectation. Zimin, Ibsen-Jensen, and Chatterjee (2014), motivated by earlier results on problems targeting specific risk measures, and Kock and Thyrgaard (2017) considered problems where one targets a functional that can be written as a function of the mean and the variance. Tran-Thanh and Yu (2014) and Cassel, Mannor, and Zeevi (2018) did not restrict themselves to functionals of the latter type, but consider bandit problems where the target can be a general risk functional. These articles use various types of regret frameworks. Tran-Thanh and Yu (2014) considered a "pure-exploration" regret function into which the errors made during the assignment period do not enter. Maillard (2013), Zimin, Ibsen-Jensen, and Chatterjee (2014), Kock and Thyrgaard (2017) and Vakili, Boukouvalas, and Zhao (2018) considered a "cumulative" regret function that is closely related to the regret used in classical multi-armed bandit problems (i.e., where the expectation is targeted). Sani, Lazaric, and Munos (2012), Vakili and Zhao (2016) and Cassel, Mannor, and Zeevi (2018) considered a "path-dependent" regret function. The just-mentioned articles have in common that pointwise regret upper bounds are derived for certain policies (and the regret considered). Except for Vakili and Zhao (2016) and Vakili, Boukouvalas, and Zhao (2018), who exclusively considered the mean-variance functional, matching lower bounds are not established. Therefore, apart from the mean-variance functional, it remains unclear if the policies developed are optimal. The main goal of the present article is to develop a minimax optimality theory for general functional targets. The regret function we work with is cumulative, and thus has the following important features which are relevant for many socio-economic assignment problems:

- Every subject not assigned to the best treatment contributes to the regret.
- A loss incurred for one subject cannot be compensated by future assignments.

The first bullet point is not satisfied by a "pure-exploration" regret; the second is violated by "path-dependent" regrets.

Our first contribution is to establish minimax expected regret optimality properties within the subclass of "explore-then-commit" policies (cf. Theorems 3.1 and 3.2). These are policies that strictly separate the exploration and exploitation phases: one first attempts to learn the best treatment, for example, by conducting a randomized controlled trial (RCT), on an initial segment of subjects. Based on the outcome, one then assigns all remaining subjects to the inferred best treatment (which is not guaranteed to be the optimal one). Such policies are close to current practice in many socio-economic decision problems. Garivier, Lattimore, and Kaufmann (2016) recently studied optimality properties of explore-then-commit policies in a 2-arm Gaussian setting targeting exclusively the expectation.

Our second contribution is to obtain lower bounds on maximal expected regret over the class of all policies (cf. Theorem 4.2), and to show that they are matched by uniform upper bounds for the following two policies: First, the "F-UCB" policy (an extension of the UCB1 policy of Auer, Cesa-Bianchi, and Fischer (2002)), and second the "F-aMOSS" policy (an extension of the anytime MOSS policy of Degenne and Perchet (2016)), see Theorems 4.1 and 4.3.

Our lower bounds hold under very weak assumptions. Therefore, they settle firmly what can and cannot be achieved in a functional sequential treatment assignment problem.

As a corollary to our results, comparing the regret upper bounds derived for the F-UCB and the F-aMOSS policy to the lower bound obtained for explore-then-commit policies, we reveal that in terms of maximal expected regret all explore-then-commit policies are inferior to the F-UCB and the F-aMOSS policy, and therefore, should not be used if it can be avoided. If an explore-then-commit policy has to be used, our results provide guidance on the optimal length of the exploration period.

In Sections 5 and 6, we provide numerical results (based on simulated and empirical data) comparing the regret-behavior of explore-then-commit policies with that of the F-UCB and the F-aMOSS policy. In this context we develop test-based and empirical-success-based explore-then-commit policies that might be of independent interest, because they provably possess desirable performance guarantees.

Concerning the functionals we permit our theory is very general. We verify in detail that it covers many inequality, welfare, and poverty measures, such as the Schutz coefficient, the Atkinson-, Gini-, and Kolm-indices. This discussion can be found in Appendix E. We also show that our theory covers quantiles, $U$-functionals, generalized $L$-functionals, and trimmed means. These results can be found in Appendix G. The results in these appendices are of high practical relevance, because they allow the policy maker to choose the functional-dependent constants appearing in the optimal policies in such a way that the performance guarantees apply.

All appendices to this article can be found in the online supplementary materials.

In the companion article (Kock, Preinerstorfer, and Veliyev 2020), we address the important but nontrivial question of how to construct policies that optimally incorporate covariate information. The results in the present article are crucial for obtaining those results.

## 2. Setting and Assumptions

We consider a setting, where at each point in time $t = 1, \ldots, n$ a policy maker must assign a subject to one out of $K$ treatments. Each subject is only treated once.2 Thus, the index $t$ can equivalently be thought of as indexing subjects instead of time. The observational structure is the one of a multi-armed bandit problem: After assigning a treatment, its outcome is observed, but the policy maker does not observe the counterfactuals. Having observed the outcomes of treatments $1, \ldots, t - 1$, subject $t$ arrives, and must be assigned to a treatment. The assignment can be based on the information gathered from all previous

---

2 We emphasize that the sequential setting is different from the "longitudinal" or "dynamic" one in, for example, Robins (1997), Lavori, Dawson, and Rush (2000), Murphy, van der Laan, and Robins (2001), and Murphy (2003, 2005), where the same subjects are treated repeatedly.
assignments and their outcomes, and, potentially, randomization. Thus, the dataset is gradually constructed in the course of the treatment program. Without knowing a priori the identity of the “best” treatment, the policy maker seeks to assign subjects to treatments so as to minimize maximal expected regret (defined in Equation (3)).

This setting is a sequential version of the potential outcomes framework with multiple treatments. Note also that restricting attention to problems where only one out of the K treatments can be assigned does not exclude that a treatment consists of a combination of several other treatments (e.g., a combination of several drugs)—one simply defines this combined treatment as a separate treatment at the expense of increasing the set of treatments.

The precise setup is as follows: let the random variable $Y_{i,t}$ denote the potential outcome of assigning subject $t \in \{1, \ldots, n\}$ to treatment $i \in I := \{1, \ldots, K\}$. That is, the potential outcomes of subject $t$ are $Y_t := (Y_{1,t}, \ldots, Y_{K,t})$. We assume that $a \leq Y_{i,t} \leq b$, where $a < b$ are real numbers. Furthermore, for every $t$, let $G_t$ be a random variable, which can be used for randomization in assigning the $t$th subject. Throughout, we assume that $Y_t$ for $t \in \mathbb{N}$ are independent and identically distributed (iid); and we assume that the sequence $G_t$ is iid, and is independent of the sequence $Y_t$. Note that no assumptions are imposed concerning the dependence between the components of each random vector $Y_t$. We think of the randomization measure, that is, the distribution of $G_t$, as being fixed, for example, the uniform distribution on $[0,1]$. We denote the cumulative distribution function (cdf) of $Y_t$ by $F^i \in D_{\text{cdf}}([a, b])$, where $D_{\text{cdf}}([a, b])$ denotes the set of all cdfs $F$ such that $F(a^-) = 0$ and $F(b) = 1$. The cdfs $F^i$ for $i = 1, \ldots, K$ are unknown to the policy maker.

A policy is a triangular array of (measurable) functions $\pi_t := \{\pi_{n,t} : n \in \mathbb{N}, 1 \leq t \leq n\}$. Here $\pi_{n,t}$ denotes the assignment of the $t$th subject out of $n$ subjects. In each row of the array, that is, for each $n \in \mathbb{N}$, the assignment $\pi_{n,t}$ can depend only on previously observed treatment outcomes and randomizations (previous and current). Formally,

$$\pi_{n,t} : ([a, b] \times \mathbb{R})^{t-1} \times \mathbb{R} \to I.$$  

(1)

Given a policy $\pi$ and $n \in \mathbb{N}$, the input to $\pi_{n,t}$ is denoted as $(Z_{t-1}, G_t)$. Here $Z_{t-1}$ is defined recursively: The first treatment $\pi_{n,1}$ is a function of $G_1$ alone, as no treatment outcomes have been observed yet (we may interpret $(Z_0, G_1) = G_1$). The second treatment is a function of $Z_1 := (Y_{\pi_{n,1}(G_1),1}, G_1)$, the outcome of the first treatment and the first randomization, and of $G_2$. For $t \geq 3$, we have

$$Z_{t-1} := (Y_{\pi_{n,1}(Z_{t-2}; G_{t-2}),t-1}, G_{t-1}, Z_{t-2}) = (Y_{\pi_{n,1}(Z_{t-2}; G_{t-2}),t-1}, G_{t-1}, \ldots, Y_{\pi_{n,1}(G_{t-1}),1}, G_{t-1}).$$

(2)

The $2(t-1)$-dimensional random vector $Z_{t-1}$ can be interpreted as the information available after the $(t-1)$th treatment outcome was observed. We emphasize that $Z_{t-1}$ depends on the policy $\pi$ via $\pi_{n,1}, \ldots, \pi_{n,t-1}$. In particular, $Z_{t-1}$ also depends on $n$, which we do not show in our notation. For convenience, the dependence of $\pi_{n,t}(Z_{t-1}, G_t)$ on $Z_{t-1}$ and $G_t$ is often suppressed, that is, we often abbreviate $\pi_{n,t}(Z_{t-1}, G_t)$ by $\pi_{n,t}$ if it is clear from the context that the actual assignment $\pi_{n,t}(Z_{t-1}, G_t)$ is meant, instead of the function defined in Equation (1).

**Remark 2.1 (Concerning the dependence of $\pi_{n,t}$ on the horizon $n$).** We have chosen to allow the assignments $\pi_{n,1}, \ldots, \pi_{n,n}$ to depend on $n$, the total number of assignments to be made. Consequently, for $n_1 < n_2$ it may be that $\{\pi_{n_1,t} : 1 \leq t \leq n_1\}$ does not coincide with the first $n_1$ elements of $\{\pi_{n_2,t} : 1 \leq t \leq n_2\}$. This is crucial, as a policy maker who knows $n$ may choose different sequences of allocations for different $n$. For example, one may wish to explore the efficacies of the available treatments in more detail if one knows that the total sample size is large, such that there is much opportunity to benefit from this knowledge later on. We emphasize that while our setup allows us to study policies that make use of $n$, we devote much attention to policies that do not. The latter subclass of policies is important. For example, a policy maker may want to run a treatment program for a year, say, but it is unknown in advance how many subjects will arrive to be treated. In such a situation, one needs a policy that works well irrespective of the unknown horizon. Such policies are called “anytime policies,” as $\pi_t := \pi_{n,t}$ does not depend on $n$.

The ideal solution of the policy maker would be to assign every subject to the “best” treatment. In the present article, this is understood in the sense that the outcome distribution for the best treatment maximizes a given functional

$$\mathbf{T} : D_{\text{cdf}}([a, b]) \to \mathbb{R}.$$  

(2)

We do not assume that the maximizer is unique, that is, $\max_{F \in D_{\text{cdf}}} \mathbf{T}(F)$ need not be a singleton. The specific functional chosen by the policy maker will depend on the application, and encodes the particular distributional characteristics the policy maker is interested in. For a streamlined presentation of our results it is helpful to keep the functional $\mathbf{T}$ abstract at this point (see Section 2.1 for an example, and a brief overview of examples we study in detail in appendices).

The ideal solution of the policy maker of assigning each subject to the best treatment is infeasible, simply because it is not known in advance which treatment is best. Therefore, every policy will make mistakes. To compare different policies, we define the (cumulative) regret of a policy $\pi$ at horizon $n$ as

$$R_n(\pi) = R_n(\pi; F^1, \ldots, F^K, Z_{n-1}, G_n) = \max_{t \in \mathcal{I}} \left\{ \mathbf{T}(F^t) - \mathbf{T}(F^{\pi_{n,t}(Z_{t-1}, G_t)}) \right\};$$

(3)

that is, for every individual subject that is not assigned to the best treatment one incurs a loss. One important feature of $R_n(\pi)$ is that the losses incurred at time $t$ cannot be nullified by later assignments. As discussed in the introduction, cumulative regret functions have previously been used by Maillard (2013), Zimin, Ibsen-Jensen, and Chatterjee (2014), Kock and Thyrsgaard (2017), and Vakili, Boukouvalas, and Zhao (2018), the latter explicitly emphasizing the practical relevance of this regret.
notion in the context of clinical trials where the loss in each individual assignment needs to be controlled.

The unknown outcome distributions \( F_1, \ldots, F^K \) are assumed to vary in a prespecified class of cdfs. Following the minimax-paradigm, we evaluate policies according to their worst-case behavior over such classes. We refer to Manski and Tetenov (2016) for further details concerning the minimax point-of-view in the context of treatment assignment problems, and for a comparison with other approaches such as the Bayesian. Formally, we seek a policy \( \pi \) that minimizes maximal expected regret, that is, a policy that minimizes

\[
\sup_{F_i \in \mathcal{D}} \mathbb{E}[R_n(\pi)],
\]

where \( \mathcal{D} \) is a subset of \( D_{\text{diff}}([a, b]) \). The supremum is taken over all potential outcome vectors \( Y_i \) such that the marginals \( Y_i, 1 \) for \( i = 1, \ldots, K \) have a cdf in \( \mathcal{D} \). The set \( \mathcal{D} \) will typically be nonparametric, and corresponds to the assumptions one is willing to impose on the cdfs of each treatment outcome, that is, on \( F_1, \ldots, F^K \). Note that the maximal expected regret of a policy \( \pi \) as defined in the previous display depends on the horizon \( n \). We will study this dependence on \( n \). In particular, we will study the rate at which the maximal expected regret increases in \( n \) for a given policy \( \pi \); furthermore, we will study the question of which kind of policy is optimal in the sense that the rate is optimal.

The following assumption is the main requirement we impose on the functional \( T \) and the set \( \mathcal{D} \). We denote the supremum metric on \( D_{\text{diff}}([a, b]) \) by \( \| \cdot \|_{\infty} \), that is, for cdfs \( F \) and \( G \) we let \( \| F - G \|_{\infty} = \sup_{x \in \mathbb{R}} |F(x) - G(x)| \).

**Assumption 2.1.** The functional \( T : D_{\text{diff}}([a, b]) \to \mathbb{R} \) and the nonempty set \( \mathcal{D} \subseteq D_{\text{diff}}([a, b]) \) satisfy

\[
|T(F) - T(G)| \leq C \| F - G \|_{\infty}
\]

for every \( F \in \mathcal{D} \) and every \( G \in D_{\text{diff}}([a, b]) \) for some \( C > 0 \).

**Remark 2.2 (Restricted-Lipschitz continuity).** Assumption 2.1 implies that the functional \( T \) is Lipschitz continuous when restricted to \( \mathcal{D} \) (the domain being equipped with \( \| \cdot \|_{\infty} \)). We emphasize, however, that if \( \mathcal{D} \neq D_{\text{diff}}([a, b]) \), the functional \( T \) is not necessarily required to be Lipschitz-continuous on all of \( D_{\text{diff}}([a, b]) \). This is due to the asymmetry inherent in the condition imposed in Equation (5), where \( F \) varies only in \( \mathcal{D} \), but \( G \) varies in all of \( D_{\text{diff}}([a, b]) \).

**Remark 2.3.** A simple approximation argument\(^4\) shows that if Assumption 2.1 is satisfied with \( \mathcal{D} \) and \( C \), then Assumption 2.1 is also satisfied with \( \mathcal{D} \) replaced by the closure of \( \mathcal{D} \subseteq D_{\text{diff}}([a, b]) \) (the ambient space \( D_{\text{diff}}([a, b]) \) being equipped with the metric \( \| \cdot \|_{\infty} \)) and the same constant \( C \).

\(^4\) Let \( \bar{F} \in D_{\text{diff}}([a, b]) \) be such that \( \| F_m - \bar{F} \|_{\infty} \to 0 \) as \( m \to \infty \) for a sequence \( F_m \in \mathcal{D} \), and let \( G \in D_{\text{diff}}([a, b]) \). Then, \( |T(\bar{F}) - T(G)| \leq |T(\bar{F}) - T(F_m)| + |T(F_m) - T(G)| \), which, by Assumption 2.1, is not greater than \( 2C\| \bar{F} - F_m \|_{\infty} + C\| F_m - G \|_{\infty} \to C\| \bar{F} - G \|_{\infty} \) as \( m \to \infty \).

**Remark 2.4.** The set \( \mathcal{D} \) encodes assumptions imposed on the cdfs of each treatment outcome. In particular, the larger \( \mathcal{D} \), the less restrictive is \( F_i \in \mathcal{D} \) for \( i \in \mathcal{I} \). Ideally, one would thus like \( \mathcal{D} = D_{\text{diff}}([a, b]) \), which, however, is too much to ask for some functionals. Furthermore, there is a trade-off between the sizes of \( \mathcal{C} \) and \( \mathcal{D} \), in the sense that a larger class \( \mathcal{D} \) typically requires a larger constant \( C \). The reader who wants to get an impression of some of the classes of cdfs we consider may want to consult Appendix E.1, where important classes of cdfs are defined.

### 2.1. Functionals That Satisfy Assumption 2.1: A Summary of Results in Appendices E and G

In the present article, we do not contribute to the construction of functionals for specific questions. Rather, we take the functional as given. To choose an appropriate functional, the policy maker can already draw on a very rich and still expanding body of literature; see Lambert (2001), Chakravarty (2009), or Cowell (2011) for textbook-treatments. To equip the reader with a specific and important example of a functional \( T \), one may think of the Gini-welfare measure (cf. Sen 1974)

\[
T(F) = \int xdF(x) - \frac{1}{2} \int \int |x_1 - x_2|dF(x_1)dF(x_2).
\]

Because all of our results impose Assumption 2.1, a natural question concerns its generality. We show in Appendix E that Assumption 2.1 is satisfied for many important inequality, welfare, and poverty measures (together with formal results concerning the sets \( \mathcal{D} \) along with corresponding constants \( C \)). For example, it is shown that for the above Gini-welfare measure, Assumption 2.1 is satisfied with \( \mathcal{D} = D_{\text{diff}}([a, b]) \), that is, without any restriction on the treatment cdfs \( F_1, \ldots, F^K \) (apart from having support \([a, b]) \), and with constant \( C = 2(b - a) \). At this point we highlight some further functionals that satisfy Assumption 2.1:

1. **The inequality measures** we discuss in Appendix E.2 include the Schutz-coefficient (Schutz 1951; Rosenbluth 1951), the Gini-index, the class of linear inequality measures of Mehran (1976), the generalized entropy family of inequality indices including Theil’s index, the Atkinson family of inequality indices (Atkinson 1970), and the family of Kolm-indices (Kolm 1976). In many cases, we discuss both relative and absolute versions of these measures.

2. **In Appendix E.3,** we provide results for welfare measures based on inequality measures.

3. **The poverty measures** we discuss in Section E.4 are the headcount ratio, the family of poverty measures of Sen (1976) in the generalized form of Kakwani (1980), and the family of poverty measures suggested by Foster, Greer, and Thorbecke (1984).

The results in Appendices E.2–E.4 mentioned above are obtained from and supplemented by a series of general results that we develop in Appendix G. These results verify Assumption 2.1 for \( U \)-functionals defined in Equation (G.3) (i.e., population versions of \( U \)-statistics, e.g., the mean or the variance), quantiles, generalized \( L \)-functionals due to Serfling (1984) defined in Equation (G.13), and trimmed \( U \)-functionals defined
in Equation (G.18). These techniques are of particular interest in case one wants to apply our results to functionals T that we do not explicitly discuss in Appendix E.\footnote{The results in Appendices E and G could also be of independent interest, because they immediately allow the construction of uniformly valid (over \( \mathcal{D} \)) confidence intervals and tests in finite samples.}

### 2.2. Further Notation and an Additional Assumption

Given a policy \( \pi \) and \( n \in \mathbb{N} \), we denote by

\[
S_{i,n}(t) := \sum_{j=1}^{t} \mathbb{I}[\pi_{n,j}(Z_{t-1}, G_i) = i]
\]

the number of times treatment \( i \) has been assigned up to time \( t \), and we abbreviate \( S_{i,n}(n) = S_i(n) \). Defining the loss incurred due to assigning treatment \( i \) instead of an optimal one by \( \Delta_i := \max_{k \in \mathcal{I}} T(F^k) - T(F^i) \), the regret \( R_n(\pi) \), which was defined in Equation (3), can equivalently be written as

\[
R_n(\pi) = \sum_{i : \Delta_i > 0} \sum_{t=1}^{n} \mathbb{I}[\pi_{n,t}(Z_{t-1}, G_i) = i] = \sum_{i : \Delta_i > 0} \Delta_i S_i(n).
\]

On the event \( \{S_{i,n}(t) > 0\} \) we define the empirical cdf based on the outcomes of all subjects in \( \{1, \ldots, t\} \) that have been assigned to treatment \( i \)

\[
\hat{F}_{i,t,n}(x) := S_{i,n}^{-1}(t) \sum_{s \leq t} \mathbb{I}[Y_{i,s} \leq x], \quad \text{for every } x \in \mathbb{R},
\]

where \( \pi_{n,s}(Z_{s-1}, G_i) = i \) depend on previously observed treatment outcomes.

We shall frequently need an assumption that guarantees that the functional \( T \) evaluated at empirical cdfs, such as \( \hat{F}_{i,t,n} \) just defined in Equation (9), is measurable.

**Assumption 2.2.** For every \( m \in \mathbb{N} \), the function on \( [a, b]^m \) that is defined via \( x \mapsto T(m^{-1} \sum_{j=1}^{m} \mathbb{I}[x_j \leq \{\cdot\}], \text{that is, } T \)

Finally, and following up on the discussion in Remark 2.1, we shall introduce some notational simplifications in case a policy \( \pi \) is such that \( \pi_{n,t} \) is independent of \( n \), that is, is an anytime policy. It is then easily seen that the random quantities \( S_{i,n}(t) \) and \( \hat{F}_{i,t,n} \) do not depend on \( n \) (as long as \( t \) and \( n \) are such that \( n \geq t \)). Therefore, for such policies, we shall drop the index \( n \) in these quantities.

### 3. Explore-Then-Commit Policies

A natural approach to assigning subjects to treatments in our sequential setup would be to first conduct a randomized controlled trial (RCT) to study which treatment is best, and then to use the acquired knowledge to assign the inferred best treatment to all remaining subjects. Such policies are special cases of explore-then-commit policies, which we study in this section.

Informally, an explore-then-commit policy deserves its name as it (i) uses the first \( n_1 \) subjects to explore, in the sense that every treatment is assigned, in expectation, at least proportionally to \( n_1 \); and (ii) then commits to a single (inferred best) treatment after the first \( n_1 \) treatments have been used for exploration. Here, \( n_1 \) may depend on the horizon \( n \). Formally, we define an explore-then-commit policy as follows.

**Definition 3.1 (Explore-then-commit policy).** A policy \( \pi \) is an explore-then-commit policy, if there exists a function \( n_1 : \mathbb{N} \rightarrow \mathbb{N} \) and an \( \eta \in (0,1) \), such that for every \( n \in \mathbb{N} \) we have that \( n_1(n) \leq n \), and such that the following conditions hold for every \( n \geq K \):

1. Exploration condition: We have that

\[
\inf_{i=1}^{K} \inf_{v \in \mathcal{D}} \mathbb{E}[S_{i,n_1(n)}] \geq \eta n_1(n).
\]

Here, the first infimum is taken over all potential outcome vectors \( Y_t \) such that the marginals \( Y_{i,t} \) for \( i = 1, \ldots, K \) have a cdf in \( \mathcal{D} \).

[That is, regardless of the (unknown) underlying marginal distributions of the potential outcomes, each treatment is assigned, in expectation, at least \( \eta n_1(n) \) times among the first \( n_1(n) \) subjects.]

2. Commitment condition: There exists a function \( \pi_n^c : ([a, b] \times \mathbb{R})^{n_1(n)} \rightarrow \mathcal{I} \) such that, for every \( t = n_1(n) + 1, \ldots, n \), we have

\[
\pi_{n,t}(\varepsilon_{t-1}, G) = \pi_n^c(\varepsilon_{n_1(n)}) \quad \text{for every } \varepsilon_{t-1} \in ([a, b] \times \mathbb{R})^{t-1}
\]

and every \( g \in \mathbb{R} \),

\[
\pi_{n,t}(\varepsilon_{t-1}, G) = \pi_n^c(\varepsilon_{n_1(n)}) \quad \text{for every } \varepsilon_{t-1} \in ([a, b] \times \mathbb{R})^{t-1}
\]

where \( \varepsilon_{n_1(n)} \) is the vector of the last \( 2n_1(n) \) coordinates of \( z_{t-1} \).

[That is, the subjects \( t = n_1(n) + 1, \ldots, n \) are all assigned to the same treatment, which is selected based on the \( n_1(n) \) outcomes and randomizations observed during the exploration period.]

It would easily be possible to let the commitment rule \( \pi_n^c \) depend on further external randomization. For simplicity, we omit formalizing such a generalization. We shall now discuss some important examples of explore-then-commit policies.

**Example 3.1.** A policy that first conducts an RCT based on a sample of \( n_1(n) \leq n \) subjects, followed by any assignment rule for subjects \( n_1(n) + 1, \ldots, n \) that satisfies the commitment condition in Definition 3.1, is an explore-then-commit policy, provided the concrete randomization scheme used in the RCT encompasses sufficient exploration. In particular, \( \pi_{n,t}(Z_{t-1}, G_i) = G_i \) with \( \mathbb{P}(G_i = i) := \frac{1}{K} \) for every \( 1 \leq t \leq n_1(n) \) and every \( i \in \mathcal{I} \) satisfies the exploration condition in Definition 3.1 with \( \eta = \frac{K}{n_1(n)} \); more generally, Definition 3.1 holds if \( \eta := \inf_{i \in \mathcal{I}} \mathbb{P}(G_i = i) > 0 \). Alternatively, a policy that enforces balancedness in the exploration phase through assigning subjects \( t = 1, \ldots, n_1(n) \) to treatments “cyclically,” that is, \( \pi_{n,t}(Z_{t-1}, G_i) = (t \mod K) + 1 \), satisfies the exploration
condition in Definition 3.1 with \( \eta = 1/(2K) \) if \( n_1(n) \geq K \) for every \( n \geq K \). Concrete choices of commitment rules could be based on the outcome of a hypothesis test (choosing \( n_1 \) based on a power calculation), or empirical success rules (cf. Manski 2004) assigning an element of \( \arg \max_{\tau \in \mathcal{T}, \mathcal{N}(\mathcal{T}, n_1(n), n) \rightarrow \mathbb{R}} \) to subjects \( n_1(n) + 1, \ldots, n \). In Section 5.1, we develop two such concrete rules with regret guarantees.

We now establish regret lower bounds for the class of explore-then-commit policies. To exclude trivial cases, we assume that \( \mathcal{D} \) (which is typically convex) contains a line segment on which the functional \( T \) is not everywhere constant.

**Assumption 3.1.** The functional \( T : D_{\text{cf}}([a, b]) \rightarrow \mathbb{R} \) satisfies Assumption 2.1, and \( \mathcal{D} \) contains two elements \( H_1 \) and \( H_2 \), such that

\[
J := \tau H_1 + (1 - \tau) H_2 \in \mathcal{D} \quad \text{for every } \tau \in [0, 1],
\]

and such that \( T(H_1) \neq T(H_2) \).

Since there only have to exist two cdfs \( H_1 \) and \( H_2 \) as in Assumption 3.1, this is a condition that is practically always satisfied.

The next theorem considers general explore-then-commit policies, as well as the subclass of policies where \( n_1(n) \leq n^* \) holds for every \( n \in \mathbb{N} \) for some \( n^* \in \mathbb{N} \). This subclass models situations, where the horizon \( n \) is unknown or ignored in planning the experiment, and the envisioned number of subjects used for exploration \( n^* \) is fixed in advance (here \( n_1(n) = n^* \) for every \( n \geq n^* \)), and \( n_1(n) = n \), else); the subclass also models situations where the sample size that can be used for experimentation is limited due to budget constraints.

**Theorem 3.1.** Suppose \( K = 2 \) and that Assumption 3.1 holds. Then the following statements hold:

1. There exists a constant \( c_1 > 0 \), such that, for every explore-then-commit policy \( \pi \) that satisfies the exploration condition with \( \eta \in (0, 1) \), and for any randomization measure, it holds that

\[
\sup_{\pi \in \mathcal{D}, \tau \in (0, 1]} \mathbb{E}[R_n(\pi)] \geq c_1 \eta n^{2/3} \quad \text{for every } n \geq 2.
\]

2. For every \( n^* \in \mathbb{N} \) there exists a constant \( c_1 = c_1(n^*) \), such that, for every explore-then-commit policy \( \pi \) that satisfies (i) the exploration condition with \( \eta \in (0, 1) \) and (ii) \( n_1(\cdot) \leq n^* \), and for any randomization measure, it holds that

\[
\sup_{\pi \in \mathcal{D}, \tau \in (0, 1]} \mathbb{E}[R_n(\pi)] \geq c_1 n \quad \text{for every } n \geq 2.
\]

The first part of Theorem 3.1 shows that, under the minimal assumption of \( \mathcal{D} \) containing a line segment on which \( T \) is not constant, any explore-then-commit policy must incur maximal expected regret that increases at least of order \( n^{2/3} \) in the horizon \( n \).

The second part implies in particular that when \( n \) is unknown, such that the exploration period \( n_1 \) cannot depend on it, any explore-then-commit policy must incur linear maximal expected regret. We note that this is the worst possible rate of regret, since by Assumption 2.1 no policy can have larger than linear maximal expected regret.

The lower bounds on maximal expected regret are obtained by taking the maximum only over all potential outcome vectors with marginal distributions in the line segment in Equation (11). This is a one-parametric subset of \( \mathcal{D} \) over which \( T \) nevertheless varies sufficiently to obtain a good lower bound.

We now prove that a maximal expected regret of rate \( n^{2/3} \) is attainable in the class of explore-then-commit policies, that is, we show that the lower bound in the first part of Theorem 3.1 cannot be improved upon. We consider the following policy.

**Policy 1:** Explore-then-commit empirical-success policy \( \widehat{\pi} \)

\[
\begin{align*}
&\text{for } t = 1, \ldots, n_1(n) := \min(K \left\lceil n^{2/3} \right\rceil, n) \text{ do} \\
&\quad \text{assign } \widehat{\pi}_{n, t}(Z_{t-1}, G_t) = G_t, \text{ with } G_t \text{ uniformly distributed on } I \\
&\text{end} \\
&\text{for } t = n_1(n) + 1, \ldots, n \text{ do} \\
&\quad \text{assign } \widehat{\pi}_{n, t}(Z_{t-1}, G_t) = \min \arg \max_{\pi \in \mathcal{D}, \tau \in (0, 1]} \mathbb{E}[\mathcal{F}_1(n_1(n), n)] \\
&\text{end}
\end{align*}
\]

Note that the policy \( \widehat{\pi} \) is an explore-then-commit policy that requires knowledge of the horizon \( n \), which by Theorem 3.1 is necessary for obtaining a rate slower than \( n \). The outer minimum in the second for loop in the policy is just taken to break ties (if necessary). Our result concerning \( \widehat{\pi} \) is as follows (an identical statement can be established for a version of \( \widehat{\pi} \) with cyclical assignment during the exploration phase as discussed in Remark 3.1; the proof follows along the same lines, and we skip the details).

**Theorem 3.2.** Under Assumptions 2.1 and 2.2, the explore-then-commit empirical-success policy \( \widehat{\pi} \) satisfies

\[
\sup_{\pi \in \mathcal{D}, \tau \in (0, 1]} \mathbb{E}[R_n(\widehat{\pi})] \leq 6CKn^{2/3} \quad \text{for every } n \in \mathbb{N}.
\]

Theorems 3.1 and 3.2 together prove that within the class of explore-then-commit policies, the policy \( \widehat{\pi} \) is rate optimal in \( n \). An upper bound as in Theorem 3.2 for the special case of the mean functional can be found in Chapter 6 of Lattimore and Szepesvári (2020). We shall now show that policies which do not separate the exploration and commitment phase can obtain lower maximal expected regret. In this sense, the natural idea of separating exploration and commitment phases turns out to be suboptimal from a decision-theoretic point-of-view in functional sequential treatment assignment problems.\footnote{The finding that for large classes of functional targets explore-then-commit policies are suboptimal in terms of maximal expected regret does, of course, by no means discredit RCTs and subsequent testing for other}
4. Functional UCB-Type Policies and Regret Bounds

In this section, we define and study two policies based on upper-confidence-bounds. We start with the functional upper confidence bound (F-UCB) policy. It is inspired by the UCB1 policy of Auer, Cesa-Bianchi, and Fischer (2002) for multi-armed bandit problems targeting the mean. Extensions of the UCB1 policy to targeting risk functionals have been considered by Sani, Lazaric, and Munos (2012), Maillard (2013), Zimin, Ibsen-Jensen, and Chatterjee (2014), Vakili and Zhao (2016), and Vakili, Boukouvalas, and Zhao (2018). The F-UCB policy can target any functional (and reduces to the UCB1 policy in case one targets the mean). It has the practical advantage of not needing to know the horizon $n$, cf. Remark 2.1 (recall also the notation introduced in Section 2.2). Furthermore, no external randomization is required, which will therefore be notionally suppressed as an argument to the policy. The policy is defined as follows, where $C$ is the constant from Assumption 2.1.

**Policy 2:** F-UCB policy $\hat{\pi}$

**Input:** $\beta > 2$

**for** $t = 1, \ldots, K$ **do**

| assign $\hat{\pi}_i(Z_{t-1}) = t$ |

**end**

**for** $t \geq K + 1$ **do**

| assign $\hat{\pi}_i(Z_{t-1}) = \min \arg \max_{i \in \mathcal{I}} \{ T(\hat{\pi}_i(Z_{t-1})) + C \sqrt{\beta \log(t)/(2S_i(t-1))} \} |

**end**

After the $K$ initialization rounds, the F-UCB policy assigns a treatment that (i) is promising, in the sense that $T(\hat{\pi}_i(Z_{t-1}))$ is large, or (ii) has not been well explored, in the sense that $S_i(t-1)$ is small. The parameter $\beta$ is chosen by the researcher and indicates the weight put on assigning scarcely explored treatments, that is, treatments with low $S_i(t-1)$. We use the notation $\log(x) := \max(\log(x), 1)$ for $x > 0$.

**Theorem 4.1.** Under Assumptions 2.1 and 2.2, the F-UCB policy $\hat{\pi}$ satisfies

$$\sup_{\pi \in \mathcal{P}} \mathbb{E}[R_n(\pi)] \leq c\sqrt{Kn \log(n)} \quad \text{for every } n \in \mathbb{N},$$

where $c = c(\beta, C) = C \sqrt{2\beta + (\beta + 2)/(\beta - 2)}$.

Note that the choice $\beta = 2 + \sqrt{2}$ minimizes $c(\beta, C)$ and implies $c \leq \sqrt{4TC}$.

In case of the mean functional, an upper bound as in Theorem 4.1 can be obtained from Theorem 1 in Auer, Cesa-Bianchi, and Fischer (2002) as explained after Theorem 2 in Audibert and Bubeck (2009), see also the discussion in Section 2.4.3 of Bubeck and Cesa-Bianchi (2012). The proof of Theorem 4.1 is inspired by their arguments. However, we cannot exploit the specific structure of the mean functional and related concentration inequalities. Instead we rely on the high-level condition of Assumption 2.1 and the Dvoretzky–Kiefer–Wolfowitz–Massart inequality as established by Massart (1990) to obtain suitable concentration inequalities, see Equation (G.1) in Appendix G. Since adaptive sampling introduces dependence, we also need to take care of the fact that the empirical cdfs defined in (9) are not directly based on a fixed number of iid random variables. This is done via the optional skipping theorem of Doob (1936), see Appendix C.2.1. For functionals that can be written as a Lipschitz-continuous function of the first and second moment (a situation where Assumption 2.1 holds), an upper bound of the same order as in Theorem 4.1 has been obtained in Kock and Thyrsgaard (2017) for a successive-elimination type policy.

The lower bound in Theorem 3.1 combined with the upper bound in Theorem 4.1 shows that the maximal expected regret incurred by any explore-then-commit policy grows much faster in $n$ than that of the F-UCB policy. What is more, the F-UCB policy achieves this without making use of the horizon $n$. Thus, in particular when $n$ is unknown, a large improvement is obtained over any explore-then-commit policy, as the order of the regret decreases from $n$ to $\sqrt{n\log(n)}$. Hence, in terms of maximal expected regret, the policy maker is not recommended to separate the exploration and commitment phases.

Theorem 4.1 leaves open the possibility that one can construct policies with even slower growth rates of maximal expected regret. We now turn to establishing a lower bound on maximal expected regret within the class of all policies. In particular, the theorem also applies to policies that incorporate the horizon $n$.

**Theorem 4.2.** Suppose $K = 2$ and that Assumption 3.1 holds. Then there exists a constant $c_1 > 0$, such that for any policy $\pi$ and any randomization measure, it holds that

$$\sup_{\pi \in \mathcal{P}_{[\{1,2\}^n]}} \mathbb{E}[R_n(\pi)] \geq c_1 n^{1/2} \quad \text{for every } n \in \mathbb{N}. \quad (14)$$

Under the same assumptions used to establish the lower bound on maximal expected regret in the class of explore-then-commit policies, Theorem 4.2 shows that any policy must incur maximal expected regret of order at least $n^{1/2}$. In combination with Theorem 4.1 this shows that, up to a multiplicative factor of $\sqrt{\log(n)}$, no policy exists that has a better dependence of maximal expected regret on $n$ than the F-UCB policy. In this sense the F-UCB policy is near minimax (rate-) optimal.

For the special case of the mean functional a lower bound as in Theorem 4.2 was given in Theorem 7.1 in Auer et al. (1995). Their proof is based on suitably chosen Bernoulli cdfs with parameters about $1/2$, and thus provides a lower bound over all sets $\mathcal{P}$ containing these cdfs, in particular over $\mathcal{P}_{[\{0,1\}]}$. Depending on the functional considered, however, Bernoulli

---

8High-probability bounds as in Theorem 8 in Audibert, Munos, and Szepesvári (2009) can also be obtained for the F-UCB policy, see Theorem C.3 in Appendix C.2.3.
Theorem 4.3. Under Assumptions 2.1 and 2.2, the F-aMOSS policy satisfies
\[ \sup_{\beta > 1/4} \frac{d(\beta)}{\sqrt{\ln n}} = \frac{(\beta W_0(\beta/4\beta))^{1/2}}{1 - (4\beta W_0(\beta/4\beta))^{1/2}} \leq C(4.83 + 6.66 \times d(\beta) + 2\sqrt{\ln n}), \]
for every \( n \in \mathbb{N} \), where \( W_0 \) the inverse of \( w \mapsto \exp(w) \) on \((0, \infty)\).

To prove the result, we generalize to the functional setup a novel argument recently put forward by Garivier et al. (2018) for obtaining a regret upper bound for the anytime MOSS policy of Degenne and Perchet (2016). As in the proof of Theorem 4.1, we need to replace arguments relying on concentration inequalities for the mean, and rely heavily on optional skipping arguments. Furthermore, in contrast to Garivier et al. (2018), we do not only consider the case \( \beta = 1/2 \), but we show that the argument actually goes through for \( \beta > 1/4 \), also expanding the range \( \beta > 1/2 \) considered in Degenne and Perchet (2016). This establishes theoretical guarantees for parameter values close to 1/4, which turned out best in their numerical results (but for which no regret guarantees were provided). Finally, we note that while the upper bound just given is of the order \( \sqrt{n} \), and improves on the upper bound for the F-UCB policy in this sense, this is bought at the price of a larger multiplicative constant than that obtained in Theorem 4.1.

5. Numerical Illustrations

We now illustrate the theoretical results established in this article by means of simulation experiments. Throughout this section, the treatment outcome distributions \( F \) will be taken from the Beta family, a parametric subset of \( D_{\text{diff}}([0, 1]) \), which has a long history in modeling income distributions (see, e.g., Thurow 1970; McDonald 1984; McDonald and Ransom 2008). An appealing characteristic of the Beta family is its ability to replicate many “shapes” of distributions. We emphasize that the policies investigated do not exploit that the unknown treatment outcome distributions are elements of the Beta family.

Our numerical results cover different functionalities \( T \), with a focus on situations where the policy maker targets the distribution that maximizes welfare, and where we consider the case \( a = 0 \) and \( b = 1 \). In all our examples the feasible set for the marginal distributions of the treatment outcomes is \( \mathcal{D} = D_{\text{diff}}([0, 1]) \).

The specific welfare measures we consider are as follows (and correspond to the Gini-, Schutz-, and Atkinson-inequality measure, respectively, through the transformations detailed in Appendix E,3, to which we refer the reader for more background information):

1. Gini-index-based welfare measure: \( W(F) = \mu(F) - \frac{1}{2} \int \int |x_1 - x_2|dF(x_1)dF(x_2) \), where \( \mu(F) := \int xdF(x) \) denotes the mean of \( F \).

2. Schutz-coefficient-based welfare measure: \( W(F) = \mu(F) - \frac{1}{2} \int \frac{F(x)}{(1-F(x))}dx \). [Noting that \( W(F) = \mu(F) - S_{\text{Sd}}(F) \) with \( S_{\text{Sd}} \) as defined in Equation (E.2), Lemmas E.1 and E.9 show that Assumption 2.1 holds with \( D = D_{\text{diff}}([0, 1]) \) and \( C = 2 \).]

3. Atkinson-index-based welfare measure: \( W(F) = \frac{\int x^{1-\epsilon}dF(x)}{(1-\epsilon)} \) for a parameter \( \epsilon \in (0, 1) \cup (1, \infty) \). [Restricting attention to \( \epsilon \in (0, 1) \), the mean value theorem along with Example G.4 in Appendix G yield that Assumption 2.1 is satisfied with \( D = D_{\text{diff}}([0, 1]) \) and \( C = \frac{1}{1-\epsilon} \). We shall consider \( \epsilon \in (0.1, 0.5] \).

In this section, we compare the performance of explore-then-commit policies which do not incorporate \( n \) with the F-UCB and the F-aMOSS policy (which also do not incorporate \( n \)). For numerical results concerning explore-then-commit policies that optimally incorporate \( n \), we refer to Appendix C.6. Throughout this section (as well as in Appendix C.6), we consider the

---

9Interestingly, in the special case of the mean functional (with \( D_{\text{diff}}([0, 1]) \)), Theorem 4.3 shows that the multiplicative constant 113 given in Theorem 3 of Degenne and Perchet (2016) for \( \beta = 2.35/2 \) can be improved to \( (4.83 + 6.66 \times d(2.35/2) + \sqrt{235})/\sqrt{77} \approx 32.5 \).
case of \( K = 2 \) treatments. In the following, the symbol \( W \) shall denote one of the welfare measures just defined in the above enumeration.

### 5.1. Implementation Details

While the F-UCB policy is implemented as in Policy 2 of Section 4 with \( \beta = 2.01 \), and the F-aMOSS policy is implemented as in Policy 3 with \( \beta = 1/3.99 \), the concrete development of explore-then-commit policies with certain performance guarantees requires some additional work which we develop next.

In all explore-then-commit policies we consider, Treatments 1 and 2 are assigned cyclically in the exploration period. This ensures that the number of assignments to each treatment differs and 2 are assigned cyclically in the exploration period. This guarantees that the policy maker must still choose (i) the length of the exploration period \( n_1 \), and (ii) a commitment rule to be used after the exploration phase. The choice of \( n_1 \) (while independent of \( n \)) depends on the commitment rule, of which we now develop a test-based and an empirical-success-based variant:

1. ETC-T: Given \( \alpha \in (0,1) \), at the end of the exploration phase one rejects the hypothesis \( W(F^1) = W(F^2) \) if \( |W(F_{1,n_1}) - W(F_{2,n_1})| \gtrsim \alpha, \) where \( \alpha = \sqrt{2 \log(4/\alpha)C^2/[n_1/2]} \). If one rejects, the commitment rule is to assign arg \( \max_{1 \leq i \leq 2} W(F_{i,n_1}) \), otherwise one randomizes the treatment assignment with equal probabilities. This test has size \( \alpha \) (see Appendix C.3.1), and detects a deviation of \( \Delta := |T(F^1) - T(F^2)| \) with probability at least \( 1 - \eta \) if \( n_1 = 2 \frac{8 \log(4/\min(\alpha, \eta))C^2}{\Delta^2} \) (see Appendix C.3.2).

   In our numerical studies we set \( \eta = \alpha = 0.1 \). We consider \( \Delta \in [0.15, 0.30] \).

   Note that while choosing \( \Delta \) small allows one to detect small differences in the functionals by the above test, this comes at the price of a larger \( n_1 \). Thus, we shall see that neither \( \Delta = 0.15 \) nor \( \Delta = 0.30 \) dominates the other uniformly (over \( t \in \mathbb{N} \)) in terms of maximal expected regret. Finally, we sometimes make the dependence of ETC-T on \( \Delta \) explicit by writing ETC-T(\( \Delta \)).

2. ETC-ES: This policy assigns \( \pi_\eta^t(Z_{n_1}) := \min \arg \max_{1 \leq i \leq K} W(F_{i,n_1}) \) to subjects \( t = n_1 + 1, \ldots, n \), which is an empirical success commitment rule inspired by Manski (2004) and Manski and Tetenov (2016). Here, given a \( \delta > 0 \), \( n_1 \) is chosen such that the maximal expected regret for every subject to be treated after the exploration phase is at most \( \delta \); that is, \( n_1 \) satisfies

   \[
   \sup_{\varphi \in \Theta} \mathbb{E} \left( \max_{t \in \mathscr{I}} W(F^t) - W(F_{\eta}^t(Z_{n_1})) \right) \leq \delta.
   \]

   We prove in Appendix C.3.3 that \( n_1 = K \lceil 16(K-1)^2C^2/(\delta^2 \exp(1)) \rceil \) suffices.

---

10Investigating policies with randomized assignment in the exploration phase would necessitate running the simulations repeatedly, averaging over different draws for the assignments in the exploration phase. The numerical results are already quite computationally intensive, which is why we only investigate a cyclical assignment scheme. This scheme already reflects to a good extent the average behavior of a randomized assignment with equal assignment probabilities.

In our numerical results, we consider \( \delta \in [0.15, 0.30] \), which should be contrasted to the treatment outcomes taking values in \([0,1]\). Note that the \( n_1 \) required to guarantee a maximal expected regret of at most \( \delta \) for every subject treated after the exploration phase is decreasing in \( \delta \). Thus, we shall see that it need not be the case that choosing \( \delta \) smaller will result in lower overall maximal expected regret. We sometimes make the dependence of ETC-ES on \( \delta \) explicit by writing ETC-ES(\( \delta \)).

For \( t = 1, \ldots, 100,000 \), we determine numerically the maximal expected regret when \( F^t \) varies in the Beta family. Since maximizing expected regret over all Beta distributions would be numerically infeasible, we have chosen to maximize expected regret over a subset of all Beta distributions. For a detailed description of this subset and the implementation, see Appendix C.4. We stress that since none of the three policies above needs to know \( n \), the numerical results also contain the maximal expected regret of the policies for any sample size less than \( n = 100,000 \).

### 5.2. Results

The left panel of Figure 1 illustrates the maximal expected regret for the F-UCB, F-aMOSS, ETC-T and ETC-ES policies in the case of Gini-welfare. Each point on the six graphs is the maximum of expected regret over the 210 different distributions considered at a given \( t \). In accordance with Theorems 3.1, 4.1, and 4.3, the maximal expected regret of the policies in the explore-then-commit family is generally higher than the one of the F-UCB and the F-aMOSS policy. For \( t = 100,000 \), the maximal expected regret of F-UCB is 498, and 159 for F-aMOSS, while the corresponding numbers for ETC-T(0.15), ETC-ES(0.15), ETC-T(0.30) and ETC-ES(0.30) are 4248, 777, 7281, and 836, respectively. Note also that no matter the values of \( \Delta \) and \( \delta \), the maximal expected regret of ETC-ES(\( \delta \)) is much lower than the one of the ETC-T(\( \Delta \)) policy.11 In fact, we shall see for all functionals considered that the F-aMOSS policy generally incurs the lowest maximal expected regret, followed by the F-UCB policy and subsequently by the ETC-ES policies, which in turn perform much better than ETC-T policies.

The shape of the graphs of the maximal expected regret of the explore-then-commit policies can be explained as follows: in the exploration phase maximal expected regret is attained by a distribution \( P_1 \), say, for which the value of the Gini-welfare differs strongly at the marginals. However, such distributions are also relatively easy to distinguish, such that none of the commitment rules (testing or empirical success) assigns the suboptimal treatment after the exploration phase. This results in no more regret being incurred and thus a horizontal part on the maximal expected regret graph. For \( t \) sufficiently large, however, maximal expected regret will be attained by a distribution \( P_2 \), say, for which the marginals are sufficiently “close” to imply that the commitment rules occasionally assign the suboptimal treatment. For such a distribution, the expected regret curve will have a positive linear increase even after the commitment
time \(n_1\) and this curve will eventually cross the horizontal part of the expected regret curve pertaining to \(P_1\). This implies that maximal expected regret increases again (as seen for ETC-T(0.30) around \(t = 14,000\) and ETC-ES(0.30) around \(t = 23,000\) in the left panel of Figure 1). Such a kink also occurs for ETC-T(0.15) and eventually also for ETC-ES(0.15). Thus, the left panel of Figure 1 illustrates the tension between choosing \(n_1\) small to avoid incurring high regret in the exploration phase and, on the other hand, choosing \(n_1\) large to ensure making the correct decision at the commitment time.

We next turn to the welfare measures in the Atkinson family. The right panel of Figure 1 contains the results for the case of \(\varepsilon = 0.1\). While F-aMOSS incurs the lowest maximal expected regret uniformly over \(t = 1, \ldots, 100,000\), the most remarkable feature of the figure is that maximal expected regret of all explore-then-commit policies is eventually increasing within the sample considered. The reason for this is that \(\varepsilon = 0.1\) implies a low value of \(n_1\) such that (i) the steep increase in maximal expected regret becomes shorter and (ii) more mistakes are made at the commitment time. The ranking of the families of policies is unaltered with F-UCB and F-aMOSS dominating ETC-ES, which in turn incurs much lower regret than ETC-T. The maximal expected regret results for the Atkinson-welfare when \(\varepsilon = 0.5\), as well as the Schutz-welfare, are similar to the ones for the Gini-welfare, see Figure C.1 in Appendix C.5.

### 6. Illustrations With Empirical Data

We here compare the performance of the policies using three (nonsequentially generated) empirical datasets, each containing the outcomes of a treatment program. From every dataset, we generate synthetic sequential data by sampling from the empirical cdfs corresponding to the treatment/control groups. That is, the empirical cdfs in the datasets are taken as the respective (unknown) treatment outcome distributions \(F_1, \ldots, F_K\), from which observations are then drawn sequentially. This approach allows us to study the policies’ performance on cdfs resembling specific characteristics arising in large scale empirical applications. The datasets considered are as follows; cf. also Appendix D.

1. **The Cognitive Abilities** program studied in Hardy et al. (2015). In this RCT, the participants were split into a treatment group who participated in an online training program targeting various cognitive capacities and an active control group solving crossword puzzles. Thus, \(K = 2\). The outcome variable is a neuropsychological performance measure.
2. **The Detroit Work First** program studied in Autor and Houseman (2010) and Autor, Houseman, and Kerr (2017). Here low-skilled workers took temporary help jobs, direct hire jobs or exited the program. Thus, \(K = 3\). The outcome variable is the total earnings in quarters 2–8 after the start of the program.
3. **The Pennsylvania Reemployment Bonus** program studied originally in Bilias (2000) and also in, for example, Chernozhukov et al. (2018). The participants in the program are unemployed individuals who are either assigned to a control group, or to one of five treatment groups who receive a cash bonus if they find and retain a job within a given qualification period. Thus, \(K = 6\). The size of the cash bonus and the length of the qualification period vary across the five treatment groups. The outcome variable is unemployment duration which varies from 1 to 52 weeks.

To facilitate the comparison with the other results in the previous section, all data were scaled to \([0, 1]\), and we consider the Gini-, Schutz-, and two Atkinson-welfare measures. We focus on the Gini-based-welfare, and report the results for the remaining functionals in Appendix D. The reported expected regrets are averages over 100 replications, with \(n = 100,000\) in each setup. When interpreting the results, it is important to keep in mind that in contrast to the maximal expected regret studied in Section 5, where for each \(t\) the worst-case regret over a certain
family of distributions is reported, the focus is now on three particular datasets, that is, three instances of pointwise expected regret w.r.t. fixed distributions.

For the Detroit Work First Program, for which \( K = 3 \), the ETC-ES policies are implemented as in Section 5.1 with \( \delta \in \{0.15, 0.30\} \). For the Pennsylvania Reemployment Bonus experiment, where \( K = 6 \), this rule led to exploration periods exceeding \( n = 100,000 \). Hence, we instead considered exploration periods assigning 250, 500, 750, and 1000 observations to each of the six arms, respectively. We only implemented the ETC-T policies for the cognitive ability program for which \( K = 2 \).12 The F-UCB and F-aMOSS policies are implemented as in Section 5.

The results for the Gini-welfare measure are summarized in Figure 2. The main take-aways are: (i) F-aMOSS performs solidly across all datasets. (ii) Except for the cognitive training program data, the F-UCB policy is not among the best. Note that this is not in contradiction to the theoretical results of this article (nor the simulations in Section 5) as these are concerned with the worst-case performance of the policies. (iii) There always exists an exploration horizon such that an ETC-ES policy incurs a low expected regret (over the sample sizes considered). However, this horizon is data dependent: for the cognitive and Pennsylvania datasets long exploration horizons are preferable, while for the Work First data the opposite is the case. From the

---

12This is justified by the fact that these policies were always inferior in Section 5. Furthermore, implementing the ETC-T policies when \( K > 2 \) would require taking a stance on how to control the size of the (multiple) testing problem at the commitment time.
Pennsylvania data it is also seen that the optimal length of the exploration horizon depends on the length of the program. The figures containing the results for the remaining functionalities are contained in Appendix D. For the Schultz- and Atkinson-based welfare measure with $\epsilon = 0.5$ the results are qualitatively similar to those of the Gini-based welfare. Regarding the Atkinson-based welfare with $\epsilon = 0.1$, F-aMOSS and F-UCB now even incur the lowest regret for the cognitive data. For the Work First and Pennsylvania data F-aMOSS remains best (at the end of the program). It is interesting that for the Work First data the ordering of the two ETC-ES policies is reversed at the end of the treatment period compared to the remaining functionalities. The latter observation again underscores the difficulty in getting the length of the exploration period "right." This echoes our theoretical results showing that there is no way of constructing an ETC-based rule that would uniformly dominate the UCB-type policies.

**Supplementary Materials**

The supplementary materials contain all appendices to this article.

**Acknowledgments**

We are grateful to the editor, the associate editor, and two referees for their comments and suggestions that helped to substantially improve the article.

**References**

Abadie, A. (2002), “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models,” *Journal of the American Statistical Association*, 97, 284–292. [1311]

Abadie, A., Angrist, J., and Imbens, G. (2002), “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quanlity of Trainee Earnings,” *Econometrica*, 70, 91–117. [1311]

Agrawal, R. (1995), “Sample Mean Based Index Policies With $O(\log n)$ Regret for the Multi-Armed Bandit Problem,” *Advances in Applied Probability*, 27, 1054–1078. [1311]

Athey, S., and Wager, S. (2017), “Efficient Policy Learning,” arXiv no. 1702.02896. [1311]

Atkinson, A. B. (1970), “On the Measurement of Inequality,” *Economic Theory*, 2, 244–263. [1314]

Audibert, J., and Bubeck, S. (2009), “Minimax Policies for Adversarial and Stochastic Bandits,” in *Proceedings of the 22nd Conference on Learning Theory*, pp. 217–226. [1311,1317,1318]

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009), “Exploration—Exploitation Tradeoff Using Variance Estimates in Multi-Armed Bandits,” *Theoretical Computer Science*, 410, 1876–1902. [1317]

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002), “Finite-Time Analysis of the Multiarmed Bandit Problem,” *Machine Learning*, 47, 235–256. [1312,1317]

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995), “Gambling in a Rigged Casino: The Adversarial Multi-Armed Bandit Problem,” in *Proceedings of IEEE 36th Annual Foundations of Computer Science*, IEEE, pp. 322–331. [1311,1317]

Autor, D. H., and Houseman, S. N. (2010), “Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence From ‘Work First’,” *American Economic Journal: Applied Economics*, 2, 96–128. [1320]

Autor, D. H., Houseman, S. N., and Kerr, S. P. (2017), “The Effect of Work First Job Placements on the Distribution of Earnings: An Instrumental Variable Quantile Regression Approach,” *Journal of Labor Economics*, 35, 149–190. [1320]

Barrett, G. E., and Donald, S. G. (2009), “Statistical Inference With Generalized Gini Indices of Inequality, Poverty, and Welfare,” *Journal of Business & Economic Statistics*, 27, 1–17. [1311]

Bhattacharya, D., and Dupas, P. (2012), “Inferring Welfare Maximizing Treatment Assignment Under Budget Constraints,” *Journal of Econometrics*, 167, 168–196. [1311]

Bilias, Y. (2000), “Sequential Testing of Duration Data: The Case of the Pennsylvania ‘Reemployment Bonus’ Experiment,” *Journal of Applied Econometrics*, 15, 575–594. [1320]

Bubeck, S., and Cesa-Bianchi, N. (2012), “Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems,” *Foundations and Trends in Machine Learning*, 5, 1–122. [1311,1317]

Cassel, A., Manor, S., and Zeevi, A. (2018), “A General Approach to Multi-Armed Bandits Under Risk Criteria,” in *Proceedings of the 31st Conference on Learning Theory*, Proceedings of Machine Learning Research (Vol. 75), eds. S. Bubeck, V. Perchet, and P. Rigollet, pp. 1295–1306. [1312]

Chakravarty, S. R. (2009), *Inequality, Polarization and Poverty*, New York: Springer. [1314]

Chamberlain, G. (2000), *Econometrics and Decision Theory*, *Journal of Econometrics*, 95, 255–283. [1311]

Chernozhukov, V., Chetverikov, D., Demirov, M., Dulfo, E., Hansen, C., Newey, W., and Robins, J. (2018), “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, 21, C1–C68. [1320]

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013), “Inference on Counterfactual Distributions,” *Econometrica*, 81, 2205–2268. [1311]

Chernozhukov, V., and Hansen, C. (2005), “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73, 245–261. [1311]

Cowell, F. (2011), *Measuring Inequality*, Oxford: Oxford University Press. [1314]

Davidson, R., and Duclos, J.-Y. (2000), “Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality,” *Econometrica*, 68, 1435–1464. [1311]

Davidson, R., and Flachaire, E. (2007), “Asymptotic and Bootstrap Inference for Inequality and Poverty Measures,” *Journal of Econometrics*, 141, 141–166. [1311]

Degene, R., and Perch, V. (2016), “Anytime Optimal Algorithms in Stochastic Multi-Armed Bandits,” in *International Conference on Machine Learning*, pp. 1587–1595. [1312,1318]

Dehejia, R. H. (2005), “Program Evaluation as a Decision Problem,” *Journal of Econometrics*, 125, 141–173. [1311]

Doob, J. (1936), “Note on Probability,” *Annals of Mathematics*, 37, 363–367. [1317]

Doster, J., Greer, I., and Thorbecke, E. (1984), “A Class of Decomposable Poverty Measures,” *Econometrica*, 52, 761–766. [1314]

Garivier, A., Hadiji, H., Menard, P., and Stoltz, G. (2018), “KL-UCB-switch: Optimal Regret Bounds for Stochastic Bandits From Both a Distribution-Dependent and a Distribution-Free Viewpoints,” arXiv no. 1805.05071. [1318]

Garivier, A., Lattimore, T., and Kaufmann, E. (2016), “On Explore-Then-Commit Strategies,” in *Advances in Neural Information Processing Systems*, pp. 784–792. [1312]

Gastwirth, J. L. (1974), “Large Sample Theory of Some Measures of Income Inequality,” *Econometrica*, 42, 191–196. [1311]

Gittins, J. C. (1979), *Bandit Processes and Dynamic Allocation Indices*, *Journal of the Royal Statistical Society*, Series B, 41, 148–164. [1311]

Hardy, J. L., Nelson, R. A., Thomason, M. E., Sternberg, D. A., Katovich, K., Farzin, F., and Scanlon, M. (2015), “Enhancing Cognitive Abilities With Comprehensive Training: A Large, Online, Randomized, Active-Controlled Trial,” *PLoS ONE*, 10, e0134467. [1320]

Hirano, K., and Porter, J. R. (2009), “Asymptotcs for Statistical Treatment Rules,” *Econometrica*, 77, 1683–1701. [1311]

Hirano, K., and Porter, J. R. (2018), “Statistical Decision Rules in Econometrics,” Working Paper. [1311]

Imbens, G. W., and Wooldridge, J. M. (2009), “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86. [1317]

Kakwani, N. (1980), “On a Class of Poverty Measures,” *Econometrica*, 48, 347–446. [1314]

Kitagawa, T., and Teteno, A. (2018), “Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice,” *Econometrica*, 86, 591–616. [1311]
Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. (2016), "Batched Bandit Problems," The Annals of Statistics, 44, 660–681. [1313]
Robbins, H. (1952), "Some Aspects of the Sequential Design of Experiments," Bulletin of the American Mathematical Society, 58, 527–535. [1311]
Robins, J. M. (1997), "Causal Inference From Complex Longitudinal Data," in Latent Variable Modeling and Applications to Causality, ed. M. Berkane, New York: Springer, pp. 69–117. [1312]
Rosenbluth, G. (1951), "Note on Mr. Schütz's Measure of Income Inequality," American Economic Review, 41, 935–937. [1314]
Rostek, M. (2010), "Quantile Maximization in Decision Theory," Review of Economic Studies, 77, 339–371. [1311]
Rothe, C. (2010), "Nonparametric Estimation of Distributional Policy Effects," Journal of Econometrics, 155, 56–70. [1311]

(2012), "Partial Distributional Policy Effects," Econometrica, 80, 2269–2301. [1311]
Sani, A., Lazarić, A., and Munos, R. (2012), "Risk-Aversion in Multi-Armed Bandits," in Advances in Neural Information Processing Systems (Vol. 25), eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Curran Associates, Inc., pp. 3275–3283. [1311,1312,1317]
Schluter, C., and van Garderen, K. J. (2009), "Edgeworth Expansions and Normalizing Transforms for Inequality Measures," Journal of Econometrics, 150, 16–29. [1311]
Schutz, R. R. (1951), "On the Measurement of Income Inequality," American Economic Review, 41, 107–122. [1314]
Sen, A. (1974), "Informational Bases of Alternative Welfare Approaches: Aggregation and Income Distribution," Journal of Public Economics, 3, 387–403. [1314]

(1976), "Poverty: An Ordinal Approach to Measurement," Econometrica, 44, 219–231. [1314]
Serfling, R. J. (1984), "Generalized L-, M-, and R-Statistics," The Annals of Statistics, 12, 76–86. [1314]
Stoye, J. (2009), "Minimax Regret Treatment Choice With Finite Samples," Journal of Econometrics, 151, 70–81. [1311]

(2012), "Minimax Regret Treatment Choice With Covariates or With Limited Validity of Experiments," Journal of Econometrics, 166, 138–156. [1311]
Tetlenov, A. (2012), "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," Journal of Econometrics, 166, 157–165. [1311]
Thistle, P. D. (1990), "Large Sample Properties of Two Inequality Indices," Econometrica, 58, 725–728. [1311]
Thompson, W. R. (1933), "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," Biometrika, 25, 285–294. [1311]
Thurow, L. C. (1970), "Analyzing the American Income Distribution," American Economic Review, 60, 261–269. [1318]
Tran-Thanh, L., and Yu, J. Y. (2014), "Functional Bandits," arXiv no. 1405.2432. [1312]
Vakili, S., Boukouvalas, A., and Zhao, Q. (2018), "Decision Variance in Online Learning," arXiv no. 1807.09089. [1312,1313,1317]
Vakili, S., and Zhao, Q. (2016), "Risk-Averse Multi-Armed Bandit Problems Under Mean-Variance Measure," IEEE Journal of Selected Topics in Signal Processing, 10, 1093–1111. [1311,1312,1317]
Zimin, A., Ibsen-Jensen, R., and Chatterjee, K. (2014), "Generalized Risk-Aversion in Stochastic Multi-Armed Bandits," arXiv no. 1405.0833. [1312,1313,1317]

--- (2019), "Equality-Minded Treatment Choice," Journal of Business & Economic Statistics, 1–14.++ [1311]
Kock, A. B., Preinerstorfer, D., and Veliyev, B. (2020), "Functional Sequential Treatment Allocation With Covariates," arXiv no. 2001.10996. [1312]
Kock, A. B., and Thyrsgaard, M. (2017), "Optimal Sequential Treatment Allocation," arXiv no. 1705.09952. [1312,1313,1317]
Kolm, S.-C. (1976), "Unequal Inequalities. I," Journal of Economic Theory, 12, 416–442. [1314]
Lai,T.L.,andRobbins,H.(1985),"AsymptoticallyEfficientAdaptiveAllocation Rules," Advances in Applied Mathematics, 6, 4–22. [1311]
Lambert, P. J. (2001), The Distribution and Redistribution of Income, Manchester: Manchester University Press. [1314]
Lattimore, T., and Szepesvári, C. (2020), Bandit Algorithms, Cambridge: Cambridge University Press. [1311,1316]
Lavori, P. W., Dawson, R., and Rush, A. J. (2000), "Flexible Treatment Strategies in Chronic Disease: Clinical and Research Implications," Biological Psychiatry, 48, 605–614. [1312]
Maillard, O.-A. (2013), "Robust Risk-Averse Stochastic Multi-armed Bandits," in Algorithmic Learning Theory, eds. S. Jain, R. Munos, F. Stephan, and T. Zeugmann, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 218–233. [1311,1312,1313,1317]
Manski, C. F. (1988), "Ordinal Utility Models of Decision Making Under Uncertainty," Theory and Decision, 25, 79–104. [1311]

(2004), "Statistical Treatment Rules for Heterogeneous Populations," Econometrica, 72, 1211–1246. [1311,1316,1319]
(2019a), "Remarks on Statistical Inference for Statistical Decisions," Tech. Rep., Centre for Microdata Methods and Practice, Institute for Fiscal Studies. [1311]

(2019b), "Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," American Statistician, 73, 296–304. [1311]
Manski, C. F., and Tetenov, A. (2016), "Sufficient Trial Size to Inform Clinical Practice," Proceedings of the National Academy of Sciences of the United States of America, 113, 10518–10523. [1311,1314,1319]
Massart, P. (1990), "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality," Annals of Probability, 18, 1269–1283. [1317]
McDonald, J. B. (1984), "Some Generalized Functions for the Size Distribution of Income," Econometrica, 52, 647–663. [1318]
McDonald, J. B., and Ransom, M. (2008), "The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation of Related Measures of Inequality," in Modeling Income Distributions and Lorenz Curves, ed. D. Chotikapanich, New York: Springer, pp. 147–166. [1318]
Mehran, F. (1976), "Linear Measures of Income Inequality," Econometrica, 44, 805–809. [1314]
Mills, J. A., and Zandvakili, S. (1997), "Statistical Inference via Bootstrapping for Measures of Inequality," Journal of Applied Econometrics, 12, 133–150. [1311]
Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," Journal of the Royal Statistical Society, Series B, 65, 331–355. [1312]

(2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," Statistics in Medicine, 24, 1455–1481. [1312]
Murphy, S. A., van der Laan, M. J., and Robins, J. M. (2001), "Marginal Mean Models for Dynamic Regimes," Journal of the American Statistical Association, 96, 1410–1423. [1312]