Batched Bandit Problems*

Vianney Perchet
Univeristé
Paris Diderot
vianney.perchet@normalesup.org
sites.google.com/site/vianneyperchet/

Philippe Rigollet
Massachusetts
Institute of Technology
rigollet@math.mit.edu
www-math.mit.edu/∼rigollet/

Sylvain Chassang
Princeton University

chassang@prinecton.edu
princeton.edu/∼chassang/

Erik Snowberg
California Institute
of Technology and NBER
snowberg@caltech.edu
hss.caltech.edu/∼snowberg/

Abstract

Motivated by practical applications, chiefly clinical trials, we study the regret achievable for stochastic bandits under the constraint that the employed policy must split trials into a small number of batches. Our results show that a very small number of batches gives close to minimax optimal regret bounds. As a byproduct, we derive optimal policies with low switching cost for stochastic bandits.

AMS 2000 subject classifications: Primary—62L05; Secondary—62C20
Keywords: Multi-armed bandit problems, regret bounds, batches, multi-phase allocation, grouped clinical trials, sample size determination, switching cost

*Rigollet acknowledges the support of NSF grants DMS-1317308, CAREER-DMS-1053987 and the Meimaris family. Chassang and Snowberg acknowledge the support of NSF SES-1156154.

1 Introduction

In his seminal paper [Tho33], Thompson introduced the multi-armed bandit problem. The main motivation behind Thompson's work came from clinical trials. Bandit problems capture a fundamental exploration-exploitation dilemma that has made the framework popular for studying problems not only in clinical trials but also in economics, finance, chemical engineering, scheduling, marketing and, more recently, online advertising. This last application has been the driving force behind a recent surge of interest in many variations of bandit problems over the past decade. However, this recent avenue of research neglects important issues in clinical trials, such as the design of policies that use a small number of batches.

The basic framework of a bandit problem can be expressed as follows [Tho33, Rob52]: given two populations of patients (or *arms*), corresponding to different medical treatments, at each time $t = 1, \ldots, T$, sample from only one of the populations and receive a random reward dictated by the efficacy of the treatment. The objective is to devise a policy that maximizes the expected cumulative reward over T rounds. Thompson compared this problem to a gambler facing a slot machine, hence the terminology of "bandit" and "arms". In this problem one faces a clear tradeoff between discovering which treatment is the most effective—or *exploration*—and administering the best treatment to as many patients as possible—or *exploitation*.

In clinical trials and some other domains, it is impractical to measure rewards (or efficacy) for each patient before deciding which treatment to administer next. Instead, clinical trials are performed in a small number of sequential batches. These batches may be formalized, as in the different phases required for approval of a new drug by the U.S. Food and Drug Administration (FDA), or, more generally, they are informally expressed as a pilot, a full trial, and then diffusion to the full population that may benefit. The second step may be skipped if the first trial is successful enough. In this three-stage approach, the first, and usually second, phases focus on exploration while the third one focuses on exploitation. This

is in stark contrast with the basic problem described above that effectively consists of T batches containing a single patient. This observation leads to two important questions: How does one reunite the two frameworks? Do many smaller batches significantly improve upon the basic three-stage approach? Answering these questions has important implications not only in clinical trials but also in marketing [BM07, SBF13] and simulations [CG09].

In this paper, we focus on the two-armed case where one arm can be thought of as treatment and the other as control. This choice is motivated by standard clinical trials, and by the fact that the central ideas and intuitions are all captured by this concise framework. Extensions of this framework to K-armed bandit problems are mostly technical, (see for instance [PR13]).

NOTATION. For any two integers $n_1 < n_2$, we define the following sets of consecutive integers: $[n_1 : n_2] = \{n_1, \dots, n_2\}$ and $(n_1 : n_2] = \{n_1+1, \dots, n_2\}$. For simplicity, $[n] = [1 : n]$ for any positive integer n. For any positive number x, let $\lfloor x \rfloor$ denote the largest integer n such that $n \leq x$ and $\lfloor x \rfloor_2$ denotes the largest *even* integer m such that $m \leq x$. Additionally, if a and b are two elements of a partially ordered set we write $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$. It will be useful to view the set of all closed intervals of $\mathbb{R}$ as a partially ordered set where $\mathcal{I} \prec \mathcal{J}$ if $x < y$ for all $x \in \mathcal{I}, y \in \mathcal{J}$ (interval order).

For two sequences $(u_T)_T, (v_T)_T$, we write $u_T = \mathcal{O}(v_T)$ or $u_T \lesssim v_T$ if there exists a constant $C > 0$ such that $u_T \leq C v_T$ for any T. Moreover, we write $u_T = \Theta(v_T)$ if $u_T = \mathcal{O}(v_T)$ and $v_T = \mathcal{O}(u_T)$.

Finally, the following functions are employed: $\overline{\log}(x) = 1 \vee (\log x)$, and $\mathbb{1}(\cdot)$ denotes the indicator function.

2 Description of the Problem

2.1 Setup

We employ a two-armed bandit problem with horizon $T \geq 2$. At each time $t \in [T]$, the decision maker chooses an arm $i \in [2]$ and observes a reward that comes from a sequence of i.i.d. draws $Y_1^{(i)}, Y_2^{(i)}, \ldots$ from some unknown distribution $\nu^{(i)}$ with expected value $\mu^{(i)}$. We assume that the distributions $\nu^{(i)}$ are standardized sub-Gaussian that is $\int e^{\lambda(x-\mu^{(i)})}\nu_i(dx) \leq e^{\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. Note that these include Gaussian distributions with variance at most 1 and distributions supported on an interval of length at most 2. This definition extends to other variance parameters σ^2, which can be handled by rescaling the observations by σ.

For any integer $M \in [2 : T]$, let $\mathcal{T} = \{t_1, \ldots, t_M\}$ be an ordered sequence, or *grid*, of integers such that $1 < t_1 < \ldots < t_M = T$. It defines a partition $\mathcal{S} = \{S_1, \ldots S_M\}$ of $[T]$ where $S_1 = [1 : t_1]$ and $S_k = (t_{k-1} : t_k]$ for $k \in [2 : M]$. The set S_k is called *k-th batch*. An *M-batch policy* is a couple $(\mathcal{T}, \pi)$ where $\mathcal{T} = \{t_1, \ldots t_M\}$ is a grid and $\pi = \{\pi_t\,, t = 1, \ldots, T\}$ is a sequence of random variables $\pi_t \in [2]$ that indicates which arm to pull at each time $t = 1, \ldots, T$, under the constraint that π_t depends only on observations from batches strictly anterior to the current batch. Formally, for each $t \in [T]$, let $J(t) \in [M]$ be the index of the *current batch* $S_{J(t)}$. Then, for $t \in S_{J(t)}$, π_t can only depend on observations $\{Y_s^{(\pi_s)} \,:\, s \in S_1 \cup \ldots \cup S_{J(t)-1}\} = \{Y_s^{(\pi_s)} \,:\, s \leq t_{J(t)-1}\}$.

Denote by $\star \in [2]$ the uniquely optimal arm defined by $\mu^{(\star)} = \max_{i\in[2]} \mu^{(i)}$. Moreover, let $\dagger \in [2]$ denote the suboptimal arm. We denote by $\Delta := \mu^{(\star)} - \mu^{(\dagger)} > 0$ the gap between the optimal expected reward and the sub-optimal expected reward.

The performance of a policy π is measured by its (cumulative) *regret* at time T defined by

$$R_T = R_T(\pi) = T\mu^{(\star)} - \sum_{t=1}^{T} \mathbb{E}\, \mu^{(\pi_t)}\,.$$

In particular, if we denote the number of times arm i was pulled (strictly) before time $t \geq 2$

by $T_i(t) = \sum_{s=1}^{t-1} \mathbb{1}(\pi_s = i)\,, i \in [2]$, then regret can be rewritten as

$$R_T = \Delta \mathbb{E} T_\dagger(T+1)\,.$$

2.2 Previous Results

As mentioned in the introduction, the bandit problem has been extensively studied in the case where $M = T$, that is, when the decision maker can use all past data at each time $t \in [T]$. Bounds on the cumulative regret R_T for stochastic multi-armed bandits come in two flavors: *minimax* or *adaptive*. Minimax bounds hold regardless of the value of Δ. The first results of this kind are attributed to Vogel [Vog60a, Vog60b] (see also [FZ70, Bat81]) who proves that $R_T = \Theta(T^{1/2})$ in the two-armed case.

Adaptive policies exhibit regret bounds that may be of order much smaller than $\sqrt{T}$ when the problem is easy, that is when Δ is large. Such bounds were proved in the seminal paper of Lai and Robbins [LR85] in an asymptotic framework (see also [CGM+13]). While leading to tight constants, this framework washes out the correct dependency in Δ of the logarithmic terms. In fact, recent research [ACBF02, AB10, AO10, PR13] has revealed that $R_T = \Theta\big(\Delta T \wedge \overline{\log}(T\Delta^2)/\Delta\big)$.

Note that $\Delta T \wedge \overline{\log}(T\Delta^2)/\Delta \leq \sqrt{T}$ for any $\Delta > 0$. This suggests that adaptive bounds are more desirable than minimax bounds. Intuitively such bounds are achievable in the classical setup because at each round t, more information about the Δ can be collected. As we shall see, this is not always possible in the batched setup and we have to resort to a compromise.

While optimal regret bounds are well understood for standard multi-armed bandit problems when $M = T$, a systematic analysis of the batched case does not exist. It is worth pointing out that UCB2 [ACBF02] and IMPROVED-UCB [AO10] are actually M-batch policies where $M = \Theta(\log T)$. Since these policies already achieve optimal adaptive bounds,

it implies that employing a batched policy can only become a constraint when the number of batches M is small, that is when $M \ll \log T$. Similar observations can be made in the minimax framework. Indeed, recent M-batch policies [CBDS13, CBGM13], where $M = \Theta(\log \log T)$, lead to a regret bounded $\mathcal{O}\left(\sqrt{T \log \log \log(T)}\right)$ which is nearly optimal, up to the logarithmic terms.

The sub-logarithmic range, where $M \ll \log T$ is precisely the interesting one in applications such as clinical trials, where M should be considered constant, and a constraint. In particular, we wish to bound the regret for small values of M, such as 2, 3, or 4. As we shall see below, the value of M has a significant impact on the cumulative regret.

2.3 Literature

This paper connects to two lines of work: batched sequential estimation and multistage clinical trials. Batched sequential estimation indirectly starts with a paper of George Danzig [Dan40] that establishes nonexistence of statistical tests with a certain property (and comes with a famous anecdote [CJW07]). To overcome this limitation, Stein [Ste45] developed a two-stage test that satisfies this property but requires a random number of observations. A similar question was studied by Ghurye and Robbins [GR54] where the question of batch size was explicitly addressed. That same year, answering a question suggested by Robbins, Somerville [Som54] used a minimax criterion to find the optimal size of batches in a two-batch setup. Specifically, Somerville [Som54] and Maurice [Mau57] studied the two-batch bandit problem in a minimax framework under a Gaussian assumption. They prove that an "explore-then-commit" type policy has regret of order $T^{2/3}$ for any value of the gap Δ. In this paper, we recover this rate in the case of two batches (see subsection 4.3) and further extend the results to more batches and more general distributions.

Soon after, and inspired by this line of work, Colton [Col63, Col65] introduced a Bayesian perspective on the problem and initiated a long line of work (see [HS02] for a fairly recent

overview of this literature). Apart from isolated works [Che96, HS02], most of this work focuses on the case of two and sometimes three batches. The general consensus is that the size of the first batch should be of order $\sqrt{T}$. As we shall see, our results below recover this order up to logarithmic term if the gap between expected rewards is of order one (see subsection 4.2). This is not surprising as the priors put on the expected rewards correspond essentially to a constant size gap.

Finally, it is worth mentioning that batched procedures have a long history in clinical trials. The literature on this topic is vast; [JT00] and [BLS13] provide detailed monographs on the frequentist treatment of this subject. The largest body of this literature focuses on batches that are of the same size, or of random size, with the latter case providing robustness. One notable exception is [Bar07], which provides an ad-hoc objective to optimize the batch size but also recovers the suboptimal $\sqrt{T}$ in the case of two batches. However, it should be noted that this literature focuses on inference questions rather than cumulative regret.

2.4 Outline

The rest of the paper is organized as follows. Section 3 introduces a general class of M-batch policies that we name *Explore-then-commit (*ETC*) policies*. The performance of generic ETC policies are detailed in Proposition 1 in Section 3.3. In Section 4, we study several instantiations of this generic policy and provide regret bounds with explicit, and often drastic, dependency on the number M of batches. Indeed, in subsection 4.3, we describe a policy whose regret decreases doubly exponentially fast with the number of batches.

Two of the instantiations provide adaptive and minimax types of bounds respectively. Specifically, we describe two M-batch policies, π^1 and π^2 that enjoy the following bounds on

the regret:

$$R_T(\pi^1) \lesssim \left(\frac{T}{\log(T)}\right)^{\frac{1}{M}} \frac{\overline{\log}(T\Delta^2)}{\Delta}$$
$$R_T(\pi^2) \lesssim T^{\frac{1}{2-2^{1-M}}} \log^{\alpha_M}\left(T^{\frac{1}{2^M-1}}\right), \qquad \alpha_M \in [0, 1/4)\,.$$

Note that the bound for π^1 corresponds to the optimal adaptive rate $\overline{\log}(T\Delta^2)/\Delta$ as soon as $M = \Theta(\log(T/\log(T)))$ and the bound for π^2 corresponds to the optimal minimax rate $\sqrt{T}$ as soon as $M = \Theta(\log\log T)$. As a byproduct of our results, we obtain that the adaptive optimal bounds can be obtained with a policy that switches between arms less than $\Theta(\log(T/\log(T)))$ times, while the optimal minimax bounds only require $\Theta(\log\log T)$ switches to be attained. Indeed, ETC policies can be adapted to switch at most once in each batch.

Section 5 then examines the lower bounds on regret of any M-batch policy, and shows that the policies identified are optimal, up to logarithmic terms, within the class of M-batch policies. Finally, in Section 6 we compare policies through simulations using both standard distributions and real data from a clinical trial, and show that those we identify perform well even with a very small number of batches.

3 Explore-then-commit Policies

In this section, we describe a simple principle that can be used to build policies: *explore-then-commit* (ETC). At a high level, this principle consists in pulling each arm the same number of times in each non-terminal batch, and checking after each batch whether an arm dominates the other, according to some statistical test. If this occurs, then only the arm believed to be optimal is pulled until T. Otherwise, the same procedure is repeated in the next batch. If, at the beginning of the terminal batch, no arm has been declared optimal, then the policy commits to the arm with the largest average past reward. This "go for broke" step does not account for deviations of empirical averages from the population means, but

is dictated by the use of cumulative regret as a measure of performance. Indeed, in the last batch exploration is pointless as the information it produces can never be used.

Clearly, any policy built using this principle is completely characterized by two ingredients: the testing criterion and the sizes of the batches.

3.1 Statistical Test

We begin by describing the statistical test employed in non-terminal batches. Denote by

$$\widehat{\mu}_s^{(i)} = \frac{1}{s}\sum_{\ell=1}^{s} Y_\ell^{(i)}$$

the empirical mean after $s \geq 1$ pulls of arm i. This estimator allows for the construction of a collection of upper and lower confidence bounds for $\mu^{(i)}$. They take the form

$$\widehat{\mu}_s^{(i)} + \mathsf{B}_s^{(i)}\,, \quad \text{and} \quad \widehat{\mu}_s^{(i)} - \mathsf{B}_s^{(i)}\,,$$

where $\mathsf{B}_s^{(i)} = 2\sqrt{2\log\big(T/s\big)/s}$, with the convention that $\mathsf{B}_0^{(i)} = \infty$. Indeed, it follows from Lemma B.1 that for any $\tau \in [T]$,

$$\mathbb{P}\Big\{\exists s \leq \tau \,:\, \mu^{(i)} > \widehat{\mu}_s^{(i)} + \mathsf{B}_s^{(i)}\Big\} \vee \mathbb{P}\Big\{\exists s \leq \tau \,:\, \mu^{(i)} < \widehat{\mu}_s^{(i)} - \mathsf{B}_s^{(i)}\Big\} \leq \frac{4\tau}{T}\,. \tag{3.1}$$

These bounds enable us to design the following family of tests $\{\varphi_t\}_{t\in[T]}$ with values in $\{1, 2, \perp\}$ where $\perp$ indicates that the test was inconclusive in determining the optimal arm. Although this test will be implemented only at times $t \in [T]$ at which each arm has been pulled exactly $s = t/2$ times, for completeness, we define the test at all times. For $t \geq 1$

define

$$\varphi_t = \begin{cases} i \in \{1,2\} & \text{if } T_1(t) = T_2(t) = t/2\,, \text{and } \widehat{\mu}^{(i)}_{t/2} - \mathsf{B}^{(i)}_{t/2} > \widehat{\mu}^{(j)}_{t/2} + \mathsf{B}^{(j)}_{t/2}, j \neq i, \\ \perp & \text{otherwise.} \end{cases}$$

The errors of such tests are controlled as follows.

Lemma 3.1. *Let $\mathcal{S} \subset [T]$ be a deterministic subset of even times such that $T_1(t) = T_2(t) = t/2$, for $t \in \mathcal{S}$. Partition $\mathcal{S}$ into $\mathcal{S}_- \cup \mathcal{S}_+$, $\mathcal{S}_- \prec \mathcal{S}_+$, where*

$$\mathcal{S}_- = \left\{ t \in \mathcal{S} \,:\, \Delta < 16\sqrt{\frac{\log(2T/t)}{t}} \right\}, \quad \mathcal{S}_+ = \left\{ t \in \mathcal{S} \,:\, \Delta \geq 16\sqrt{\frac{\log(2T/t)}{t}} \right\}.$$

Moreover, let $\bar{t}$ denote the smallest element of $\mathcal{S}_+$. Then, the following holds

$$(i)\ \mathbb{P}(\varphi_{\bar{t}} \neq \star) \leq \frac{4\bar{t}}{T}\,, \quad (ii)\ \mathbb{P}(\exists\, t \in \mathcal{S}_- \,:\, \varphi_t = \dagger) \leq \frac{4\bar{t}}{T}\,.$$

Proof. Assume without loss of generality that $\star = 1$.

(i) By definition

$$\{\varphi_{\bar{t}} \neq 1\} = \left\{ \widehat{\mu}^{(1)}_{\bar{t}/2} - \mathsf{B}^{(1)}_{\bar{t}/2} \leq \widehat{\mu}^{(2)}_{\bar{t}/2} + \mathsf{B}^{(2)}_{\bar{t}/2} \right\} \subset \{E^1_{\bar{t}} \cup E^2_{\bar{t}} \cup E^3_{\bar{t}}\},$$

where $E^1_t = \left\{\mu^{(1)} \geq \widehat{\mu}^{(1)}_{t/2} + \mathsf{B}^{(1)}_{t/2}\right\}$, $E^2_t = \left\{\mu^{(2)} \leq \widehat{\mu}^{(2)}_{t/2} - \mathsf{B}^{(2)}_{t/2}\right\}$ and $E^3_t = \left\{\mu^{(1)} - \mu^{(2)} < 2\mathsf{B}^{(1)}_{t/2} + 2\mathsf{B}^{(2)}_{t/2}\right\}$. It follows from (3.1) with $\tau = \bar{t}/2$ that, $\mathbb{P}(E^1_{\bar{t}}) \vee \mathbb{P}(E^2_{\bar{t}}) \leq 2\bar{t}/T$.

Next, for any $t \in \mathcal{S}_+$, in particular for $t = \bar{t}$, it holds

$$E^3_t \subset \left\{ \mu^{(1)} - \mu^{(2)} < 16\sqrt{\frac{\log(2T/t)}{t}} \right\} = \emptyset\,.$$

(ii) We now focus on the case $t \in \mathcal{S}_-$, where $\Delta < 16\sqrt{\log(2T/t)/t}$. In this case,

$$\bigcup_{t\in\mathcal{S}_-} \{\varphi_t = 2\} = \bigcup_{t\in\mathcal{S}_-} \{\widehat{\mu}^{(2)}_{t/2} - \mathsf{B}^{(2)}_{t/2} > \widehat{\mu}^{(1)}_{t/2} + \mathsf{B}^{(1)}_{t/2}\} \subset \bigcup_{t\in\mathcal{S}_-} \{E^1_t \cup E^2_t \cup F^3_t\}\,,$$

where, E^1_t, E^2_t are defined above and $F^3_t = \{\mu^{(1)} - \mu^{(2)} < 0\} = \emptyset$ since $\star = 1$. It follows from (3.1) with $\tau = \bar{t}$ that

$$\mathbb{P}\Big(\bigcup_{t\in\mathcal{S}_-} E^1_t\Big) \vee \mathbb{P}\Big(\bigcup_{t\in\mathcal{S}_-} E^2_t\Big) \le \frac{2\bar{t}}{T}\,.$$

□

3.2 Go for Broke

In the last batch, the ETC principle will "go for broke" by selecting the arm i with the largest average. Formally, at time t, let $\psi_t = i$ iff $\widehat{\mu}^{(i)}_{T_i(t)} \ge \widehat{\mu}^{(j)}_{T_j(t)}$, with ties broken arbitrarily. While this criterion may select the suboptimal arm with higher probability than the statistical test described in the previous subsection, it also increases the probability of selecting the correct arm by eliminating inconclusive results. This statement is formalized in the following lemma, whose proof follows immediately from Lemma B.1.

Lemma 3.2. *Fix an even time $t \in [T]$, and assume that both arms have been pulled $t/2$ times each (i.e., $T_i(t) = t/2$, for $i = 1, 2$). Going for broke leads to a probability of error*

$$\mathbb{P}\left(\psi_t \neq \star\right) \le \exp(-t\Delta^2/16)$$

3.3 Explore-then-commit Policy

When operating in the batched setup, we recall that an extra constraint is that past observations can only be inspected at a specific set of times $\mathcal{T} = \{t_1, \ldots, t_{M-1}\} \subset [T]$ that we call

a *grid*.

The generic ETC policy uses a deterministic grid $\mathcal{T}$ that is fixed beforehand, and is described more formally in Figure 1. Informally, at each decision time $t_1, \ldots, t_{M-2}$, the policy uses the statistical test criterion to determine whether an arm is better. If the test indicates this is so, the better arm is pulled until the horizon T. If no arm is declared best, then both arms are pull the same number of times in the next batch.

We denote by $\varepsilon_t \in \{1, 2\}$ the arm pulled at time $t \in [T]$, and employ an external source of randomness to generate the variables ε_t. This has no effect on the policy, and could easily be replaced by any other mechanism that pulls each arm an equal number of times, such as a mechanism that pulls one arm for the first half of the batch and the other arm for the rest. This latter mechanism may be particularly attractive if switching costs are a concern. However, the randomized version is aligned with practice in clinical trials where randomized experiments are the rule. Typically, if N is an even number, let $(\varepsilon_1, \ldots, \varepsilon_N)$ be uniformly distributed over the subset $\mathcal{V}_N = \big\{v \in \{1, 2\}^N : \sum_i \mathbb{1}(v_i = 1) = N/2\big\}$.[1]

For the terminal batch S_M, if no arm was determined to be optimal in prior batches, the ETC policy will go for broke by selecting the arm i such that $\widehat{\mu}^{(i)}_{T_i(t_{M-1})} \geq \widehat{\mu}^{(j)}_{T_j(t_{M-1})}$, with ties broken arbitrarily.

To describe the regret incurred by a generic ETC policy, we introduce extra notation. For any $\Delta \in (0, 1)$, let $\tau(\Delta) = T \wedge \vartheta(\Delta)$ where $\vartheta(\Delta)$ is the smallest integer such that

$$\Delta \geq 16\sqrt{\frac{\log[2T/\vartheta(\Delta)]}{\vartheta(\Delta)}}\,.$$

Notice that it implies that $\tau(\Delta) \geq 2$ and

$$\tau(\Delta) \leq \frac{256}{\Delta^2}\overline{\log}\left(\frac{T\Delta^2}{128}\right)\,. \tag{3.2}$$

[1] One could consider odd numbers for the deadlines t_i but this creates rounding problems that only add complexity without insight. In the general case, we define $\mathcal{V}_N = \big\{v \in \{1, 2\}^N : \big|\sum_i \mathbb{1}(v_i = 1) - \sum_i \mathbb{1}(v_i = 2)\big| \leq 1\big\}$.

Input:

- Horizon: T.
- Number of batches: $M \in [2:T]$.
- Grid: $\mathcal{T} = \{t_1, \ldots, t_{M-1}\} \subset [T]$, $t_0 = 0$, $t_M = T$, $|S_m| = t_m - t_{m-1}$ is even for $m \in [M-1]$.

Initialization:

- Let $\varepsilon^{[m]} = (\varepsilon_1^{[m]}, \ldots, \varepsilon_{|S_m|}^{[m]})$ be uniformly distributed over[a] $\mathcal{V}_{|S_m|}$, for $m \in [M]$.
- The index ℓ of the batch in which a best arm was identified is initialized to $\ell = \circ$.

Policy:

1. For $t \in [1:t_1]$, choose $\pi_t = \varepsilon_t^{[1]}$
2. For $m \in [2:M-1]$,
   (a) If $\ell \neq \circ$, then $\pi_t = \varphi_{t_\ell}$ for $t \in (t_{m-1}:t_m]$.
   (b) Else, compute $\varphi_{t_{m-1}}$
      i. If $\varphi_{t_{m-1}} = \perp$, select an arm at random, that is, $\pi_t = \varepsilon_t^{[m]}$ for $t \in (t_{m-1}:t_m]$.
      ii. Else, $\ell = m-1$ and $\pi_t = \varphi_{t_{m-1}}$ for $t \in (t_{m-1}:t_m]$.
3. For $t \in (t_{M-1}, T]$,
   (a) If $\ell \neq \circ$, $\pi_t = \varphi_{t_\ell}$.
   (b) Otherwise, go for broke, i.e., $\pi_t = \psi_{t_{M-1}}$.

[a] In the case where $|S_m|$ is not an even number, we use the general definition of footnote 1 for $\mathcal{V}_{|S_m|}$.

Figure 1: Generic explore then commit policy with grid $\mathcal{T}$.

The time $\tau(\Delta)$ is, up to a multiplicative constant, the theoretical time at which the optimal arm can be declared best with large enough probability. As Δ is unknown, the grid will not usually contain this value, thus the relevant quantity is the first time posterior to $\tau(\Delta)$ in a

grid. Specifically, given a grid $\mathcal{T} = \{t_1, \dots, t_{M-1}\} \subset [T]$, define

$$m(\Delta, \mathcal{T}) = \begin{cases} \min\{m \in \{1, \dots, M-1\} : t_m \geq \tau(\Delta)\} & \text{if } \tau(\Delta) \leq t_{M-1} \\ M-1 & \text{otherwise} \end{cases} \tag{3.3}$$

The first proposition gives an upper bound for the regret incurred by a generic ETC policy run with a given set of times $\mathcal{T} = \{t_1, \dots, t_{M-1}\}$.

Proposition 1. *Let the time horizon $T \in \mathbb{N}$, the number of batches $M \in [2, T]$ and the grid $\mathcal{T} = \{t_1, \dots, t_{M-1}\} \subset [T]$ be given, $t_0 = 0$. For any $\Delta \in [0, 1]$, the generic* ETC *policy described in Figure 1 incurs a regret bounded as*

$$R_T(\Delta, \mathcal{T}) \leq 9\Delta t_{m(\Delta,\mathcal{T})} + T\Delta e^{-\frac{t_{M-1}\Delta^2}{16}} \mathbb{1}(m(\Delta, \mathcal{T}) = M-1)\,. \tag{3.4}$$

Proof. Throughout the proof, we denote $\bar{m} = m(\Delta, \mathcal{T})$ for simplicity. We first treat the case where $t_{\bar{m}} < M-1$. Note that $t_{\bar{m}}$ denotes the theoretical time on the grid at which the statistical test will declare $\star$ to be the best arm with high probability.

Define the following events:

$$A_m = \bigcap_{n=1}^{m} \{\varphi_{t_n} = \perp\}\,, \quad B_m = \{\varphi_{t_m} = \dagger\}\,, \quad C_m = \{\varphi_{t_m} \neq \star\}\,.$$

Note that on A_m, both arms have been pulled an equal number of times at time t_m. The test may also declare $\star$ to be the best arm at times prior to $t_{\bar{m}}$ without incurring extra regret. Regret can be incurred in one of the following three manners:

(i) by exploring before time $\bar{t}_m$.

(ii) by choosing arm $\dagger$ before time $t_{\bar{m}}$: this happens on event B_m

(iii) by not committing to the optimal arm $\star$ at the optimal time $t_{\bar{m}}$: this happens on event $C_{\bar{m}}$.

Error (i) is unavoidable and may occur with probability close to one. It corresponds to the exploration part of the policy and leads to an additional term $t_{\bar{m}}\Delta/2$ in the regret. Committing an error of the type (ii) or (iii) can lead to a regret of at most $T\Delta$ so we need to ensure that they occur with low probability. Therefore, the regret incurred by the policy is bounded as

$$R_T(\Delta, \mathcal{T}) \leq \frac{t_{\bar{m}}\Delta}{2} + T\Delta\mathbb{E}\Big[\mathbb{1}(\bigcup_{m=1}^{\bar{m}-1} A_{m-1} \cap B_m) + \mathbb{1}(B_{\bar{m}-1} \cap C_{\bar{m}})\Big]\,, \tag{3.5}$$

with the convention that A_0 is the whole probability space.

Next, observe that $\bar{m}$ is chosen such that

$$16\sqrt{\frac{\log(2T/t_{\bar{m}})}{t_{\bar{m}}}} \leq \Delta < 16\sqrt{\frac{\log(2T/t_{\bar{m}-1})}{t_{\bar{m}-1}}}\,.$$

In particular, $t_{\bar{m}}$ plays the role of $\bar{t}$ in Lemma 3.1. This yields using part (i) of Lemma 3.1 that

$$\mathbb{P}(B_{\bar{m}-1} \cap C_{\bar{m}}) \leq \frac{4t_{\bar{m}}}{T}$$

Moreover, using part (ii) of the same lemma, we get

$$\mathbb{P}\Big(\bigcup_{m=1}^{\bar{m}-1} A_{m-1} \cap B_m\Big) \leq \frac{4t_{\bar{m}}}{T}\,.$$

Together with (3.5) we get that the ETC policy has regret bounded by $R_T(\Delta, \mathcal{T}) \leq 9\Delta t_{\bar{m}}$.

We now consider the case where $t_{m(\Delta,\mathcal{T})} = M-1$. Lemma 3.2 yields that the go for broke test $\psi_{t_{M-1}}$ errs with probability at most $\exp(-t_{M-1}\Delta^2/16)$. The same argument as before gives that the expected regret is bounded as

$$R_T(\Delta, \mathcal{T}) \leq 9\Delta t_{m(\Delta,\mathcal{T})} + T\Delta e^{-\frac{t_{M-1}\Delta^2}{16}}\,.$$

□

Proposition 1 serves as a guide to choosing a grid by showing how that choice reduces to an optimal discretization problem. Indeed, the grid $\mathcal{T}$ should be chosen in such a way that $t_{m(\Delta,\mathcal{T})}$ is not much larger than the theoretically optimal $\tau(\Delta)$.

4 Functionals, Grids and Bounds

The regret bound of Proposition 1 critically depends on the choice of the grid $\mathcal{T} = \{t_1, \dots, t_{M-1}\} \subset [T]$. Ideally, we would like to optimize the right-hand side of (3.4) with respect to the t_ms. For a fixed Δ, this problem is easy and it is enough to choose $M = 2$, $t_1 \simeq \tau(\Delta)$ to obtain optimal regret bounds of the order $R^*(\Delta) = \log(T\Delta^2)/\Delta$. For unknown Δ, the problem is not well defined: as observed by [Col63, Col65], it consists in optimizing a function $R(\Delta, \mathcal{T})$ for all Δ and there is no choice that is uniformly better than others. To overcome this limitation, we minimize pre-specified real-valued functional of $R(\cdot, \mathcal{T})$. Examples of such functionals are of the form:

$$
\begin{aligned}
F_{\mathsf{xs}}\big[R_T(\cdot,\mathcal{T})\big] &= \sup_{\Delta}\{R_T(\Delta,\mathcal{T}) - CR^*(\Delta)\}, C>0 && \text{Excess regret}\\
F_{\mathsf{cr}}\big[R_T(\cdot,\mathcal{T})\big] &= \sup_{\Delta}\frac{R_T(\Delta,\mathcal{T})}{R^*(\Delta)} && \text{Competitive ratio}\\
F_{\mathsf{mx}}\big[R_T(\cdot,\mathcal{T})\big] &= \sup_{\Delta} R_T(\Delta,\mathcal{T}) && \text{Maximum}\\
F_{\mathsf{by}}\big[R_T(\cdot,\mathcal{T})\big] &= \int R_T(\Delta,\mathcal{T})\mathrm{d}\pi(\Delta) && \text{Bayesian}
\end{aligned}
$$

where in F_{by}, π is a given prior distribution on Δ. Note that the prior is on Δ here, rather than directly on the expected rewards as in the traditional Bayesian bandit literature [BF85].

One can also consider combination of the Bayesian criterion with other criteria. For example:

$$\int \frac{R_T(\Delta, \mathcal{T})}{R^*(\Delta)} \mathrm{d}\pi(\Delta)\,.$$

As Bayesian criteria are beyond the scope of this paper we focus on the first three criteria.

Optimizing different functionals leads to different grids. In the rest of this section, we define and investigate the properties of optimal grids associated with each of the three criteria.

4.1 Excess Regret and the Arithmetic Grid

We begin with a simple grid that consists in choosing a uniform discretization of $[T]$. Such a grid is particularly prominent in the group testing literature [JT00]. As we will see, even in a favorable setup, the regret bounds yielded by this grid are not good. Assume for simplicity that $T = 2KM$ for some positive integer K, so that the grid is defined by $t_m = mT/M$.

In this case, the right-hand side of (3.4) is bounded *below* by $\Delta t_1 = \Delta T/M$. For small M, this lower bound is linear in $T\Delta$, which is a trivial bound on the regret. To obtain a valid upper bound for the excess regret, note that

$$t_{m(\Delta,\mathcal{T})} \le \tau(\Delta) + \frac{T}{M} \le \frac{256}{\Delta^2} \overline{\log}\left(\frac{T\Delta^2}{128}\right) + \frac{T}{M}\,.$$

Moreover, if $m(\Delta, \mathcal{T}) = M - 1$ then Δ is of the order of $\sqrt{1/T}$ thus $T\Delta \lesssim 1/\Delta$. Together with (3.4), it yields the following theorem.

Theorem 1. *The* ETC *policy implemented with the arithmetic grid defined above ensures that, for any* $\Delta \in [0, 1]$,

$$R_T(\Delta, \mathcal{T}) \lesssim \left(\frac{1}{\Delta} \overline{\log}(T\Delta^2) + \frac{T\Delta}{M}\right) \wedge T\Delta\,.$$

In particular, if $M = T$, we recover the optimal rate. However, it leads to a bound on the excess regret of the order of ΔT when T is large and M is constant.

We will see in Section 5 that the bound of Theorem 1 is, in fact, optimal up to logarithmic factors. It implies that the arithmetic grid is optimal for excess regret, but most importantly, that this criterion is inadequate for this the batched bandit problem when M is small.

4.2 Competitive Ratio and the Geometric Grid

Let us consider the geometric grid $\mathcal{T} = \{t_1, \dots, t_{M-1}\}$ where $t_m = \lfloor a^m \rfloor_2$ and $a \geq 2$ is a parameter to be chosen later. Equation (3.4) gives the following upper bounds on the regret. On the one hand, if $m(\Delta, \mathcal{T}) \leq M - 2$, then

$$R_T(\Delta, \mathcal{T}) \leq 9\Delta a^{m(\Delta,\mathcal{T})} \leq 9a\Delta\tau(\Delta) \leq \frac{2304a}{\Delta} \overline{\log}\left(\frac{T\Delta^2}{128}\right).$$

On the other hand, if $m(\Delta, \mathcal{T}) = M - 1$, then $\tau(\Delta) > t_{M-2}$ and Equation (3.4) together with Lemma B.2 yield

$$R_T(\Delta, \mathcal{T}) \leq 9\Delta a^{M-1} + T\Delta e^{-\frac{a^{M-1}\Delta^2}{32}} \leq \frac{2336a}{\Delta} \overline{\log}\left(\frac{T\Delta^2}{32}\right).$$

for $a \geq 2\left(\frac{T}{\log T}\right)^{1/M} \geq 2$. In particular, we have proved the following theorem.

Theorem 2. *The* ETC *policy implemented with the geometric grid defined above for the value* $a := 2\left(\frac{T}{\log T}\right)^{1/M}$, *when* $M \leq \log(T/(\log T))$ *ensures that, for any* $\Delta \in [0, 1]$,

$$R_T(\Delta, \mathcal{T}) \lesssim \left(\frac{T}{\log T}\right)^{\frac{1}{M}} \frac{\overline{\log}\,(T\Delta^2)}{\Delta} \wedge T\Delta$$

Note that for a logarithmic number of rounds, $M = \Theta(\log T)$, the above theorem leads

to the optimal regret bound

$$R_T(\Delta, \mathcal{T}) \lesssim \frac{\overline{\log}\,(T\Delta^2)}{\Delta} \wedge T\Delta$$

This bound shows that the the discretization according to the geometric grid leads to a deterioration of the regret bound by a factor $(T/\log(T))^{\frac{1}{M}}$, which can be interpreted as a uniform bound on the competitive ratio. For $M = 2$ and $\Delta = 1$ for example, this leads to the $\sqrt{T}$ regret bound observed in the Bayesian literature and that is also optimal in the minimax sense. However, this minimax optimal bound is not valid for all values of Δ. Indeed, maximizing over $\Delta > 0$ yields

$$\sup_{\Delta} R_T(\mathcal{T}, \Delta) \lesssim T^{\frac{M+1}{2M}} \log^{\frac{M-1}{2M}}((T/\log(T))^{\frac{1}{M}}),$$

which yields the minimax rate $\sqrt{T}$ as soon as $M \geq \log(T/\log(T))$, as expected. It turns out that the decay in M can be made even faster if one focuses on the maximum risk by employing a grid designed for this purpose: the "minimax grid".

4.3 Maximum Risk and the Minimax Grid

The objective of this grid is to minimize the maximum risk, and thus recover the classical distribution independent minimax bound in $\sqrt{T}$ when there are no constraints on the number of batches (that is, when $M = T$). The intuition behind this gird comes from examining Proposition 1, in which the main term to control in order to bound the regret is $\Delta t_{m(\Delta,\mathcal{T})}$. Consider now a grid $\mathcal{T} = \{t_1, \ldots, t_{M-1}\}$, where the t_ms are defined recursively as $t_{m+1} = f(t_m)$. Such a recurrence ensures that $t_{m(\Delta,\mathcal{T})} \leq f(\tau(\Delta) - 1)$ by definition. Since we want minimize the maximum risk, we want $\Delta f(\tau(\Delta))$ to be the smallest possible term that is constant with respect to Δ. Such a choice is ensured by choosing $f(\tau(\Delta) - 1) = a/\Delta$ or, equivalently, by choosing $f(x) = a/\tau^{-1}(x + 1)$ for a suitable notion of inverse. It yields

$\Delta t_{m(\Delta,\mathcal{T})} \le a$, so that the parameter a is actually a bound on the regret. It has to be chosen large enough so that the regret $T \sup_\Delta \Delta e^{-t_{M-1}\Delta^2/8} = 2T/\sqrt{et_{M-1}}$ incurred in the go-for-broke step is also of the order of a. The formal definition below uses not only this delicate recurrence but also takes care of rounding problems.

Let $u_1 = a$, for some real number a to be chosen later, and $u_j = f(u_{j-1})$ where

$$f(u) = a\sqrt{\frac{u}{\log\left(\frac{2T}{u}\right)}}\,, \tag{4.6}$$

for all $j \in \{2, \dots, M-1\}$. The *minimax grid* $\mathcal{T} = \{t_1, \dots, t_{M-1}\}$ has points given by $t_m = \lfloor u_m \rfloor_2, m \in [M-1]$.

Observe now that if $m(\Delta, \mathcal{T}) \le M-2$, then it follows from (3.4) that $R_T(\Delta, \mathcal{T}) \le 9\Delta t_{m(\Delta,\mathcal{T})}$. Moreover, since $\tau(\Delta)$ is defined to be the smallest integer such that $\Delta \ge 16a/f(\tau(\Delta))$, we have

$$\Delta t_{m(\Delta,\mathcal{T})} \le \Delta f(\tau(\Delta) - 1) \le 16a\,.$$

Next, as discussed above, if a is chosen to be no less than $2\sqrt{2}T/(16\sqrt{et_{M-1}})$, then the regret is also bounded by $16a$ when $m(\Delta, \mathcal{T}) = M-1$. Therefore, in both cases, the regret is bounded by $16a$.

Before finding a that satisfies the above conditions, note that it follows from Lemma B.3 that

$$t_{M-1} \ge \frac{u_{M-1}}{2} \ge \frac{a^{S_{M-2}}}{30 \log^{\frac{S_{M-3}}{2}}\left(2T/a^{S_{M-5}}\right)}\,,$$

as long as $15a^{S_{M-2}} \le 2T$, where we used the notation $S_k := 2 - 2^{-k}$. Therefore, we need to choose a such that

$$a^{S_{M-1}} \ge \sqrt{\frac{15}{16e}} T \log^{\frac{S_{M-3}}{4}}\left(\frac{2T}{a^{S_{M-5}}}\right) \quad \text{and} \quad 15a^{S_{M-2}} \le 2T\,.$$

It follows from Lemma B.4 that the choice $a := (2T)^{\frac{1}{S_{M-1}}} \log^{\frac{1}{4} - \frac{3}{4}\frac{1}{2^{M-1}}}\left((2T)^{\frac{15}{2^{M-1}}}\right)$ ensures

both conditions, as soon as $2^M \leq \log(2T)/6$. We emphasize that $M = \lfloor \log_2(\log(2T)/6) \rfloor$ yields

$$\log^{\frac{1}{4}-\frac{3}{4}\frac{1}{2^M-1}}\left((2T)^{\frac{15}{2^M-1}}\right) \leq 2\,.$$

As a consequence, in order to get the optimal minimax rate of $\sqrt{T}$, one only needs $\lfloor \log_2 \log(T) \rfloor$ batches. If more batches are available, then our policy implicitly combines some of them. The above proves the following theorem:

Theorem 3. *The* ETC *policy with respect to the minimax grid defined above for the value*

$$a = (2T)^{\frac{1}{2-2^{1-M}}} \log^{\frac{1}{4}-\frac{3}{4}\frac{1}{2^M-1}}\left((2T)^{\frac{15}{2^M-1}}\right)\,,$$

ensures that, for any M such that $2^M \leq \log(2T)/6$,

$$\sup_{0\leq\Delta\leq 1} R_T(\Delta, \mathcal{T}) \lesssim T^{\frac{1}{2-2^{1-M}}} \log^{\frac{1}{4}-\frac{3}{4}\frac{1}{2^M-1}}\left(T^{\frac{1}{2^M-1}}\right).$$

For $M \geq \log_2 \log(T)$, this leads to the optimal minimax regret bound $\sup_\Delta R_T(\Delta, \mathcal{T}) \lesssim \sqrt{T}$.

Table 1 gives, for illustration purposes, the regret bounds (without constant factors) and the decision times of the ETC policy with respect to the minimax grid for values $M = 2, 3, 4, 5$.

M	$\sup_\Delta R_T(\Delta, \mathcal{T}) = t_1$	t_2	t_3	t_4
2	$T^{2/3}$			
3	$T^{4/7}\log^{1/7}(T)$	$T^{6/7}\log^{-1/7}(T)$		
4	$T^{8/15}\log^{1/5}(T)$	$T^{12/15}\log^{-1/5}(T)$	$T^{14/15}\log^{-2/5}(T)$	
5	$T^{16/31}\log^{7/31}(T)$	$T^{24/31}\log^{-5/31}(T)$	$T^{28/31}\log^{-11/31}(T)$	$T^{30/31}\log^{-14/31}(T)$

Table 1: Regret bounds and decision times of the ETC policy with respect to the minimax grid for values $M = 2, 3, 4, 5$

Note that this policy can be adapted to have only $\log_2 \log T$ switches and still achieve the optimal rate of $\sqrt{T}$. This compares favorably with the best current policies constrained to have $\log_2 \log(T)$ switches, which bound regret in $\mathcal{O}(\sqrt{T \log\log\log T})$, as in [CBDS13].

5 Lower Bounds

In this section, we address the optimality of the regret bounds derived above for the specific instance of the functionals F_{xs}, F_{cr} and F_{mx}. The results below do not merely characterize optimality (up to logarithmic terms) of the chosen grid within the class of ETC policies, but also optimality of the final policy among the class of *all M-batch policies*.

5.1 Lower Bound for the Excess Regret

As mentioned above, we show that even though the regret bound of Theorem 1 obtained using the arithmetic grid is of the trivial order $T\Delta$ when M is small, the rate $T\Delta/M$ is actually the best excess regret that any policy could achieve.

Theorem 4. *Fix $T \geq 2$ and $M \in [2 : T]$. For any M-batch policy $(\mathcal{T}, \pi)$, there exists $\Delta \in (0, 1]$ such that the policy has regret bounded below as*

$$R_T(\Delta, \mathcal{T}) \gtrsim \frac{1}{\Delta} + \frac{T}{M}\,.$$

Proof. Fix $\Delta_k = \frac{1}{\sqrt{t_k}}, k = 1 \ldots, M$. It follows from Proposition A.1 that

$$\begin{aligned}\sup_{\Delta\in(0,1]} \left\{ R_T(\Delta, \mathcal{T}) - \frac{1}{\Delta} \right\} &\geq \max_{1\leq k\leq M} \sum_{j=1}^{M} \left\{ \frac{\Delta_k t_j}{4} \exp(-t_{j-1}\Delta_k^2/2) - \frac{1}{\Delta_k} \right\} \\ &\geq \max_{1\leq k\leq M} \left\{ \frac{t_{k+1}}{4\sqrt{e t_k}} - \sqrt{t_k} \right\}\end{aligned}$$

Since $t_{k+1} \geq t_k$, the last quantity above is minimized if all the terms are all of order 1. It yields

$$t_{k+1} = t_k + a\,,$$

for some positive constant a. Since $t_M = T$, we get that $t_j \sim jT/M$ and taking $\Delta = 1$ yields

$$\sup_{\Delta \in (0,1]} \left\{ R_T(\Delta, \mathcal{T}) - \frac{1}{\Delta} \right\} \geq \frac{t_1}{4} \gtrsim \frac{T}{M} \,.$$

□

5.2 Lower Bound for the Competitive Ratio

In subsection 4.2, we established a bound on the competitive ratio that holds uniformly in Δ:

$$\frac{\Delta R_T(\Delta, \mathcal{T})}{\overline{\log}(T\Delta^2)} \lesssim \left(\frac{T}{\log T} \right)^{\frac{1}{M}} \,.$$

In this subsection, we wish to assess the quality of this bound. We show that it essentially cannot be improved, apart from logarithmic factors.

Theorem 5. *Fix $T \geq 2$ and $M \in [2 : T]$. For any M-batch policy $(\mathcal{T}, \pi)$, there exists $\Delta \in (0, 1)$ such that the policy has regret bounded below as*

$$\Delta R_T(\Delta, \mathcal{T}) \gtrsim T^{\frac{1}{M}} \,.$$

Proof. Fix $\Delta_k = \frac{1}{\sqrt{t_k}}, k = 1 \ldots, M$. It follows from Proposition A.1 that

$$\begin{aligned} \sup_{\Delta \in (0,1]} \left\{ \Delta R_T(\Delta, \mathcal{T}) \right\} &\geq \max_{1 \leq k \leq M} \sum_{j=1}^{M} \left\{ \frac{\Delta_k^2 t_j}{4} \exp(-t_{j-1}\Delta_k^2/2) \right\} \\ &\geq \max_{1 \leq k \leq M} \left\{ \frac{t_{k+1}}{4\sqrt{e} t_k} \right\} \end{aligned}$$

The last quantity above is minimized if the terms in the maximum are all of the same order, which yields

$$t_{k+1} = a t_k \,,$$

for some positive constant a. Since $t_M = T$, we get that $t_j \sim T^{j/M}$ and taking $\Delta = 1$ yields

$$\sup_{\Delta \in (0,1]} \Delta R_T(\Delta, \mathcal{T}) \geq \frac{t_1}{4} \gtrsim T^{1/M} .$$

□

5.3 Lower Bound for Maximum Regret

We conclude this section by a lower bound on the maximum regret that matches the upper bound of Theorem 3, up to logarithmic factors.

Theorem 6. *Fix $T \geq 2$ and $M \in [2 : T]$. For any M-batch policy $(\mathcal{T}, \pi)$, there exists $\Delta \in (0, 1)$ such that the policy has regret bounded below as*

$$R_T(\Delta, \mathcal{T}) \gtrsim T^{\frac{1}{2-2^{1-M}}} .$$

Proof. Fix $\Delta_k = \frac{1}{\sqrt{t_k}}, k = 1 \ldots, M$. It follows from Proposition A.1 that

$$\begin{aligned} \sup_{\Delta \in (0,1]} R_T(\Delta, \mathcal{T}) &\geq \max_{1 \leq k \leq M} \sum_{j=1}^{M} \left\{ \frac{\Delta_k t_j}{4} \exp(-t_{j-1}\Delta_k^2/2) \right\} \\ &\geq \max_{1 \leq k \leq M} \left\{ \frac{t_{k+1}}{4\sqrt{e t_k}} \right\} \end{aligned}$$

The last quantity above is minimized if all the terms are all of the same order, which yields

$$t_{k+1} = a\sqrt{t_k} ,$$

for some positive constant a. Since $t_M = T$, we get that $t_j \sim T^{2-2^{1-M}}$ and taking $\Delta = 1$ yields

$$\sup_{\Delta \in (0,1]} R_T(\Delta, \mathcal{T}) \geq \frac{t_1}{4} \gtrsim T^{\frac{1}{2-2^{1-M}}} .$$

□

6 Simulations

In this final section we compare, in simulations, the various policies (grids) introduced above. These are additionally compared with UCB2 [ACBF02], which, as noted above, can be seen as an M batch trial with $M = \Theta(\log T)$. The simulations are based both on data drawn from standard distributions and from a real medical trial: specifically data from Project AWARE, an intervention that sought to reduce the rate of sexually transmitted infections (STI) among high-risk individuals [LDLea13].

Out of the three policies introduced here, the minimax grid often does the best at minimizing regret. While all three policies are often bested by UCB2, it is important to note that the latter algorithm uses an order of magnitude more batches. This makes using UCB2 for medical trials functionally impossible. For example, in the real data we examine, the data on STI status was not reliably available until at least six months after the intervention. Thus, a three batch trial would take 1.5 years to run—as intervention and data collection would need to take place three times, at six months a piece. However, in contrast, UCB2 would use as many as 56 batches, meaning the overall experiment would take at least 28 years. Despite this extreme difference in time scales, the geometric and minimax grids produce similar levels of average regret.

6.1 Effect of Reward Distributions

We begin by examining how different distributions affect the average regret produced by different policies, for many values of the total sample, T in Figure 2. For each value of T in the figure, a sample is drawn, grids are computed based on M and T, the policy is implemented, and average regret is calculated based on the choices in the policy. This is

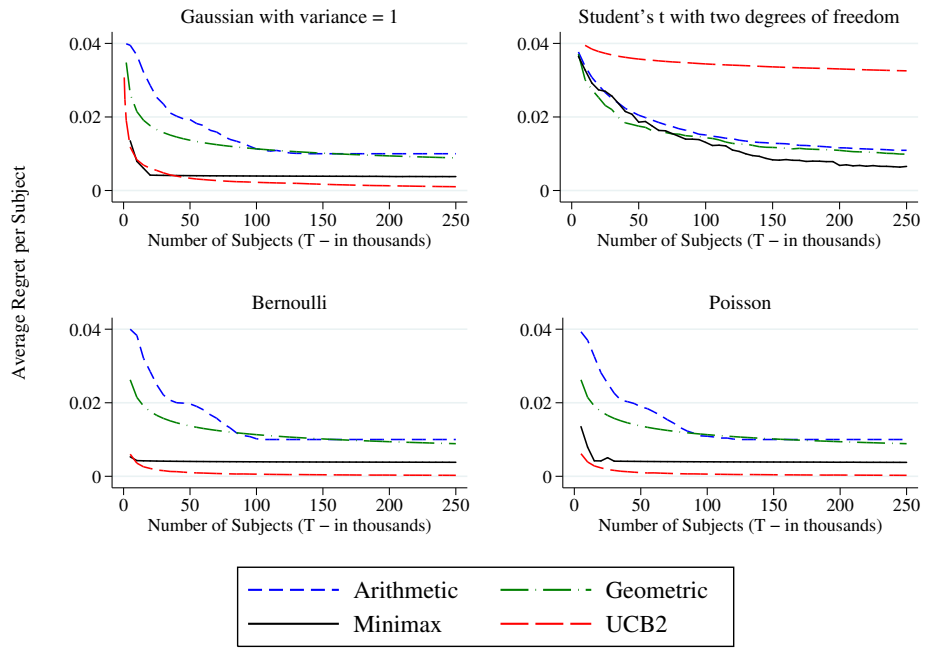

Figure 2: Performance of Policies with Different Distributions and $M = 5$. (For all distributions $\mu^{(\dagger)} = 0.5$, and $\mu^{(\star)} = 0.5 + \Delta = 0.6$.)

repeated 100 times for each value of T. Thus, each panel compares average regret for different policies as a function of the total sample T.

In all panels, the number of batches is set at $M = 5$ for all policies except UCB2. The panels each consider one of four distributions: two continuous—Gaussian and Student's t-distribution—and two discrete—Bernoulli and Poisson. In all cases, and no matter the number of subjects T, we set the difference between the arms at $\Delta = 0.1$.

A few patterns are immediately apparent. First, the arithmetic grid produces relatively constant average regret above a certain number of subjects. The intuition is straightforward: when T is large enough, the ETC policy will tend to commit after the first batch, as the first evaluation point will be greater than $\tau(\Delta)$. As in the case of the arithmetic grid, the size of this first batch is a constant proportion of the overall subject pool, average regret will be constant once T is large enough.

Second, the minimax grid also produces relatively constant average regret, although this holds for smaller values of T, and produces lower regret than the geometric or arithmetic case when M is small. This indicates, using the intuition above, that the minimax grid excels at choosing the optimal batch size to allow a decision to commit very close to $\tau(\Delta)$. This advantage versus the arithmetic and geometric grids is clear, and it can even produce lower regret than UCB2, but with an order of magnitude fewer batches. However, according to the theory above, with the minimax grid average regret is bounded by a more steeply decreasing function than is apparent in the figures. The discrepancy is due to the fact that the bounding of regret is loose for relatively small T. As T grows, average regret does decrease, but more slowly than the bound, so eventually the bound is tight at values greater than shown in the figure.

Third, and finally, the UCB2 algorithm generally produces lower regret than any of the policies considered in this manuscript for all distributions but the heavy-tailed Student's t-distribution. This increase in performance comes at a steep practical cost: many more batches. For example, with draws from a Gaussian distribution, and T between 10,000 and 40,000, the minimax grid performs better than UCB2. Throughout this range, the number of batches is fixed at $M = 5$ for the minimax grid, but UCB2 uses an average of 40–46 batches. The average number of batches used by UCB2 increases with T, and with $T = 250,000$ it reaches 56.

The fact that UCB2 uses so many more batches than the geometric grid may seem a bit surprising as both use geometric batches, leading UCB2 to have $M = \Theta(\log T)$. The difference occurs because the geometric grid uses *exactly* M batches, while the total number of batches in UCB2 is dominated by the constant terms in the range of T we consider. It should further be noted that although the level of regret is higher for the geometric grid, it is higher by a relatively constant factor.

6.2 Effect of the Gap Δ

The patterns in Figure 2 largely hold irrespective of the distribution used to generate the simulated data. Thus, in this subsection we focus on a single distribution: the exponential (to add variety), in Figure 3. What varies is the difference in mean value between the two arms, $\Delta \in \{.01, .5\}$.

In both panels of 3 the mean of the second arm is set to $\mu^{(\dagger)} = 0.5$, so Δ in these panels is 2% and 100%, respectively, of $\mu^{(\dagger)}$. This affects both the maximum average regret $T\Delta/T = \Delta$, and the number of subjects it will take to determine, using the statistical test in Section 3, which arm to commit to.

When the value of Δ is small (0.01), then in small to moderate samples T, the performance of the geometric grid and UCB2 are equivalent. When samples get large, then all of the minimax grid, the geometric grid, and UCB2 have similar performance. However, as before, UCB2 uses an order of magnitude larger number of batches—between 38–56, depending on the number T of subjects. As in Figure 2, the arithmetic grid performs poorly, but as expected based on the intuition built from the previous subsection, more subjects are needed before the performance of this grid stabilizes at a constant value. Although not shown, middling values of Δ (for example, $\Delta = 0.1$) produce the same patterns as those shown in the panels of Figure 2 (except for the panel using Student's t).

When the value of Δ is relatively large (0.5), then there is a reversal of the pattern found when Δ is relatively small. In particular the geometric grid performs poorly—worse, in fact, than the arithmetic grid—for small samples, but when the number of subjects is large, the performance of the minimax grid, geometric grid, and UCB2 are comparable. Nevertheless, the latter uses an order of magnitude more batches.

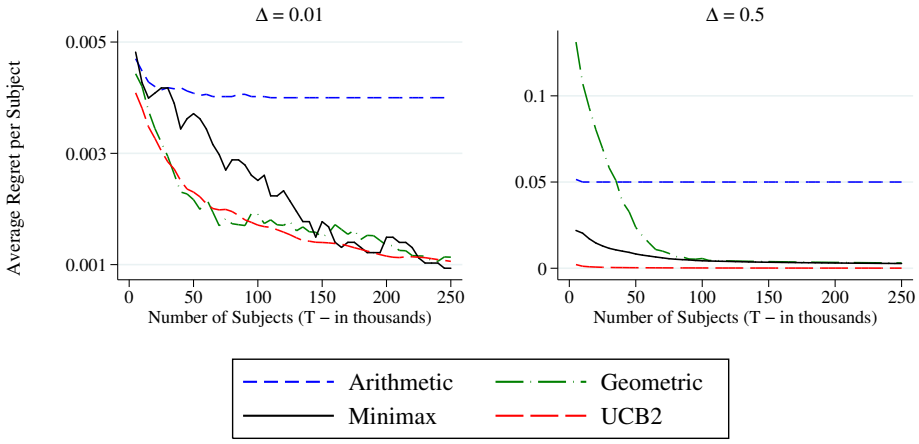

Figure 3: Performance of Policies with different Δ and $M = 5$. (For all panels $\mu^{(\dagger)} = 0.5$, and $\mu^{(\star)} = 0.5 + \Delta$.)

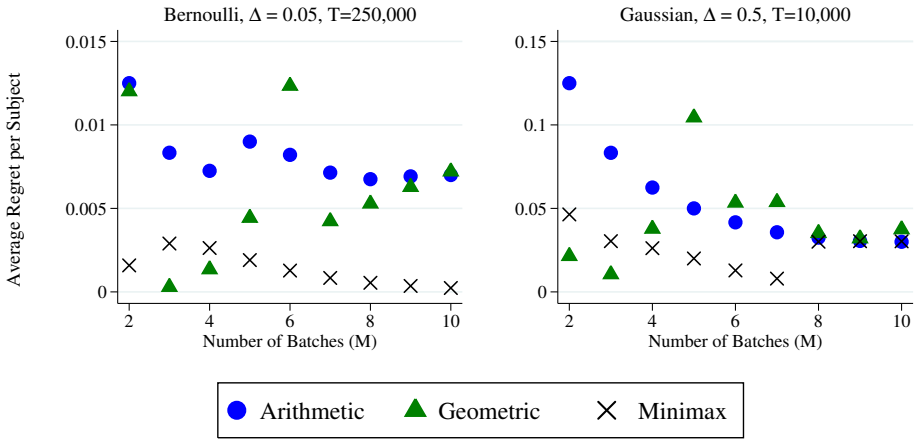

Figure 4: Performance of Policies with Different Numbers of Batches. (For all panels $\mu^{(\dagger)} = 0.5$, and $\mu^{(\star)} = 0.5 + \Delta$.)

6.3 Effect of the Number of Batches (M)

There is likely to be some variation in how well different numbers of batches perform. This is explored in Figure 4. The minimax grid's performance is consistent between $M = 2$ to $M = 10$. However, as M gets large relative to the number of subjects T and gap between

the arms Δ, all grids perform approximately equally. This occurs because as the size of the batches decrease, all grids end up with decision points near $\tau(\Delta)$.

These simulations also reveal an important point about implementation: the value of a, the termination point of the first batch, suggested in Theorems 2 and 3 is not feasible when M is "too big", that is, if it is comparable to $\log(T/(\log T))$ in the case of the geometric grid, or comparable to $\log_2 \log T$ in the case of the minimax grid. When this occurs, using this initial value of a may lead to the last batch being entirely outside of the range of T. We used the suggested a whenever feasible, but, when it was not, we selected a such that the last batch finished exactly at $T = t_M$. In the simulations displayed in Figure 4, this occurs with the geometric grid for $M \geq 7$ in the first panel, and $M \geq 6$ in the second panel. For the minimax grid, this occurs for $M \geq 8$ in the second panel. For the geometric grid, this improves performance, and for the minimax grid it slightly decrease performance. In both cases this is due to the relatively small sample, and how changing the formula for computing the grid positions decision points relative to $\tau(\Delta)$.

6.4 Real Data

Our final simulations use data from Project AWARE, a medical intervention to try to reduce the rate of sexually transmitted infections (STI) among high-risk individuals [LDLea13]. In particular, when participants went to a clinic to get an instant blood test for HIV, they were randomly assigned to receive an information sheet—control, or arm 2—or extensive "AWARE" counseling—treatment, or arm 1. The main outcome of interest was whether a participant had an STI upon six-month follow up.

The data from this trial is useful for simulations for several reasons. First, the time to observed outcome makes it clear that only a small number of batches is feasible. Second, the difference in outcomes between the arms Δ was slight, making the problem difficult. Indeed, the differences between arms was not statistically significant at conventional levels

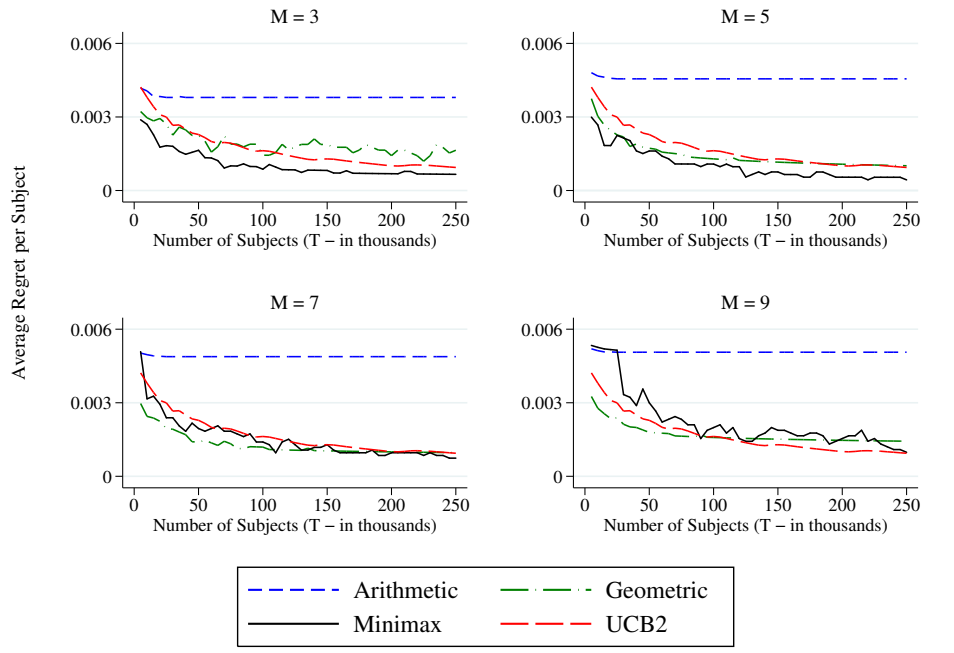

Figure 5: Performance of Policies using data from Project AWARE.

within the studied sample. Third, the trial itself was fairly large by medical trial standards, enrolling over 5,000 participants.

To simulate trials based on this data, we randomly draw observations, with replacement, from the Project AWARE participant pool. We then assign these participants to different batches, based on the outcomes of previous batches. The results of these simulations, for different numbers of participants and different numbers of batches, can be found in Figure 5. The arithmetic grid once again provides the intuition. Note that the performance of this grid degrades as the number of batches M is increased. This occurs because Δ is so small that the ETC policy does not commit until the last round, it "goes for broke". However, when doing so, the policy rarely makes a mistake. Thus, more batches cause the grid to "go for broke" later and later, resulting in worse performance.

The geometric grid and minimax grid perform similarly to UCB2, with minimax performing best with a very small number of batches ($M = 3$), and geometric performing best

with a moderate number of batches ($M = 9$). In both cases, this small difference comes from the fact that one grid or the other "goes for broke" at a slightly earlier time. As before UCB2 uses between 40–56 batches. Given the six-month time between intervention and outcome measures this suggests that a complete trial could be accomplished in 1.5 years using the minimax grid, but would take up to 28 years—a truly infeasible amount of time—using UCB2.

It is worth noting that there is nothing special in medical trials about the six-month delay between intervention and observing outcomes. Cancer drugs often measure variables like the 1- or 3-year survival rate, or increase in average survival off of a baseline that may be greater than a year. In these cases, the ability to get relatively low regret with a small number of batches is extremely important.

A Tools for Lower Bounds

Our results below hinge on lower bound tools that were recently adapted to the bandit setting in [BPR13]. Specifically, we reduce the problem of deciding which arm to pull to that of hypothesis testing. Consider the following two candidate setup for the rewards distributions: $P_1 = \mathcal{N}(\Delta, 1) \otimes \mathcal{N}(0, 1)$ and $P_2 = \mathcal{N}(0, 1) \otimes \mathcal{N}(\Delta, 1)$. This means that under P_1, successive pulls of arm 1 yield $\mathcal{N}(\Delta, 1)$ rewards and successive pulls of arm 2 yield $\mathcal{N}(0, 1)$ rewards. The opposite is true for P_2. In particular, arm $i \in \{1, 2\}$ is optimal under P_i.

At a given time $t \in [T]$ the choice of a policy $\pi_t \in [2]$ is a test between P_1^t and P_2^t where P_i^t denotes the distribution of observations available at time t under P_i. Let $R(t, \pi)$ denote the regret incurred by policy π at time t. We have $R(t, \pi) = \Delta \mathbb{1}(\pi_t \neq i)$. As a result, denoting by E_i^t the expectation with respect to P_i^t, we get

$$\begin{aligned} E_1^t\big[R(t,\pi)\big] \vee E_2^t\big[R(t,\pi)\big] &\geq \frac{1}{2}\big(E_1^t\big[R(t,\pi)\big] + E_2^t\big[R(t,\pi)\big]\big) \\ &= \frac{\Delta}{2}\big(P_1^t(\pi_t = 2) + P_2^t(\pi_t = 1)\big)\,. \end{aligned}$$

Next, we use the following lemma (see [Tsy09, Chapter 2]).

Lemma A.1. *Let P_1 and P_2 be two probability distributions such that $P_1 \ll P_2$. Then for any measurable set A, one has*

$$P_1(A) + P_2(A^c) \geq \frac{1}{2}\exp\big(-\mathsf{KL}(P_1, P_2)\big)\,.$$

where $\mathsf{KL}(P_1, P_2)$ denotes the Kullback-Leibler divergence between P_1 and P_2 and is defined by

$$\mathsf{KL}(P_1, P_2) = \int \log\Big(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\Big)\mathrm{d}P_1\,.$$

In our case, observations, are generated by an M-batch policy π. Recall that $J(t) \in [M]$ denotes the index of the current batch. Since π can only depend on observations $\{Y_s^{(\pi_s)} : s \in$

$[t_{J(t)-1}]\}$, P_i^t is a product distribution of at most $t_{J(t)-1}$ marginals. It is a standard exercise to show, whatever arm is observed in these past observations, $\mathsf{KL}(P_1^t, P_2^t) = t_{J(t)-1}\Delta^2/2$. Therefore, we get

$$E_1^t\big[R(t,\pi)\big] \vee E_2^t\big[R(t,\pi)\big] \geq \frac{1}{4}\exp\big(-t_{J(t)-1}\Delta^2/2\big)\,.$$

Summing over t yields immediately yields the following theorem

Proposition A.1. *Fix $\mathcal{T} = \{t_1, \ldots, t_M\}$ and let $(\mathcal{T}, \pi)$ be an M-batch policy. There exists reward distributions with gap Δ such that $(\mathcal{T}, \pi)$ has regret bounded below as*

$$R_T(\Delta, \mathcal{T}) \geq \Delta \sum_{j=1}^{M} \frac{t_j}{4}\exp\big(-t_{j-1}\Delta^2/2\big)\,,$$

where by convention $t_0 = 0$.

Equipped with this proposition, we can prove a variety of lower bounds as in Section 5.

B Technical Lemmas

Recall that a process $\{Z_t\}_{t\geq 0}$ is a sub-Gaussian martingale difference sequence if $\mathbb{E}\big[Z_{t+1}\big|Z_1, \ldots, Z_t\big] = 0$ and $\mathbb{E}\big[e^{\lambda Z_{t+1}}\big] \leq e^{\lambda^2/2}$ for every $\lambda > 0, t \geq 0$

Lemma B.1. *Let Z_t be a sub-Gaussian martingale difference sequence then, for every $\delta > 0$ and every integer $t \geq 1$,*

$$\mathbb{P}\left\{\bar{Z}_t \geq \sqrt{\frac{2}{t}\log\left(\frac{1}{\delta}\right)}\right\} \leq \delta.$$

Moreover, for every integer $\tau \geq 1$,

$$\mathbb{P}\left\{\exists\, t \leq \tau,\ \bar{Z}_t \geq 2\sqrt{\frac{2}{t}\log\left(\frac{4}{\delta}\frac{\tau}{t}\right)}\right\} \leq \delta.$$

PROOF. The first inequality follows from a classical Chernoff bound. To prove the maximal inequality, define $\varepsilon_t = 2\sqrt{\frac{2}{t}\log\left(\frac{4}{\delta}\frac{\tau}{t}\right)}$. Note that by Jensen's inequality, for any $\alpha > 0$, the process $\{\exp(\alpha s \bar{Z}_s)\}_s$ is a sub-martingale. Therefore, it follows from Doob's maximal inequality [Doo90, Theorem 3.2, p. 314] that for every $\eta > 0$ and every integer $t \geq 1$,

$$\mathbb{P}\left\{\exists\ s \leq t,\ s\bar{Z}_s \geq \eta\right\} = \mathbb{P}\left\{\exists\ s \leq t,\ \exp(\alpha s \bar{Z}_s) \geq \exp(\alpha\eta)\right\} \leq \mathbb{E}\left[\exp(\alpha t \bar{Z}_t)\right]\exp(-\alpha\eta)\,.$$

Next, since Z_t is sub-Gaussian, we have $\mathbb{E}\left[\exp(\alpha t \bar{Z}_t)\right] \leq \exp(\alpha^2 t/2)$. Together with the above display and optimizing with respect to $\alpha > 0$ yields

$$\mathbb{P}\left\{\exists\ s \leq t,\ s\bar{Z}_s \geq \eta\right\} \leq \exp\left(-\frac{\eta^2}{2t}\right).$$

Next, using a peeling argument, one obtains

$$\begin{aligned}
\mathbb{P}\left\{\exists\ t \leq \tau,\ \bar{Z}_t \geq \varepsilon_t\right\} &\leq \sum_{m=0}^{\lfloor \log_2(\tau)\rfloor} \mathbb{P}\Big\{\bigcup_{t=2^m}^{2^{m+1}-1}\{\bar{Z}_t \geq \varepsilon_t\}\Big\} \\
&\leq \sum_{m=0}^{\lfloor \log_2(\tau)\rfloor} \mathbb{P}\Big\{\bigcup_{t=2^m}^{2^{m+1}}\{\bar{Z}_t \geq \varepsilon_{2^{m+1}}\}\Big\} \leq \sum_{m=0}^{\lfloor \log_2(\tau)\rfloor} \mathbb{P}\Big\{\bigcup_{t=2^m}^{2^{m+1}}\{t\bar{Z}_t \geq 2^m\varepsilon_{2^{m+1}}\}\Big\} \\
&\leq \sum_{m=0}^{\lfloor \log_2(\tau)\rfloor} \exp\left(-\frac{(2^m\varepsilon_{2^{m+1}})^2}{2^{m+2}}\right) = \sum_{m=0}^{\lfloor \log_2(\tau)\rfloor} \frac{2^{m+1}}{\tau}\frac{\delta}{4} \leq \frac{2^{\log_2(\tau)+2}}{\tau}\frac{\delta}{4} \leq \delta.
\end{aligned}$$

Hence the result. ∎

Lemma B.2. *Fix two positive integers T and $M \leq \log(T)$. It holds that*

$$T\Delta e^{-\frac{a^{M-1}\Delta^2}{32}} \leq 32a\frac{\overline{\log}\left(\frac{T\Delta^2}{32}\right)}{\Delta}, \quad \textit{if}\ \ a \geq \left(\frac{MT}{\log T}\right)^{\frac{1}{M}}.$$

Proof. We first fix the value of a and we observe that $M \leq \log(T)$ implies that $a \geq e$. In

order to simplify the reading of the first inequality of the statement, we shall introduce the following quantities. Let $x > 0$ and $\theta > 0$ be defined by $x := T\Delta^2/32$ and $\theta := a^{M-1}/T$ so that the first inequality is rewritten as

$$xe^{-\theta x} \leq a\,\overline{\log}(x)\,. \tag{B.1}$$

We are going to prove that this inequality is true for all $x > 0$ given that θ and a satisfy some relation. This, in turn, gives a condition solely on a ensuring that the statement of the lemma is true for all $\Delta > 0$.

Equation (B.1) immediately holds if $x \leq e$ since $a\,\overline{\log}(x) = a \geq e$. Similarly, $xe^{-\theta x} \leq 1/(\theta e)$. Thus Equation (B.1) holds for all $x \geq 1/\sqrt{\theta}$ as soon as $a \geq a^* := 1/(\theta\,\overline{\log}(1/\theta))$. We shall assume from now on that this inequality holds. It therefore remains to check whether Equation (B.1) holds for $x \in [e, 1/\sqrt{\theta}]$. For $x \leq a$, the derivative of the right hand side is $\frac{a}{x} \geq 1$ while the derivative of the left hand side is always smaller than 1. As a consequence, Equation (B.1) holds for every $x \leq a$, in particular for every $x \leq a^*$.

To sum up, whenever

$$a \geq a^* = \frac{T}{a^{M-1}} \frac{1}{\overline{\log}(\frac{T}{a^{M-1}})}\,,$$

Equation (B.1) holds on $(0, e]$, on $[e, a^*]$ and on $[1/\sqrt{\theta}, +\infty)$, thus on $(0, +\infty)$ since $a^* \geq 1/\sqrt{\theta}$. Next if $a^M \geq MT/\log T$, we obtain

$$\frac{a}{a^*} = \frac{a^M}{T}\,\overline{\log}\left(\frac{T}{a^{M-1}}\right) \geq \frac{M}{\log(T)}\log\left(T\left(\frac{\log T}{MT}\right)^{\frac{M-1}{M}}\right) = \frac{1}{\log(T)}\log\left(T\left(\frac{\log(T)}{M}\right)^{M-1}\right).$$

As a consequence, the fact that $\log(T)/M \geq 1$ implies that $a/a^* \geq 1$, which entails the result. □

Lemma B.3. *Fix $a \geq 1, b \geq e$ and let $u_1, u_2, \ldots$ be defined by $u_1 = a$ and $u_{k+1} = a\sqrt{\frac{u_k}{\log(b/u_k)}}$.*

Define $S_k = 0$ *for* $k < 0$ *and*

$$S_k = \sum_{j=0}^{k} 2^{-j} = 2 - 2^{-k}\,, \quad \textit{for } k \geq 0\,.$$

Then, for any M *such that* $15a^{S_{M-2}} \leq b$*, it holds that for all* $k \in [M-3]$*,*

$$u_k \geq \frac{a^{S_{k-1}}}{15 \log^{\frac{S_{k-2}}{2}} \left(b/a^{S_{k-2}}\right)}.$$

Moreover, for $k \in [M-2 : M]$*, we also have*

$$u_k \geq \frac{a^{S_{k-1}}}{15 \log^{\frac{S_{k-2}}{2}} \left(b/a^{S_{M-5}}\right)}.$$

Proof. Define $z_k = \log(b/a^{S_k})$. It is not hard to show that $z_k \leq 3z_{k+1}$ iff $a^{S_{k+2}} \leq b$. In particular, $a^{S_{M-2}} \leq b$ implies that $z_k \leq 3z_{k+1}$ for all $k \in [0 : M-4]$. Next, we have

$$u_{k+1} = a\sqrt{\frac{u_k}{\log(b/u_k)}} \geq a\sqrt{\frac{a^{S_{k-1}}}{15 z_{k-2}^{\frac{S_{k-2}}{2}} \log(b/u_k)}} \tag{B.2}$$

Next, observe that since $b/a^{S_{k-1}} \geq 15$ for all $k \in [0, M-1]$, we have for all such k,

$$\log(b/u_k) \leq \log(b/a^{S_{k-1}}) + \log 15 + \frac{S_{k-2}}{2} \log z_{k-2} \leq 5z_{k-1}$$

It yields

$$z_{k-2}^{\frac{S_{k-2}}{2}} \log(b/u_k) \leq 15 z_{k-1}^{\frac{S_{k-2}}{2}} z_{k-1} = 15 z_{k-1}^{S_{k-1}}$$

Plugging this bound into (B.2) completes the proof for $k \in [M-3]$.

Finally, if $k \geq M-2$, we have by induction on k from $M-3$,

$$u_{k+1} = a\sqrt{\frac{u_k}{\log(b/u_k)}} \geq a\sqrt{\frac{a^{S_{k-1}}}{15 z_{M-5}^{\frac{S_{k-2}}{2}} \log(b/u_k)}}$$

Moreover, since $b/a^{S_{k-1}} \geq 15$ for $k \in [M-3, M-1]$, we have for such k,

$$\log(b/u_k) \leq \log(b/a^{S_{k-1}}) + \log 15 + \frac{S_{k-2}}{2} \log z_{M-5} \leq 3 z_{M-5}\,.$$

□

Lemma B.4. *If $2^M \leq \log(4T)/6$, the following specific choice*

$$a := (2T)^{\frac{1}{S_{M-1}}} \log^{\frac{1}{4} - \frac{3}{4}\frac{1}{2^M-1}} \left((2T)^{\frac{15}{2^M-1}}\right)$$

ensures that

$$a^{S_{M-1}} \geq \sqrt{\frac{15}{16e}} T \log^{\frac{S_{M-3}}{4}} \left(\frac{2T}{a^{S_{M-5}}}\right) \tag{B.3}$$

and

$$15 a^{S_{M-2}} \leq 2T\,. \tag{B.4}$$

Proof. First, the result is immediate for $M = 2$, because

$$\frac{1}{4} - \frac{3}{4}\frac{1}{2^2-1} = 0\,.$$

For $M > 2$, notice that $2^M \leq \log(4T)$ implies that

$$a^{S_{M-1}} = 2T \log^{\frac{S_{M-3}}{4}} \left((2T)^{\frac{15}{2^M-1}}\right) \geq 2T \left[16 \frac{15}{2^M-1} \log(2T)\right]^{1/4} \geq 2T\,.$$

Therefore, $a \geq (2T)^{1/S_{M-1}}$, which in turn implies that

$$a^{S_{M-1}} = 2T \log^{\frac{S_{M-3}}{4}} \left((2T)^{1-\frac{S_{M-5}}{S_{M-1}}} \right) \geq \sqrt{\frac{15}{16e}} T \log^{\frac{S_{M-3}}{4}} \left(\frac{2T}{a^{S_{M-5}}} \right).$$

This completes the proof of (B.3).

We now turn to the proof of (B.4) which follows if we prove

$$15^{S_{M-1}} (2T)^{S_{M-2}} \log^{\frac{S_{M-3}S_{M-2}}{4}} \left((2T)^{\frac{15}{2^{M-1}}} \right) \leq (2T)^{S_{M-1}}. \tag{B.5}$$

Using the trivial bounds $S_{M-k} \leq 2$, we get that the left-hand side of (B.4) is smaller than

$$15^2 \log \left((2T)^{\frac{15}{2^{M-1}}} \right) \leq 2250 \log \left((2T)^{2^{1-M}} \right).$$

Moreover, it is easy to see that $2^M \leq \log(2T)/6$ implies that the right-hand side in the above inequality is bounded by $(2T)^{2^{1-M}}$ which concludes the proof of (B.5) and thus of (B.4). □

References

[AB10] Jean-Yves Audibert and Sébastien Bubeck, *Regret bounds and minimax policies under partial monitoring*, J. Mach. Learn. Res. **11** (2010), 2785–2836.

[ACBF02] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer, *Finite-time analysis of the multiarmed bandit problem*, Mach. Learn. **47** (2002), no. 2-3, 235–256.

[AO10] P. Auer and R. Ortner, *UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem*, Periodica Mathematica Hungarica **61** (2010), no. 1, 55–65, A revised version is available at http://personal.unileoben.ac.at/rortner/Pubs/UCBRev.pdf.

[Bar07] Jay Bartroff, *Asymptotically optimal multistage tests of simple hypotheses*, Ann. Statist. **35** (2007), no. 5, 2075–2105. MR 2363964 (2009e:62324)

[Bat81] J. A. Bather, *Randomized allocation of treatments in sequential experiments*, J. Roy. Statist. Soc. Ser. B **43** (1981), no. 3, 265–292, With discussion and a reply by the author.

[BF85] D.A. Berry and B. Fristedt, *Bandit problems: Sequential allocation of experiments*, Monographs on Statistics and Applied Probability Series, Chapman & Hall, London, 1985.

[BLS13] Jay Bartroff, Tze Leung Lai, and Mei-Chiung Shih, *Sequential experimentation in clinical trials*, Springer Series in Statistics, Springer, New York, 2013, Design and analysis. MR 2987767

[BM07] Dimitris Bertsimas and Adam J. Mersereau, *A learning approach for interactive marketing to a customer segment*, Operations Research **55** (2007), no. 6, 1120–1135.

[BPR13] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, *Bounded regret in stochastic multi-armed bandits*, COLT 2013 - The 26th Conference on Learning Theory, Princeton, NJ, June 12-14, 2013 (Shai Shalev-Shwartz and Ingo Steinwart, eds.), JMLR W&CP, vol. 30, 2013, pp. 122–134.

[CBDS13] Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir, *Online learning with switching costs and other adaptive adversaries*, Advances in Neural Information Processing Systems 26 (C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, eds.), Curran Associates, Inc., 2013, pp. 1160–1168.

[CBGM13] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour, *Regret minimization for reserve prices in second-price auctions*, Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '13, SIAM, 2013, pp. 1190–1204.

[CG09] Stephen E. Chick and Noah Gans, *Economic analysis of simulation selection problems*, Management Science **55** (2009), no. 3, 421–437.

[CGM+13] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz, *Kullback–leibler upper confidence bounds for optimal sequential allocation*, Ann. Statist. **41** (2013), no. 3, 1516–1541.

[Che96] Y. Cheng, *Multistage bandit problems*, J. Statist. Plann. Inference **53** (1996), no. 2, 153–170.

[CJW07] Richard Cottle, Ellis Johnson, and Roger Wets, *George B. Dantzig (1914–2005)*, Notices Amer. Math. Soc. **54** (2007), no. 3, 344–362.

[Col63] Theodore Colton, *A model for selecting one of two medical treatments*, Journal of the American Statistical Association **58** (1963), no. 302, pp. 388–400.

[Col65] ———, *A two-stage model for selecting one of two treatments*, Biometrics **21** (1965), no. 1, pp. 169–180.

[Dan40] George B. Dantzig, *On the non-existence of tests of student's hypothesis having power functions independent of σ*, The Annals of Mathematical Statistics **11** (1940), no. 2, 186–192.

[Doo90] J. L. Doob, *Stochastic processes*, Wiley Classics Library, John Wiley & Sons, Inc., New York, 1990, Reprint of the 1953 original, A Wiley-Interscience Publication.

[FZ70] J. Fabius and W. R. Van Zwet, *Some remarks on the two-armed bandit*, The Annals of Mathematical Statistics **41** (1970), no. 6, 1906–1916.

[GR54] S. G. Ghurye and Herbert Robbins, *Two-stage procedures for estimating the difference between means*, Biometrika **41** (1954), 146–152.

[HS02] Janis Hardwick and Quentin F. Stout, *Optimal few-stage designs*, J. Statist. Plann. Inference **104** (2002), no. 1, 121–145.

[JT00] Christopher Jennison and Bruce W. Turnbull, *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, Boca Raton, FL, 2000.

[LDLea13] Metsch LR, Feaster DJ, Gooden L, and et al, *Effect of risk-reduction counseling with rapid hiv testing on risk of acquiring sexually transmitted infections: The aware randomized clinical trial*, JAMA **310** (2013), no. 16, 1701–1710.

[LR85] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in Applied Mathematics **6** (1985), 4–22.

[Mau57] Rita J. Maurice, *A minimax procedure for choosing between two populations using sequential sampling*, Journal of the Royal Statistical Society. Series B (Methodological) **19** (1957), no. 2, pp. 255–261.

[PR13] Vianney Perchet and Philippe Rigollet, *The multi-armed bandit problem with covariates*, Ann. Statist. **41** (2013), no. 2, 693–721.

[Rob52] Herbert Robbins, *Some aspects of the sequential design of experiments*, Bulletin of the American Mathematical Society **58** (1952), no. 5, 527–535.

[SBF13] Eric M. Schwartz, Eric Bradlow, and Peter Fader, *Customer acquisition via display advertising using multi-armed bandit experiments*, Tech. report, University of Michigan, 2013.

[Som54] Paul N. Somerville, *Some problems of optimum sampling*, Biometrika **41** (1954), no. 3/4, pp. 420–429.

[Ste45] Charles Stein, *A two-sample test for a linear hypothesis whose power is independent of the variance*, The Annals of Mathematical Statistics **16** (1945), no. 3, 243–258.

[Tho33] William R. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25** (1933), no. 3/4, 285–294.

[Tsy09] Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR 2724359 (2011g:62006)

[Vog60a] Walter Vogel, *An asymptotic minimax theorem for the two armed bandit problem*, Ann. Math. Statist. **31** (1960), 444–451.

[Vog60b] ———, *A sequential design for the two armed bandit*, The Annals of Mathematical Statistics **31** (1960), no. 2, 430–443.