The Annals of Statistics
2016, Vol. 44, No. 2, 660–681
DOI: 10.1214/15-AOS1381

BATCHED BANDIT PROBLEMS

BY VIANNEY PERCHET[*,1], PHILIPPE RIGOLLET[†,2],
SYLVAIN CHASSANG[‡,3] AND ERIK SNOWBERG[§,3]

Université Paris Diderot and INRIA, Massachusetts Institute of Technology[†], Princeton University[‡] and California Institute of Technology and NBER[§]*

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. We propose a simple policy, and 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.

1. Introduction. All clinical trials are run in *batches*: groups of patients are treated simultaneously, with the data from each batch influencing the design of the next. This structure arises as it is impractical to measure outcomes (rewards) for each patient before deciding what to do next. Despite the fact that this system is codified into law for drug approval, it has received scant attention from statisticians. What can be achieved with a small number of batches? How big should these batches be? How should results in one batch affect the structure of the next?

We address these questions using the multi-armed bandit framework. This encapsulates an "exploration vs. exploitation" dilemma fundamental to ethical clinical research [30, 34]. In the basic problem, there are two populations of patients (or *arms*), corresponding to different treatments. At each point in time $t = 1, \dots, T$, a decision maker chooses to sample one, and receives a random reward dictated by the efficacy of the treatment. The objective is to devise a series of choices—a policy—maximizing the expected cumulative reward over T rounds. There is thus 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*.

The importance of batching extends beyond clinical trials. In recent years, the bandit framework has been used to study problems in economics, finance, chemical engineering, scheduling, marketing and, more recently, internet advertising.

Received May 2015; revised August 2015.

[1] Supported by ANR Grant ANR-13-JS01-0004.

[2] Supported by NSF Grants DMS-13-17308 and CAREER-DMS-10-53987.

[3] Supported by NSF Grant SES-1156154.

MSC2010 subject classifications. Primary 62L05; secondary 62C20.

Key words and phrases. Multi-armed bandit problems, regret bounds, batches, multi-phase allocation, grouped clinical trials, sample size determination, switching cost.

This last application has been the driving force behind a recent surge of interest in many variations of bandit problems over the past decade. Yet, even in internet advertising, technical constraints often force data to be considered in batches; although the size of these batches is usually based on technical convenience rather than on statistical reasoning. Discovering the optimal structure, size and number of batches has applications in marketing [8, 31] and simulations [14].

In clinical trials, batches may be formal—the different phases required for approval of a new drug by the US Food and Drug Administration—or informal—with a pilot, a full trial, and then diffusion to the full population that may benefit. In an informal setup, the second step may be skipped if the pilot is successful enough. In this three-stage approach, the first, and usually second, phases focus on exploration, while the third focuses on exploitation. This is in stark contrast to the basic bandit problem described above, which effectively consists of T batches, each containing a single patient.

We describe a policy that performs well with a small fixed number of batches. A fixed number of batches reflects clinical practice, but presents mathematical challenges. Nonetheless, we identify batch sizes that lead to a minimax regret bounds as low as the best non-batched algorithms. We further show that these batch sizes perform well empirically. Together, these features suggest that near-optimal policies could be implemented with only small changes to current clinical practice.

2. Description of the problem.

2.1. *Notation.* For any positive integer n, define $[n] = \{1, \dots, n\}$, and for any $n_1 < n_2$, $[n_1 : n_2] = \{n_1, \dots, n_2\}$ and $(n_1 : n_2] = \{n_1 + 1, \dots, n_2\}$. For any positive number x, let $\lfloor x \rfloor$ denote the largest integer n such that $n \le x$ and $\lfloor x \rfloor_2$ denotes the largest *even* integer m such that $m \le x$. Additionally, for any real numbers a and b, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Further, define $\overline{\log}(x) = 1 \vee (\log x)$. $\mathbb{1}(\cdot)$ denotes the indicator function.

If $\mathcal{I}$, $\mathcal{J}$ are closed intervals of $\mathbb{R}$, then $\mathcal{I} \prec \mathcal{J}$ if $x < y$ for all $x \in \mathcal{I}$, $y \in \mathcal{J}$.

Finally, 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 $|u_T| \le C|v_T|$ for any T. Moreover, we write $u_T = \Theta(v_T)$ if $u_T = \mathcal{O}(v_T)$ and $v_T = \mathcal{O}(u_T)$.

2.2. *Framework.* We employ a two-armed bandit framework with horizon $T \ge 2$. Central ideas and intuitions are well captured by this concise framework. Extensions to K-armed bandit problems are mostly technical (see, for instance, [28]).

At each time $t \in [T]$, the decision maker chooses an arm $i \in \{1, 2\}$ and observes a reward that comes from a sequence of i.i.d. draws $Y_1^{(i)}, Y_2^{(i)}, \dots$ from some unknown distribution $\nu^{(i)}$ with expected value $\mu^{(i)}$. We assume that the distributions $\nu^{(i)}$ are standardized sub-Gaussian, that is, $\int e^{\lambda(x-\mu^{(i)})}\nu_i(dx) \le e^{\lambda^2/2}$ 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. Rescaling extends the framework to other variance parameters σ^2.

For any integer $M \in [2 : T]$, let $\mathcal{T} = \{t_1, \dots, t_M\}$ be an ordered sequence, or *grid*, of integers such that $1 < t_1 < \cdots < t_M = T$. It defines a partition $\mathcal{S} = \{S_1, \dots, S_M\}$ 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 *kth batch*. An *M-batch policy* is a couple $(\mathcal{T}, \pi)$ where $\mathcal{T} = \{t_1, \dots, t_M\}$ is a grid and $\pi = \{\pi_t, t = 1, \dots, T\}$ is a sequence of random variables $\pi_t \in \{1, 2\}$, indicating which arm to pull at each time $t = 1, \dots, T$, which depend only on observations from batches strictly prior to the current one. 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 $\{Y_s^{(\pi_s)} : s \in S_1 \cup \cdots \cup S_{J(t)-1}\} = \{Y_s^{(\pi_s)} : s \le t_{J(t)-1}\}$.

Denote by $\star \in \{1, 2\}$ the optimal arm defined by $\mu^{(\star)} = \max_{i \in \{1,2\}} \mu^{(i)}$, by $\dagger \in \{1, 2\}$ the suboptimal arm, and by $\Delta := \mu^{(\star)} - \mu^{(\dagger)} > 0$ the gap between the optimal expected reward and the suboptimal expected reward.

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

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

Denoting by $T_i(t) = \sum_{s=1}^{t} \mathbb{1}(\pi_s = i), i \in \{1, 2\}$ the number of times arm i was pulled before time $t \ge 2$, regret can be rewritten as $R_T = \Delta \mathbb{E} T_\dagger(T)$.

2.3. *Previous results.* Bandit problems are well understood in the case where $M = T$, that is, when the decision maker can use all available 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 uniformly in Δ over a suitable subset of the positive real line such as the intervals $(0, 1)$ or even $(0, \infty)$. The first results of this kind are attributed to Vogel [36, 37], who proved that $R_T = \Theta(\sqrt{T})$ in the two-armed case (see also [6, 20]).

Adaptive policies exhibit regret bounds that may be much smaller than the order of $\sqrt{T}$ when Δ is large. Such bounds were proved in the seminal paper of Lai and Robbins [25] in an asymptotic framework (see also [10]). While leading to tight constants, this framework washes out the correct dependency on Δ of the logarithmic terms. In fact, recent research [1–3, 28] has revealed that $R_T = \Theta(\Delta T \wedge \overline{\log}(T\Delta^2)/\Delta)$.

Nonetheless, a systematic analysis of the batched case does not exist, even though UCB2 [2] and IMPROVED-UCB [3] are implicitly M-batch policies with $M = \Theta(\log T)$. These algorithms achieve optimal adaptive bounds. Thus, employing a batched policy is only a constraint when the number of batches M is much smaller than $\log T$, as is often the case in clinical practice. Similarly, in the minimax framework, M-batch policies, with $M = \Theta(\log\log T)$, lead to the optimal regret bound (up to logarithmic terms) of $\mathcal{O}(\sqrt{T \log\log\log T})$ [11, 12]. The sublogarithmic range $M \ll \log T$ is essential in applications where M is small and constant, like clinical trials. In particular, we wish to bound the regret for small values of M, such as 2, 3 or 4.

2.4. *Literature.* This paper connects to two lines of work: batched sequential estimation [17, 18, 21, 33] and multistage clinical trials. Somerville [32] and Maurice [26] 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; a result we recover and extend (see Section 4.3).

Colton [15, 16] introduced a Bayesian perspective, initiating a long line of work (see [22] for a recent overview). Most of this work focuses on the case of two-three batches, with isolated exceptions [13, 22]. Typically, this work claims the size of the first batch should be of order $\sqrt{T}$, which agrees with our results, up to a logarithmic term (see Section 4.2).

Batched procedures have a long history in clinical trials (see, for instance, [23] and [5]). Usually, batches are of the same size, or of random size, with the latter case providing robustness. This literature also focuses on inference questions rather than cumulative regret. A notable exception provides an ad-hoc objective to optimize batch size but recovers the suboptimal $\sqrt{T}$ in the case of two batches [4].

2.5. *Outline.* Section 3 introduces a general class of M-batch policies we call *explore-then-commit* (ETC) *policies*. These policies are close to clinical practice within batches. The performance of generic ETC policies are detailed in Proposition 1, found 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 of batches M. Indeed, in Section 4.3, we describe a policy in which 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:

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

$$R_T(\pi^2) \lesssim T^{1/(2-2^{1-M})} \log^{\alpha_M}\bigl(T^{1/(2^M-1)}\bigr), \qquad \alpha_M \in [0, 1/4).$$

Note that the bound for π^1 corresponds to the optimal adaptive rate $\overline{\log}(T\Delta^2)/\Delta$ when $M = \Theta(\log(T/\log(T)))$ and the bound for π^2 corresponds to the optimal minimax rate $\sqrt{T}$ when $M = \Theta(\log\log T)$. The latter is entirely feasible in clinical settings. As a byproduct of our results, we show 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. 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 the policies we identify perform well even with a very small number of batches.

3. Explore-then-commit policies.

In this section, we describe a simple structure that can be used to build policies: *explore-then-commit* (ETC). This structure consists of pulling each arm the same number of times in each non-terminal batch, and checking after each batch whether, according to some statistical test, one arm dominates the other. If one dominates, then only that arm is pulled until T. If, at the beginning of the terminal batch, neither arm has been declared dominant, then the policy commits to the arm with the largest average past reward. This "go for broke" step is dictated by regret minimization: in the last batch exploration is pointless as the information it produces can never be used.

Any policy built using this principle is completely characterized by two elements: the testing criterion and the sizes of the batches.

3.1. *Statistical test.*

We begin by describing the statistical test employed before 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 $\mu^{(i)}$ of the form

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

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

$$\mathbb{P}\{\exists s \leq \tau : \mu^{(i)} > \widehat{\mu}_s^{(i)} + \mathsf{B}_s^{(i)}\} \vee \mathbb{P}\{\exists s \leq \tau : \mu^{(i)} < \widehat{\mu}_s^{(i)} - \mathsf{B}_s^{(i)}\} \leq \frac{4\tau}{T}. \tag{1}$$

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. This test is only implemented at times $t \in [T]$ at which each arm has been pulled exactly $s = t/2$ times. However, for completeness, we define the test at all times t. For $t \geq 1$, define

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

The errors of such tests are controlled as follows.

LEMMA 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*

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

Let $\bar{t}$ denote the smallest element of $\mathcal{S}_+$. Then

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

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

(i) By definition,

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

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

Finally, for any $t \in \mathcal{S}_+$, in particular for $t = \bar t$, we have

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

(ii) Focus on the case $t \in \mathcal{S}_-$, where $\Delta < 16\sqrt{\log(2T/t)/t}$. Here,

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

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

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

3.2. *Go for broke.* In the last batch, the ETC structure 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. The proof follows immediately from Lemma B.1.

LEMMA 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*

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

3.3. *Explore-then-commit policy.* In a batched process, an extra constraint is that past observations can only be inspected at a specific set of times $\mathcal{T} = \{t_1, \dots, t_{M-1}\} \subset [T]$, called 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 $t_1, \dots, t_{M-2}$, the policy implements the statistical test. If one arm is determined to be better than the other, it is pulled until T. If no arm is declared best, then both arms are pulled 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. With N an even num-

```
Input:

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

Initialization:

• Let $\varepsilon^{[m]} = (\varepsilon_1^{[m]}, \dots, \varepsilon_{|S_m|}^{[m]})$ be uniformly distributed over[a] $\mathcal{V}_{|S_m|}$, for $m \in$
  $[M]$.
• The index $\ell$ 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, that is, $\pi_t = \psi_{t_{M-1}}$.
```

[a] In the case where $|S_m|$ is not an even number, we use the general definition of footnote 4 for $\mathcal{V}_{|S_m|}$.

FIG. 1. *Generic explore-then-commit policy with grid $\mathcal{T}$.*

ber, let $(\varepsilon_1, \dots, \varepsilon_N)$ be uniformly distributed over the subset $\mathcal{V}_N = \{v \in \{1,2\}^N : \sum_i \mathbb{1}(v_i = 1) = N/2\}$.[4] This randomization has no effect on the policy, and could easily be replaced by any other mechanism that pulls each arm an equal number of times. For example, a mechanism that pulls one arm for the first half of the batch, and the other for the second half, may be used if switching costs are a concern.

In the terminal batch S_M, if no arm was determined to be optimal in any prior batch, 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 the above definition implies that $\tau(\Delta) \geq 2$ and

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

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

$$m(\Delta, \mathcal{T}) = \begin{cases} \min\{m \in \{1, \dots, M-1\} : t_m \geq \tau(\Delta)\}, & \text{if } \tau(\Delta) \leq t_{M-1}, \\ M-1, & \text{otherwise.} \end{cases} \tag{3}$$

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

$$R_T(\Delta, \mathcal{T}) \leq 9\Delta t_{m(\Delta,\mathcal{T})} + T\Delta e^{-(t_{M-1}\Delta^2)/16}\mathbb{1}\big(m(\Delta, \mathcal{T}) = M-1\big). \tag{4}$$

PROOF. Denote $\bar{m} = m(\Delta, \mathcal{T})$. Note that $t_{\bar{m}}$ denotes the theoretical time on the grid at which the statistical test will declare $\star$ to be (with high probability) the better arm.

[4] Odd numbers for the deadlines t_i could be considered, at the cost of rounding problems and complexity, by defining $\mathcal{V}_N = \{v \in \{1,2\}^N : |\sum_i \mathbb{1}(v_i = 1) - \sum_i \mathbb{1}(v_i = 2)| \leq 1\}$.

We first examine the case where $t_{\bar{m}} < M - 1$. Define the following events:

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

Regret can be incurred in one of the following three manners:

(i) by exploring before time $t_{\bar{m}}$,
(ii) by choosing arm † before time $t_{\bar{m}}$: this happens on event B_m,
(iii) by not committing to the optimal arm ⋆ 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. 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

$$R_T(\Delta, \mathcal{T}) \le \frac{t_{\bar{m}}\Delta}{2} + T\Delta\mathbb{E}\left[\mathbb{1}\left(\bigcup_{m=1}^{\bar{m}-1} A_{m-1} \cap B_m\right) + \mathbb{1}(B_{\bar{m}-1} \cap C_{\bar{m}})\right], \tag{5}$$

with the convention that A_0 is the whole probability space.

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

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

In particular, $t_{\bar{m}}$ plays the role of $\bar{t}$ in Lemma 1. Thus, using part (i) of Lemma 1,

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

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

$$\mathbb{P}\left(\bigcup_{m=1}^{\bar{m}-1} A_{m-1} \cap B_m\right) \le \frac{4t_{\bar{m}}}{T}.$$

Together with (5) this implies regret is bounded by $R_T(\Delta, \mathcal{T}) \le 9\Delta t_{\bar{m}}$.

In the case where $t_{m(\Delta,\mathcal{T})} = M - 1$, Lemma 2 shows that the go for broke test errs with probability at most $\exp(-t_{M-1}\Delta^2/16)$, which gives that

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

using the same arguments as before. □

Proposition 1 helps choose a grid by showing how that choice reduces to an optimal discretization problem.

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 (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 [15, 16], 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 functionals of $R(\cdot, \mathcal{T})$. The functionals we focus on are:

$$F_{\mathsf{xs}}\big[R_T(\cdot, \mathcal{T})\big] = \sup_{\Delta \in [0,1]} \big\{R_T(\Delta, \mathcal{T}) - CR^*(\Delta)\big\}, \qquad C > 0 \qquad \text{Excess regret,}$$

$$F_{\mathsf{cr}}\big[R_T(\cdot, \mathcal{T})\big] = \sup_{\Delta \in [0,1]} \frac{R_T(\Delta, \mathcal{T})}{R^*(\Delta)} \qquad \text{Competitive ratio,}$$

$$F_{\mathsf{mx}}\big[R_T(\cdot, \mathcal{T})\big] = \sup_{\Delta \in [0,1]} R_T(\Delta, \mathcal{T}) \qquad \text{Maximum.}$$

Optimizing different functionals leads to different optimal grids. We investigate the properties of these functionals and grids in the rest of this section.[5]

4.1. *Excess regret and the arithmetic grid.* We begin with the simple grid consisting in a uniform discretization of $[T]$. This is particularly prominent in the group sequential testing literature [23]. As we will see, even in a favorable setup, it yields poor regret bounds.

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 (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 regret. To obtain a valid upper bound, note that

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

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 (4), this yields the following theorem.

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

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

[5] One could also consider the Bayesian criterion $F_{\mathsf{by}}[R_T(\cdot, \mathcal{T})] = \int R_T(\Delta, \mathcal{T})\, d\pi(\Delta)$ where π is a given prior distribution on Δ, rather than on the expected rewards as in the traditional Bayesian bandit literature [7].

The optimal rate is recovered if $M = T$. However, the arithmetic grid leads to a bound on the excess regret of the order of ΔT when T is large and M constant.

In Section 5, the bound of Theorem 1 is shown to be optimal for excess regret, up to logarithmic factors. Clearly, this criterion provides little useful guidance on how to attack the batched bandit problem when M is small.

4.2. *Competitive ratio and the geometric grid.* The geometric grid is defined as $\mathcal{T} = \{t_1, \dots, t_{M-1}\}$, where $t_m = \lfloor a^m \rfloor_2$, and $a \ge 2$ is a parameter to be chosen later. To bound regret using (4), note that if $m(\Delta, \mathcal{T}) \le M - 2$, then

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

and if $m(\Delta, \mathcal{T}) = M - 1$, then $\tau(\Delta) > t_{M-2}$. Then, (4), together with Lemma B.2 yields

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

for $a \ge 2(\frac{T}{\log T})^{1/M} \ge 2$. We have proved the following theorem.

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

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

For a logarithmic number of batches, $M = \Theta(\log T)$, the geometric grid 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 geometric grid leads to a deterioration of the regret bound by a factor $(T/\log(T))^{1/M}$, which can be interpreted as a uniform bound on the competitive ratio. For example, for $M = 2$ and $\Delta = 1$, this leads to the $\sqrt{T}$ regret bound observed in the Bayesian literature, which 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

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

which yields the minimax rate $\sqrt{T}$ when $M \ge \log(T/\log(T))$, as expected from prior results. The decay in M can be made even faster if one focuses on the maximum risk, by employing our "minimax grid."

4.3. *Maximum risk and the minimax grid.* The objective of this grid is to minimize the maximum risk, and to recover the classical distribution independent minimax bound in $\sqrt{T}$. The intuition behind this grid comes from Proposition 1, in which $\Delta t_{m(\Delta,\mathcal{T})}$ is the most important term to control. Consider a grid $\mathcal{T} = \{t_1, \dots, t_{M-1}\}$, where the t_m's are defined recursively as $t_{m+1} = f(t_m)$ so that, by definition, $t_{m(\Delta,\mathcal{T})} \le f(\tau(\Delta) - 1)$. As we minimize the maximum risk, $\Delta f(\tau(\Delta))$ should be the smallest possible term, and constant with respect to Δ. This 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 the inverse. This yields $\Delta t_{m(\Delta,\mathcal{T})} \le a$, so that the parameter a is actually a bound on the regret. This parameter also has to be 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 $a > 0$ to be chosen later, and $u_j = f(u_{j-1})$ where

$$f(u) = a\sqrt{\frac{u}{\log((2T)/u)}} \tag{6}$$

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

If $m(\Delta, \mathcal{T}) \le M - 2$, then it follows from (4) that $R_T(\Delta, \mathcal{T}) \le 9\Delta t_{m(\Delta,\mathcal{T})}$, and as $\tau(\Delta)$ is the smallest integer such that $\Delta \ge 16a/f(\tau(\Delta))$, we have

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

As discussed above, if a is greater 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 an a satisfying the above conditions, note that it follows from Lemma B.3 that, as long as $15a^{S_{M-2}} \le 2T$,

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

with the notation $S_k := 2 - 2^{-k}$. Therefore, we need to choose a such that

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

It follows from Lemma B.4 that the choice

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

ensures both conditions when $2^M \le \log(2T)/6$. We emphasize that

$$\log^{1/4-(3/4)1/(2^M-1)}\big((2T)^{15/(2^M-1)}\big) \le 2 \qquad \text{with } M = \big\lfloor \log_2\big(\log(2T)/6\big)\big\rfloor.$$

TABLE 1
Regret and decision times of the ETC *policy with the minimax grid for* $M = 2, 3, 4, 5$. *In the table,* $l_T = \log(T)$

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

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. We have proved the following theorem.

THEOREM 3. *The* ETC *policy over the minimax grid with*

$$a = (2T)^{1/(2-2^{1-M})} \log^{1/4-(3/4)1/(2^M-1)}\big((2T)^{15/(2^M-1)}\big)$$

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

$$\sup_{0 \le \Delta \le 1} R_T(\Delta, \mathcal{T}) \lesssim T^{1/(2-2^{1-M})} \log^{1/4-(3/4)1/(2^M-1)}\big(T^{1/(2^M-1)}\big),$$

which is minimax optimal, that is, $\sup_\Delta R_T(\Delta, \mathcal{T}) \lesssim \sqrt{T}$, *for* $M \ge \log_2 \log(T)$.

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

The ETC policy with the minimax grid can easily be adapted to have only $O(\log\log T)$ switches, and yet still achieve regret of optimal order $\sqrt{T}$. To do so, in each batch one arm should be pulled for the first half of the batch, and the other for the second half, leading to only one switch within the batch, until the policy commits to a single arm. To ensure that a switch does not occur between batches, the first arm pulled in a batch should be set to the last arm pulled in the previous batch, assuming that the policy has not yet committed. This strategy is relevant in applications such as labor economics and industrial policy, where switching from an arm to the other may be expensive [24]. In this context, our policy compares favorably with the best current policies constrained to have $\log_2 \log(T)$ switches, which lead to a regret bound of order $\sqrt{T \log\log\log T}$ [11].

5. Lower bounds.

In this section, we address the optimality of the regret bounds derived above for the specific 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*.

THEOREM 4. *Fix $T \geq 2$ and $M \in [2:T]$. Any M-batch policy $(\mathcal{T}, \pi)$, must satisfy the following lower bounds*:

$$\sup_{\Delta \in (0,1]} \left\{ R_T(\Delta, \mathcal{T}) - \frac{1}{\Delta_{\mathsf{xs}}} \right\} \gtrsim \frac{T}{M},$$

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

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

PROOF. Fix $\Delta_k = \frac{1}{\sqrt{t_k}}$, $k = 1, \dots, M$. Focusing first on excess risk, it follows from Proposition A.1 that

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

As $t_{k+1} \geq t_k$, the last quantity above is minimized if all the terms are of order 1. This yields $t_{k+1} = t_k + a$, for some positive constant a. As $t_M = T$, we get that $t_j \sim jT/M$, and taking $\Delta = 1$ yields

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

Proposition A.1 also yields

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

Arguments similar to the ones for the excess regret above, give the lower bound for the competitive ratio. Finally,

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

gives the lower bound for maximum risk. □

6. Simulations. In this final section, we briefly compare, in simulations, the various policies (grids) introduced above. These are also compared with UCB2 [2], which, as noted above, can be seen as an $M = O(\log T)$ batch trial. A more complete exploration can be found in [29].

The minimax and geometric grids perform well using an order of magnitude fewer batches than UCB2. The number of batches required for UCB2 make its use

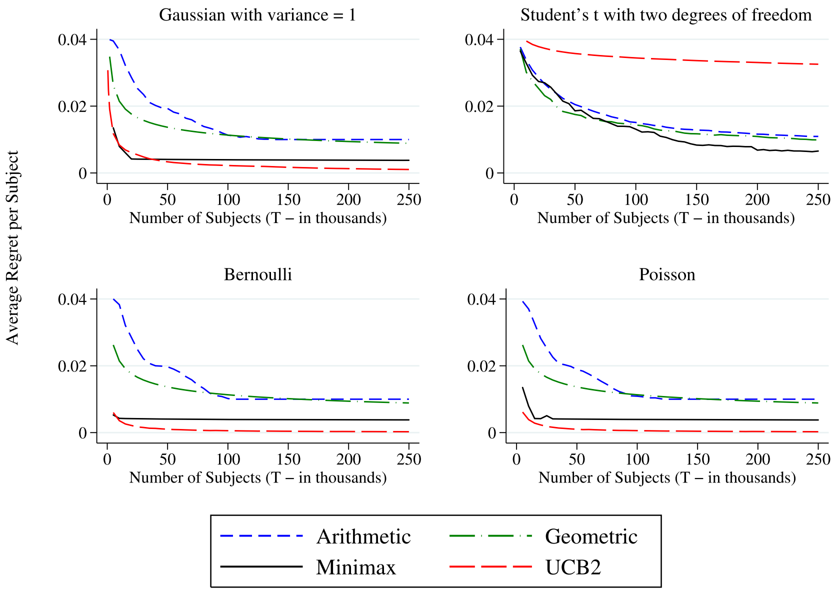

FIG. 2. *Performance of policies with different distributions and* $M = 5$. (*For all distributions* $\mu^{(\dagger)} = 0.5$, *and* $\mu^{(\star)} = 0.5 + \Delta = 0.6$.)

for medical trials functionally impossible. For example, a study that examined STI status six months after an intervention in [27] would require 1.5 years to run using minimax batch sizes, but UCB2 would use as many as 56 batches, meaning the study would take 28 years.

Specific examples of performance can be found in Figure 2. This figure compares average regret produced by different policies and many values of the total sample, T. 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 repeated 100 times for each value of T.

The number of batches is set at $M = 5$ for all policies except UCB2. Each panel considers one of four distributions: two continuous—Gaussian and Student's t-distribution—and two discrete—Bernoulli and Poisson. In all cases, 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 participants. 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 $\tau(\Delta)$. In the arithmetic grid, the size of this first batch is a constant proportion of the overall participant pool, so average regret will be constant when 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 over the arithmetic and geometric grids is clear. The minimax grid can even produce lower regret than UCB2, using an order of magnitude fewer batches.

Third, and finally, the UCB2 algorithm generally produces lower regret than any of the policies considered in this manuscript for all distributions except the heavy-tailed Student's t-distribution, for which batched policies perform significantly better. Indeed, the UCB2 is calibrated for sub-Gaussian rewards, as are batched policies. However, even with heavy-tailed distributions, the central limit theorem implies that batching a large number of observations returns averages that are sub-Gaussian; see the supplementary material [29]. Even when UCB2 performes better, 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 with only 5 batches performs better than UCB2. Throughout this range, UCB2 uses roughy 50 batches.

It is worth noting that in medical trials, there is nothing special about waiting six months for data from an intervention. Trials of cancer drugs often measure variables like the 1- or 3-year survival rate, or the increase in average survival compared to 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.

APPENDIX A: TOOLS FOR LOWER BOUNDS

Our results hinge on tools for lower bounds, recently adapted to the bandit setting in [9]. Specifically, we reduce the problem of deciding which arm to pull to that of hypothesis testing. Consider the following two candidate setups for the rewards distributions: $P_1 = \mathcal{N}(\Delta, 1) \otimes \mathcal{N}(0, 1)$ and $P_2 = \mathcal{N}(0, 1) \otimes \mathcal{N}(\Delta, 1)$, that is, 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, so arm i is optimal under P_i.

At a given time $t \in [T]$, the choice of $\pi_t \in \{1, 2\}$ 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 $R(t, \pi) = \Delta \mathbb{1}(\pi_t \neq i)$. Denote by E_i^t the expectation under P_i^t, so that

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

Next, we use the following lemma (see [35], 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*,

$$P_1(A) + P_2(A^c) \geq \tfrac{1}{2}\exp\bigl(-\mathsf{KL}(P_1, P_2)\bigr),$$

where $\mathsf{KL}(\cdot, \cdot)$ is the Kullback–Leibler divergence defined by

$$\mathsf{KL}(P_1, P_2) = \int \log\biggl(\frac{dP_1}{dP_2}\biggr)\, dP_1.$$

Here, observations are generated by an M-batch policy π. Recall that $J(t) \in [M]$ denotes the index of the current batch. As π depends 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 straightforward to show that whatever arms are observed over the history, $\mathsf{KL}(P_1^t, P_2^t) = t_{J(t)-1}\Delta^2/2$. Therefore,

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

Summing over t yields the following result.

PROPOSITION A.1. *Fix $\mathcal{T} = \{t_1, \ldots, t_M\}$ and let $(\mathcal{T}, \pi)$ be an M-batch policy. There exist reward distributions with gap Δ, such that $(\mathcal{T}, \pi)$ has regret bounded below as*, *defining $t_0 := 0$*,

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

A variety of lower bounds in Section 5 are shown using this proposition.

APPENDIX B: TECHNICAL LEMMAS

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

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

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

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

PROOF. The first inequality follows from a classical Chernoff bound. To prove the maximal inequality, define $\varepsilon_t = 2\sqrt{\frac{2}{t}\log(\frac{4}{\delta}\frac{\tau}{t})}$. 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 [19], Theorem 3.2, page 314, that for every $\eta > 0$ and every integer $t \geq 1$,

$$\mathbb{P}\{\exists s \leq t, s\bar{Z}_s \geq \eta\} = \mathbb{P}\{\exists s \leq t, e^{\alpha s\bar{Z}_s} \geq e^{\alpha\eta}\} \leq \mathbb{E}[e^{\alpha t\bar{Z}_t}]e^{-\alpha\eta}.$$

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

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

Next, using a peeling argument, one obtains

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

Hence, the result. □

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

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

PROOF. Fix the value of a and observe that $M \leq \log T$ implies that $a \geq e$. Define $x := T\Delta^2/32 > 0$ and $\theta := a^{M-1}/T > 0$. The first inequality is rewritten as

$$xe^{-\theta x} \leq a\,\overline{\log}(x). \tag{7}$$

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

Equation (7) immediately holds if $x \leq e$ as $a\,\overline{\log}(x) = a \geq e$. Similarly, $xe^{-\theta x} \leq 1/(\theta e)$. Thus (7) holds for all $x \geq 1/\sqrt{\theta}$ when $a \geq a^* := 1/(\theta\,\overline{\log}(1/\theta))$. We assume this inequality holds. Thus, we must show that (7) 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 smaller than 1. As a consequence, (7) holds for every $x \leq a$, in particular for every $x \leq a^*$. To summarize, whenever

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

equation (7) holds on $(0, e]$, on $[e, a^*]$ and on $[1/\sqrt{\theta}, +\infty)$, thus on $(0, +\infty)$ as $a^* \geq 1/\sqrt{\theta}$. Next, if $a^M \geq MT/\log T$, we obtain

$$\begin{aligned}\frac{a}{a^*} &= \frac{a^M}{T}\,\overline{\log}\biggl(\frac{T}{a^{M-1}}\biggr) \geq \frac{M}{\log(T)} \log\biggl(T\biggl(\frac{\log T}{MT}\biggr)^{(M-1)/M}\biggr)\\ &= \frac{1}{\log(T)} \log\biggl(T\biggl(\frac{\log(T)}{M}\biggr)^{M-1}\biggr).\end{aligned}$$

The result follows from $\log(T)/M \geq 1$, hence $a/a^* \geq 1$. □

LEMMA B.3. *Fix $a \geq 1, b \geq e$ and let $u_1, u_2, \dots$ 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*

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

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

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

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

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

PROOF. Define $z_k = \log(b/a^{S_k})$. It is straightforward 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

$$u_{k+1} = a\sqrt{\frac{u_k}{\log(b/u_k)}} \geq a\sqrt{\frac{a^{S_{k-1}}}{15 z_{k-2}^{S_{k-2}/2} \log(b/u_k)}}. \tag{8}$$

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

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

This yields

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

Plugging this bound into (8) completes the proof for $k \in [M-3]$.

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

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

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

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

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

$$a := (2T)^{1/S_{M-1}} \log^{1/4-(3/4)1/(2^M-1)}\bigl((2T)^{15/(2^M-1)}\bigr)$$

ensures that

$$a^{S_{M-1}} \geq \sqrt{\frac{15}{16e}} T \log^{S_{M-3}/4}\biggl(\frac{2T}{a^{S_{M-5}}}\biggr) \tag{9}$$

and

$$15 a^{S_{M-2}} \leq 2T. \tag{10}$$

PROOF. Immediate for $M = 2$. For $M > 2$, $2^M \leq \log(4T)$ implies

$$a^{S_{M-1}} = 2T \log^{S_{M-3}/4}\bigl((2T)^{15/(2^M-1)}\bigr) \geq 2T\biggl[16 \frac{15}{2^M - 1} \log(2T)\biggr]^{1/4} \geq 2T.$$

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

$$a^{S_{M-1}} = 2T \log^{S_{M-3}/4}\bigl((2T)^{1-S_{M-5}/S_{M-1}}\bigr) \geq \sqrt{\frac{15}{16e}} T \log^{S_{M-3}/4}\biggl(\frac{2T}{a^{S_{M-5}}}\biggr).$$

This completes the proof of (9). Equation (10) follows if

$$15^{S_{M-1}} (2T)^{S_{M-2}} \log^{(S_{M-3} S_{M-2})/4}\bigl((2T)^{15/(2^M-1)}\bigr) \leq (2T)^{S_{M-1}}. \tag{11}$$

Using that $S_{M-k} \leq 2$, we get that the left-hand side of (10) is smaller than

$$15^2 \log\bigl((2T)^{15/(2^M-1)}\bigr) \leq 2250 \log\bigl((2T)^{2^{1-M}}\bigr).$$

The result follows using $2^M \leq \log(2T)/6$, which implies that the right-hand side in the above inequality is bounded by $(2T)^{2^{1-M}}$. $\square$

SUPPLEMENTARY MATERIAL

Supplement to "Batched bandit problems" (DOI: 10.1214/15-AOS1381SUPP; .pdf). The supplementary material [29] contains additional simulations, including some using real data.

REFERENCES

[1] AUDIBERT, J.-Y. and BUBECK, S. (2010). Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.* **11** 2785–2836. MR2738783

[2] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

[3] AUER, P. and ORTNER, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hungar.* **61** 55–65. MR2728432

[4] BARTROFF, J. (2007). Asymptotically optimal multistage tests of simple hypotheses. *Ann. Statist.* **35** 2075–2105. MR2363964

[5] BARTROFF, J., LAI, T. L. and SHIH, M.-C. (2013). *Sequential Experimentation in Clinical Trials*: *Design and Analysis*. Springer, New York. MR2987767

[6] BATHER, J. A. (1981). Randomized allocation of treatments in sequential experiments. *J. Roy. Statist. Soc. Ser. B* **43** 265–292. MR0637940

[7] BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems*: *Sequential Allocation of Experiments*. Chapman & Hall, London. MR0813698

[8] BERTSIMAS, D. and MERSEREAU, A. J. (2007). A learning approach for interactive marketing to a customer segment. *Oper. Res.* **55** 1120–1135. MR2372281

[9] BUBECK, S., PERCHET, V. and RIGOLLET, P. (2013). Bounded regret in stochastic multi-armed bandits. *COLT* 2013, *JMLR W&CP* **30** 122–134.

[10] CAPPÉ, O., GARIVIER, A., MAILLARD, O.-A., MUNOS, R. and STOLTZ, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* **41** 1516–1541. MR3113820

[11] CESA-BIANCHI, N., DEKEL, O. and SHAMIR, O. (2013). Online learning with switching costs and other adaptive adversaries. *Adv. Neural Inf. Process. Syst.* **26** 1160–1168.

[12] CESA-BIANCHI, N., GENTILE, C. and MANSOUR, Y. (2012). Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* 1190–1204. SIAM, Philadelphia, PA. MR3202975

[13] CHENG, Y. (1996). Multistage bandit problems. *J. Statist. Plann. Inference* **53** 153–170. MR1412085

[14] CHICK, S. E. and GANS, N. (2009). Economic analysis of simulation selection problems. *Manage. Sci.* **55** 421–437.

[15] COLTON, T. (1963). A model for selecting one of two medical treatments. *J. Amer. Statist. Assoc.* **58** 388–400. MR0149637

[16] COLTON, T. (1965). A two-stage model for selecting one of two treatments. *Biometrics* **21** 169–180.

[17] COTTLE, R., JOHNSON, E. and WETS, R. (2007). George B. Dantzig (1914–2005). *Notices Amer. Math. Soc.* **54** 344–362. MR2292141

[18] DANTZIG, G. B. (1940). On the non-existence of tests of "Student's" hypothesis having power functions independent of σ. *Ann. Math. Statist.* **11** 186–192. MR0002082

[19] DOOB, J. L. (1990). *Stochastic Processes*. Wiley, New York. MR1038526

[20] FABIUS, J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41** 1906–1916. MR0278454

[21] GHURYE, S. G. and ROBBINS, H. (1954). Two-stage procedures for estimating the difference between means. *Biometrika* **41** 146–152. MR0062394

[22] HARDWICK, J. and STOUT, Q. F. (2002). Optimal few-stage designs. *J. Statist. Plann. Inference* **104** 121–145. MR1900522

[23] JENNISON, C. and TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL. MR1710781

[24] JUN, T. (2004). A survey on the bandit problem with switching costs. *Economist* **152** 513–541.

[25] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826

[26] MAURICE, R. J. (1957). A minimax procedure for choosing between two populations using sequential sampling. *J. R. Stat. Soc., B* **19** 255–261.

[27] METSCH, L. R., FEASTER, D. J., GOODEN, L. et al. (2013). Effect of risk-reduction counseling with rapid HIV testing on risk of acquiring sexually transmitted infections: The AWARE randomized clinical trial. *JAMA* **310** 1701–1710.

[28] PERCHET, V. and RIGOLLET, P. (2013). The multi-armed bandit problem with covariates. *Ann. Statist.* **41** 693–721. MR3099118

[29] PERCHET, V., RIGOLLET, P., CHASSANG, S. and SNOWBERG, E. (2015). Supplement to "Batched bandit problems." DOI:10.1214/15-AOS1381SUPP.

[30] ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246

[31] SCHWARTZ, E. M., BRADLOW, E. and FADER, P. (2013). Customer acquisition via display advertising using multi-armed bandit experiments. Technical report, Univ. Michigan.

[32] SOMERVILLE, P. N. (1954). Some problems of optimum sampling. *Biometrika* **41** 420–429. MR0066607

[33] STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16** 243–258. MR0013885

[34] 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.

[35] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[36] VOGEL, W. (1960). An asymptotic minimax theorem for the two armed bandit problem. *Ann. Math. Statist.* **31** 444–451. MR0116443

[37] VOGEL, W. (1960). A sequential design for the two armed bandit. *Ann. Math. Statist.* **31** 430–443. MR0116442

V. PERCHET
LPMA, UMR 7599
UNIVERSITÉ PARIS DIDEROT
8, PLACE FM/13
75013, PARIS
FRANCE
E-MAIL: vianney.perchet@normalesup.org

P. RIGOLLET
DEPARTMENT OF MATHEMATICS AND IDSS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MASSACHUSETTS 02139-4307
USA
E-MAIL: rigollet@math.mit.edu

S. CHASSANG
DEPARTMENT OF ECONOMICS
PRINCETON UNIVERSITY
BENDHEIM HALL 316
PRINCETON, NEW JERSEY 08544-1021
USA
E-MAIL: chassang@princeton.edu

E. SNOWBERG
DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
MC 228-77
PASADENA, CALIFORNIA 91125
USA
E-MAIL: snowberg@caltech.edu