From Finite to Countable-Armed Bandits

Anand Kalvit[1] **and Assaf Zeevi**[2]
Graduate School of Business
Columbia University
New York, USA
{[1]akalvit22,[2]assaf}@gsb.columbia.edu

Abstract

We consider a stochastic bandit problem with countably many arms that belong to a finite set of types, each characterized by a unique mean reward. In addition, there is a fixed distribution over types which sets the proportion of each type in the population of arms. The decision maker is oblivious to the type of any arm and to the aforementioned distribution over types, but perfectly knows the total number of types occurring in the population of arms. We propose a fully adaptive online learning algorithm that achieves $\mathcal{O}(\log n)$ distribution-dependent expected cumulative regret after any number of plays n, and show that this order of regret is best possible. The analysis of our algorithm relies on newly discovered concentration and convergence properties of optimism-based policies like UCB in finite-armed bandit problems with *zero gap*, which may be of independent interest.

1 Introduction

Background and motivation. The multi-armed bandit (MAB) problem is a widely studied machine learning paradigm that captures the tension between *exploration* and *exploitation* in online decision making. The problem traces its roots to 1933 when it was first studied in the context of clinical trials in [21]. It has since evolved and numerous variants of the MAB problem have seen an upsurge in applications across a plethora of domains spanning dynamic pricing, online auctions, packet routing, scheduling, e-commerce and matching markets to name a few (see [12] for a comprehensive survey). In its simplest formulation, the decision maker must sequentially play an arm at each time instant out of a set of K possible arms, each characterized by its own distribution of rewards. The objective is to maximize cumulative expected payoffs over the horizon of play. Every play of an arm results in an independent sample from its reward distribution. The decision maker, oblivious to the statistical properties of the arms, must balance exploring new arms and exploiting the best arm played thus far. The objective of maximizing cumulative rewards is often converted to minimizing *regret* relative to an oracle with perfect ex ante knowledge of the best arm. The seminal work [20] was the first to show that the optimal order of this regret is asymptotically logarithmic in the number of plays. Much of the focus since has been on the design and analysis of algorithms that can achieve near-optimal regret rates (see [5, 16, 15], etc., and references therein).

Many practical applications of the multi-armed bandit problem involve a prohibitively large number of arms, the number in some cases is even larger than the horizon of play itself. This renders finite-armed models unsuitable vehicle for the study of such settings. The simplest prototypical example of such a setting occurs in the context of online assignment problems arising in large marketplaces serving a very large population of agents that each belong to one of K possible types; e.g., if $K = 2$, the set of agent types could be {“high caliber”, “low caliber”}, {“patient”, “impatient”}, etc. Such finite-typed settings are also relevant in many applications with an exponentially large choice space and where a limited planning horizon forbids exploration-exploitation in the traditional sense (This is common in online retail where assortments of substitutable products are selected from a very

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

large product space, cf. [2]). We shall refer to problems of this nature as *countable-armed bandits (CAB)*. The CAB problem lies hedged between the finite-armed bandit problem on one end, and the so called *infinite-armed bandit problem* on the other. As the name suggests, the latter is typically characterized by a continuum of arm types and for this reason, the CAB problem is closer in spirit to the finite-armed problem despite an infinity of arms, though it has its own unique salient features.

The CAB problem is characterized by a finite set of arm types $\mathcal{T}$ and a distribution over $\mathcal{T}$ denoted by $\mathcal{D}(\mathcal{T})$. For simplicity of exposition, we assume in this paper that $|\mathcal{T}| = 2$ and state all our propositions under this assumption. This is without loss of generality and all our algorithms and theoretical guarantees readily extend to any finite cardinality $\mathcal{T}$ (see Appendix G). The statistical complexity of the CAB problem with a binary $\mathcal{T}$ is determined by three primitives: (i) the sub-optimality gap (Δ) between the mean of the superior and inferior arm types; (ii) the proportion of arms of the superior type in the infinite population of arms (α); and (iii) the duration of play (n).

Main contributions. We show that the finite-time expected cumulative regret achievable in the CAB problem, absent ex ante knowledge of (Δ, α, n), is $\mathcal{O}\left(\beta^{-1}\left(\Delta^{-1}\log n + \alpha^{-1}\Delta\right)\right)$ (Theorem 3), where $\beta \leqslant 1$ is an instance-specific constant that depends on the reward distributions associated with the arm types alone, and the big-Oh notation only hides absolute constants. To this end, we propose a fully adaptive online learning algorithm that has the aforementioned regret guarantee and show that its performance cannot essentially be improved upon. The proof of Theorem 3 relies on a newly discovered concentration property of optimism-based algorithms such as UCB in finite-armed bandit problems with *zero gap*, e.g., a two-armed bandit with $\Delta = 0$ (Theorem 4 (i)). This result is of independent interest as it disproves a folk conjecture on non-convergence of UCB in zero gap settings (Theorem 4 (ii)) and is likely to have implications for statistical inference problems involving adaptive data collected by UCB-like algorithms. Additionally, the zero gap setting also highlights a stark difference between the limiting pathwise behavior of UCB and Thompson Sampling. In particular, we observe empirically that UCB's concentration and convergence properties à la Theorem 4 are, in fact, violated by Thompson Sampling (Figure 2). A theoretical explanation for said pathological behavior of Thompson Sampling is presently lacking in literature. Before describing the CAB model formally, we survey two closely related MAB models below and note key differences with our model.

Relation to the finite-armed bandit model. In this problem, finiteness of the action set (set of arms) allows for sufficient exploration of all the arms which makes it possible to design policies that achieve near-optimal regret rates (cf. [5, 15], etc.) relative to the lower bound in [20]. In contrast, exploring every single arm in our problem is: (a) infeasible due to an infinity of available arms; and (b) clearly sub-optimal since any attempt at it would result in linear regret. The fundamental difficulty in the countable-armed problem lies in identifying a consideration set that contains at least one arm of the optimal type. In the absence of any ex ante information on (Δ, α), it is unclear whether this can be done in a manner that would guarantee sub-linear regret; and secondly, what is the minimal achievable regret. These questions capture the essence of our work in this paper.

Relation to the infinite-armed bandit model. This problem also considers an infinite population of arms and a fixed *reservoir* distribution over the set of arm types, which maps to the set of possible mean rewards. However, unlike our problem, the set of arm types here forms the continuum $[0, 1]$. The infinite-armed problem traces its roots to [7] where it was first studied under a Bernoulli reward setting with the reservoir distribution of mean rewards being Uniform on $[0, 1]$. This work spawned a rich literature on infinite-armed problems, however, to the best of our knowledge, all of the extant body of work is predicated on the assumption that the reservoir distribution satisfies a certain regularity property (or a variant thereof) in the neighborhood of the optimal mean reward (cf. [7, 22, 9, 13, 11] for a comprehensive survey). Such assumptions restrict the set of types to infinite cardinality sets. In terms of statistical complexity, this has the implication that the minimal achievable regret is polynomial in the number of plays. In contrast, the CAB model is fundamentally simpler since the set of arm types is only finite. The natural question then is if better regret rates are possible for the CAB problem at least on "well-separated" instances. This is the central question underlying our work.

In addition to the infinite-armed bandit model discussed above, there are two other related problem classes: *continuum-armed bandits* and *online stochastic optimization*. However, these problems are predicated on an entirely different set of assumptions involving the topological embedding of the arms and regularities of the mean-reward function, and share little similarity with our stochastic model. The reader is advised to refer to [17, 1, 19, 6, 18, 10], etc., for a detailed coverage of the aforementioned problem classes.

Organization of the paper. The CAB problem is formally described in § 2. Algorithms for the CAB problem and related theoretical guarantees are stated in § 3. A formal statement of the concentration and convergence properties of UCB in finite-armed bandits with zero gap is deferred to § 4. Proof sketches are included in the main text to the extent permissible, full proofs and other technical details including ancillary lemmas are relegated to the appendices.

2 Problem formulation

The set of arm types is denoted by $\mathcal{T} = \{1, 2\}$. Each type $i \in \mathcal{T}$ is characterized by a *unique* mean reward $\mu_i \in (0, 1)$ with the rewards themselves bounded in $[0, 1]$. The proportion of arms of type $\arg\max_{i\in\mathcal{T}} \mu_i$ in the population of arms is given by α. Different arms of the same type may have distinct reward distributions but their mean rewards are equal. For each $i \in \mathcal{T}$, $\mathcal{G}(\mu_i)$ denotes a finite[1] collection of reward distributions with mean μ_i associated with the type i sub-population.

Assumption 1 (Maximally supported rewards in $[0, 1]$**)** *Any CDF* $F \in \cup_{i\in\mathcal{T}}\mathcal{G}(\mu_i)$ *satisfies: (i)* $\sup\{x \in \mathbb{R} : F(x) = 0\} = 0$*, and (ii)* $\inf\{x \in \mathbb{R} : F(x) = 1\} = 1$.[2]

For example, distributions such as Bernoulli(0.1), Beta$(2, 3)$, Uniform on $[0, 1]$, etc., satisfy Assumption 1. Without loss of generality, we assume $\mu_1 > \mu_2$ and call type 1, the optimal type. $\Delta := \mu_1 - \mu_2$ denotes the separation (or gap) between the types. The index set $\mathcal{I}_n$ contains labels of all the arms that have been played up to and including time n (with $\mathcal{I}_0 := \phi$). The set of available actions at time n is given by $\mathcal{A}_n = \mathcal{I}_{n-1} \cup \{\text{new}\}$ and $\mathcal{P}(\mathcal{A}_n)$ denotes the probability simplex on $\mathcal{A}_n$. At any time n, the decision maker must either choose to play an arm from $\mathcal{I}_{n-1}$, or select the action "new" which corresponds to playing a new arm, unexplored hitherto, whose type is an unobserved, independent sample from an unknown distribution on $\mathcal{T}$ denoted by $\mathcal{D}(\mathcal{T}) = (\alpha, 1 - \alpha)$. The realized rewards are independent across arms and i.i.d. in time keeping the arm fixed. The natural filtration $\mathcal{F}_n$ is defined w.r.t. the sequence of rewards realized up to and including time n (with $\mathcal{F}_0 := \phi$). A policy $\pi = \{\pi_n : n \in \mathbb{N}\}$ is a non-anticipatory adaptive sequence that for each n prescribes an action from $\mathcal{P}(\mathcal{A}_n)$, i.e., $\pi_n : \mathcal{F}_{n-1} \to \mathcal{P}(\mathcal{A}_n)\ \forall\ n \in \mathbb{N}$. The cumulative pseudo-regret of π after n plays is given by $R_n^\pi = \sum_{m=1}^{n} \left(\mu_1 - \mu_{t(\pi_m)}\right)$, where $t(\pi_m)$ denotes the type of the arm played by π at time m. We are interested in the problem $\min_{\pi\in\Pi} \mathbb{E}R_n^\pi$, where n is the horizon of play, Π is the set of all non-anticipation policies, and the expectation is w.r.t. the randomness in π as well as $\mathcal{D}(\mathcal{T})$. We remark that $\mathbb{E}R_n^\pi$ is the same as the traditional notion of expected cumulative regret in our problem[3].

Other notation. We reemphasize that for any given arm, *label* and *type* are two distinct attributes. The number of plays up to and including time n of arm i is denoted by $N_i(n)$, and its type by $t(i) \in \mathcal{T}$. At any time n^+, $(X_{i,j})_{j=1}^{m}$ denotes the sequence of rewards realized from the first $m \leqslant N_i(n)$ plays of arm i. The natural filtration at time n^+ is formally defined as $\mathcal{F}_n := \sigma\left\{(X_{i,j})_{j=1}^{N_i(n)}; i \in \mathcal{I}_n\right\}$. The empirical mean reward from the first $N_i(n)$ plays of arm i is denoted by $\overline{X}_i(n)$. An absolute constant is understood to be one that does not depend on any problem primitive or free parameters.

3 Main results: Rate-optimal algorithms for the CAB problem

In the finite-armed bandit problem, the gap Δ is the key primitive that determines the statistical complexity of regret minimization. The literature on finite-armed bandits roughly bifurcates into two broad strands of algorithms, Δ-*aware* and Δ-*agnostic*. Explore-then-Commit (aka, Explore-then-Exploit) and ϵ_n-Greedy are two prototypical examples of the former category, while UCB and Thompson Sampling belong to the latter. In the CAB problem too, Δ plays a key role in determining the complexity of regret minimization. Since this is the first theoretical treatment of the subject matter, it is instructive to first study the Δ-aware case to gain insight into the basic premise that sets the finite and countable-armed problems apart. We investigate the case of a Δ-aware decision maker in § 3.1 and the Δ-agnostic case in § 3.2. Before proceeding to the algorithms, we first state a lower

[1] This is simply to keep the analysis simple and has no bearing on the regret guarantees of our algorithms.

[2] Define $\lambda(F_i, F_j) := \max_{(k,l)\in\{(i,j),(j,i)\}} \left(\inf\{x \in \mathbb{R} : F_k(x) = 1\} - \sup\{x \in \mathbb{R} : F_l(x) = 0\}\right)$ for arbitrary CDFs F_i, F_j. Then, we require prior knowledge of $\min_{i,j\in\mathcal{T}, i\neq j} \min_{F_i\in\mathcal{G}(\mu_i), F_j\in\mathcal{G}(\mu_j)} \lambda(F_i, F_j)$. Assumption 1 fixes $\lambda = 1$.

[3] Expected cumulative regret equals the expected cumulative pseudo-regret in the stochastic bandits setting.

bound for the CAB problem that applies for any admissible policy. In what follows, an *instance* of the CAB problem refers to the tuple $(\mathcal{G}(\mu_1), \mathcal{G}(\mu_2))$ with $|\mu_1 - \mu_2| = \Delta$, and we slightly overload the notation for expected cumulative regret to emphasize its instance-dependence.

Theorem 1 (Lower bound on achievable performance) *For any $\Delta > 0$, $\exists$ a pair of reward distributions (Q_1, Q_2) with means (μ_1, μ_2) respectively, satisfying $|\mu_1 - \mu_2| = \Delta$, and an absolute constant C, s.t. the expected cumulative regret of any asymptotically consistent*[4] *policy π on the CAB instance $\nu = (\{Q_1\}, \{Q_2\})$ satisfies for all $\alpha \leqslant 1/2$ and n large enough, $\mathbb{E}R_n^\pi(\nu) \geqslant C\Delta^{-1} \log n$.*

Remark. Theorem 1 bears resemblance to the classical lower bound of Lai and Robbins for finite-armed bandits [20], but the two results differ in a fundamental way. While $\nu = (\{Q_1\}, \{Q_2\})$ fully specifies a two-armed bandit problem, it is the *realization* of ν, i.e., an infinite sequence $(r_i)_{i\in\mathbb{N}}$ with $\mathbb{P}\left(r_i = Q_{\arg\max_{j\in\{1,2\}} \mu_j}\right) = \alpha$ and where $r_i \in \{Q_1, Q_2\}$ indicates the reward distribution of arm $i \in \mathbb{N}$, that specifies the CAB problem. As such, traditional lower bound proofs for finite-armed bandits are not directly adaptable to the CAB problem. Nonetheless, the two results retain structural similarities because the CAB problem, despite its additional complexity, remains amenable to a standard reduction to a hypothesis testing problem. It must be noted that any policy incurs linear regret when $\alpha = 0$, while zero regret when $\alpha = 1$. Theorem 1 states a uniform lower bound independent of α that applies for all $\alpha \leqslant 1/2$. Since the CAB problem with $\alpha < 1/2$ is statistically harder than its two-armed counterpart, we believe the lower bound in Theorem 1 is in fact, unachievable in the sense of the exact scaling of the $\log n$ term. However, our objective in this paper is to develop algorithms for the CAB problem that are order-optimal in n and to that end, Theorem 1 serves its stipulated purpose. Characterizing an *achievable* scaling of the lower bound and its dependence on $\alpha \in [0, 1]$ remains an open problem. We consider the restriction to the classical asymptotically consistent policy class (Definition 1, Appendix A) as more generic policy classes are unwieldy for lower bound proofs due to reasons stemming from the combinatorial nature of our problem. Full proof is given in Appendix A.

3.1 A near-optimal Δ-aware algorithm for the CAB problem

The intuition and understanding developed through this section shall be useful while studying the Δ-agnostic case later and highlights key statistical features of the CAB problem. Below, we present a simple fixed-design ETC (Explore-then-Commit) algorithm assuming ex ante knowledge of the duration of play[5] n and a separability parameter $\delta \in (0, \Delta]$. In what follows, we use *select* to indicate an arm selection action, and *play* to indicate the action of pulling a selected arm. A reward is only realized after an arm is played, not merely selected. A *new* arm refers to one that has never been selected before. $(X_{i,j})_{j=1}^m$ denotes the sequence of rewards realized from the first m plays of arm i.

Algorithm 1 ETC-∞(2): ETC for an infinite population of arms with $|\mathcal{T}| = 2$.

```
Input: $(n, \delta)$, where $\delta \in (0, \Delta]$.
Set $L = \lceil 2\delta^{-2} \log n \rceil$. Set budget $T = n$.
Initialization (Starts a new epoch): Select two new arms. Call it consideration set $\mathcal{A} = \{1, 2\}$.
$m \leftarrow \min(L, T/2)$.
Play each arm in $\mathcal{A}$ $m$ times. Update budget: $T \leftarrow T - 2m$.
if $\left|\sum_{j=1}^m (X_{1,j} - X_{2,j})\right| < \delta m$ then
    Permanently discard $\mathcal{A}$ and go to Initialization.
else
    Commit the remaining budget of play to arm $i^* \in \arg\max_{i\in\mathcal{A}} \sum_{j=1}^m X_{i,j}$.
```

Mechanics of ETC-∞(2). The horizon of play is divided into epochs of length $2m = \mathcal{O}(\log n)$ each. The algorithm starts off by selecting a pair of arms at random from the infinite population of arms and playing them m times each in the first epoch. Thereafter, the pair is classified as having either identical or distinct types via a hypothesis test through step 6. If classified as "identical," the algorithm permanently discards both the arms (never to be selected again) and replaces them with yet another newly selected pair, which is subsequently played equally in the next epoch. This process is

[4] This is a rich policy class that includes all algorithms achieving sublinear regret (defined in Appendix A).

[5] The standard exponential doubling trick can be employed to make the algorithm horizon-free, cf. [8].

repeated until a pair of arms with distinct types is identified. In the event of such a discovery, the algorithm commits the residual budget to the empirically better arm in the current consideration set.

Theorem 2 (Upper bound on the expected regret of ETC-∞(2)) *The expected cumulative regret of the policy π given by Algorithm 1 after n plays is bounded as follows:*

$$\mathbb{E}R_n^{\pi} \leqslant \min\left(\Delta n,\ \Delta\left(2+\alpha^{-1}\right)\left(2\delta^{-2}\log n+1\right)+\alpha^{-1}\left(f(n,\delta,\Delta)+2\right)\Delta\right),$$

where $f(n,\delta,\Delta)=o(1)$ in n and independent of α (Note: This result is agnostic to Assumption 1.).

Proof sketch of Theorem 2. On a pair of arms of the optimal type (type 1), any playing rule incurs zero regret in expectation, whereas the expected regret is linear in the number of plays if the pair is of the inferior type (type 2). Since it is statistically impossible to distinguish between a type 1 pair and a type 2 pair in the absence of any distributional knowledge of the associated rewards, the algorithm must identify a pair of distinct types whenever so obtained, to avoid high regret. This is precisely done through step 6 of Algorithm 1 via a hypothesis test. Since the distribution over the types, denoted by $\mathcal{D}(\mathcal{T})=(\alpha,1-\alpha)$, is stationary, the number of fresh draws of consideration sets until one with arms of distinct types is obtained is a geometric random variable (say W). Thus, it only takes $(\mathbb{E}W)(2m)=\mathcal{O}(\log n)$ plays in expectation to obtain such a pair and identify it correctly with high probability. The algorithm subsequently commits to the optimal arm in the pair with high probability. Therefore, the overall expected regret is also $\mathcal{O}(\log n)$. Full proof is relegated to Appendix B. □

Remark. The key idea used in Algorithm 1 is that of interleaving hypothesis testing (step 6) with regret minimization (step 9). In the stated version of the algorithm, the regret minimization step simply commits to the arm with the higher empirical mean reward. The framework of Algorithm 1 also allows for other regret minimizing playing rules (for e.g., ϵ_n-Greedy [5], etc.) to be used instead in step 9. The flexibility afforded by this framework shall become apparent in § 3.2.

3.2 A near-optimal Δ-agnostic algorithm for the CAB problem

Designing an adaptive, Δ-agnostic algorithm and the proof that it can achieve the lower bound in Theorem 1 (in n, modulo multiplicative constants) is the main focus of this paper. Recall that ex ante information about Δ serves a dual role in Algorithm 1: (i) in calibrating the epoch length in step 2; and (ii) determining the separation threshold for hypothesis testing in step 6. In the absence of information on Δ, it is a priori unclear if there exists an algorithm that would guarantee sublinear regret on "well-separated" instances. In Algorithm 2 below, we present a generic framework called ALG($\Xi,\Theta,2$), around which various Δ-agnostic playing rules such as UCB, Thompson Sampling, etc., can be tested. In what follows, $s\in\{1,2,...\}$ indicates a discrete time index at which an arm may be played in the current epoch. Every epoch starts from $s=1$.

Algorithm 2 ALG($\Xi,\Theta,2$): An algorithmic framework for countable-armed bandits with $|\mathcal{T}|=2$.

```
 1: Input: A Δ-agnostic playing rule $\Xi$, a deterministic sequence $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ in $\mathbb{R}$.
 2: Initialization (Starts a new epoch): Select two new arms. Call it consideration set $\mathcal{A} = \{1, 2\}$.
 3: For $s \in \{1, 2\}$, play each arm in $\mathcal{A}$ once.
 4: $m \leftarrow 1$.
 5: for $s \in \{3, 4, ...\}$ do
 6:     if $\left|\sum_{j=1}^{m}(X_{1,j} - X_{2,j})\right| < \theta_m$ then
 7:         Permanently discard $\mathcal{A}$ and go to Initialization.
 8:     else
 9:         Play an arm from $\mathcal{A}$ according to $\Xi$.
10:         $m \leftarrow \min_{i\in\mathcal{A}} N_i(s)$.
```

On the issue of sample-adaptivity in hypothesis-testing. The foremost noticeable aspect of Algorithm 2 that also sets it apart from Algorithm 1, is that the samples used for hypothesis testing in step 6 are collected *adaptively* by Ξ. For instance, if $\Xi=$ UCB1 [5], then step 9 translates to playing arm $i^*\in\arg\max_{i\in\mathcal{A}}\left(\overline{X}_i(s-1)+\sqrt{2\log(s-1)/N_i(s-1)}\right)$. This is distinct from the classical hypothesis testing setup used in step 6 of Algorithm 1, where the collected data does not exhibit such dependencies. It is well understood that adaptivity in the sampling process can lead to biased inferences (see, e.g., [14]). However, for standard choices of Ξ such as UCB or Thompson Sampling (or variants thereof), the exploratory nature of Ξ ensures that the test statistic $\sum_{j=1}^{m}(X_{1,j} - X_{2,j})$ where $m = \min_{i\in\mathcal{A}} N_i(s)$, remains agnostic to any sample-adaptivity due to Ξ. This statement is formalized and further explained in Lemma 1 (Appendix F).

Mechanics of ALG$(\Xi, \Theta, 2)$. We call a consideration set $\mathcal{A}$ of arms "heterogeneous" if it contains arms of distinct types, and "homogeneous" otherwise. Algorithm 2 has a master-slave framework in which step 6 is the master routine and Ξ serves as the slave subroutine in step 9. The purpose of step 6 is to quickly determine if $\mathcal{A}$ is homogeneous, in which case it discards $\mathcal{A}$ and restarts the algorithm afresh in a new epoch. On the other hand, whenever a heterogeneous $\mathcal{A}$ gets selected, step 6 ensures that its selection persists in expectation which allows Ξ to run "uninterrupted." This idea is formalized in Lemma 2 (Appendix F). In a nutshell, Algorithm 2 runs in epochs of random lengths that are themselves determined adaptively. At the beginning of every epoch, the algorithm selects a new consideration set $\mathcal{A}$ and deploys Ξ on it. It then determines (via the hypothesis test in step 6) whether to keep playing Ξ on $\mathcal{A}$ or to stop and terminate the epoch, based on the current sample history of $\mathcal{A}$. Upon termination, $\mathcal{A}$ is discarded and the algorithm starts afresh in a new epoch.

Calibrating Θ. ALG$(\Xi, \Theta, 2)$ identifies homogeneous $\mathcal{A}$'s by means of a hypothesis test through step 6. It starts with the null hypothesis $\mathcal{H}_0$ that the current $\mathcal{A}$ is heterogeneous and persists with it until "enough" evidence to the contrary is gathered. If $\mathcal{H}_0$ were indeed true, the Strong Law of Large Numbers (SLLN) would dictate that $\left|\sum_{j=1}^{m}(X_{1,j} - X_{2,j})\right| \sim \Delta m$, almost surely. If $\mathcal{H}_0$ were false, it would follow from the Central Limit Theorem (CLT) that $\left|\sum_{j=1}^{m}(X_{1,j} - X_{2,j})\right| = \mathcal{O}\left(\sqrt{m}\right)$. Therefore, in order to separate $\mathcal{H}_0$ from its complement, the right θ_m must satisfy: $\theta_m = o(\Delta m)$ and $\theta_m = \omega\left(\sqrt{m}\right)$. Indeed, our choice of θ_m (see (2)) satisfies these conditions and is such that $\theta_m \sim 2\sqrt{m \log m}$. We reemphasize that the calibration of Θ is independent of Δ and only *informed* by classical results (SLLN, CLT) that are themselves inapplicable since the data collection is adaptive.

High-level overview of results. We show that for a suitably calibrated input sequence Θ (see (2)), the instance-dependent expected cumulative regret of ALG(UCB1, Θ, 2) is logarithmic in the number of plays anytime, this order of regret being best possible. We also demonstrate empirically that a key concentration property of UCB1 that is pivotal to the aforementioned regret guarantee, is violated for Thompson Sampling (TS) and therefore, ALG(TS, Θ, 2) suffers linear regret. A formal statement of said concentration property of UCB1 is deferred to § 4. The regret upper bound of ALG(UCB1, Θ, 2) is stated next in Theorem 3. Following is an auxiliary proposition that is useful towards Theorem 3.

Proposition 1 (Lower bound on the true negative rate) *For each $i \in \mathcal{T} = \{1, 2\}$, let $\left(Y_j^{F_i}\right)_{j\in\mathbb{N}}$ denote an i.i.d. sequence of random variables with distribution $F_i \in \mathcal{G}(\mu_i)$ satisfying Assumption 1. Let $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ be a deterministic non-negative real-valued sequence such that $\{(\theta_m/m) : m = 1, 2, ...\}$ is monotone decreasing in m with $\theta_1 < 1$ and $\theta_m = o(m)$. Then,*

$$\beta := \min_{F_1\in\mathcal{G}(\mu_1), F_2\in\mathcal{G}(\mu_2)} \mathbb{P}\left(\bigcap_{m=1}^{\infty} \left|\sum_{j=1}^{m}\left(Y_j^{F_1} - Y_j^{F_2}\right)\right| \geqslant \theta_m\right) > 0. \tag{1}$$

Proof of Proposition 1. Refer to Appendix C (Note: Assumption 1 plays a key role here.). □

Remark. β is a continuous function of Δ with $\lim_{\Delta\to 0}\beta = 0$. In particular, β depends on Δ and the specific choice of Θ. Proposition 1 implicitly assumes $\Delta > 0$.

Theorem 3 (Upper bound on the expected regret of ALG(UCB1, Θ, 2)) *Consider the input sequence $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ given by*

$$\theta_m := \sqrt{m^2(m + m_0)^{-1}\left(4\log(m + m_0) + \gamma \log\log(m + m_0)\right)}, \tag{2}$$

where $m_0 \geqslant 0$ and $\gamma > 2$ are user-defined parameters that ensure Θ satisfies the conditions of Proposition 1 (for example, $m_0 = 11$ and $\gamma = 2.1$ is an acceptable configuration). Suppose that Assumption 1 is satisfied. Then, the expected cumulative regret of π = ALG (UCB1, Θ, 2) after any number of plays n is bounded as follows:

$$\mathbb{E}R_n^{\pi} \leqslant \min\left(\Delta n,\ 8\left(\beta\Delta\right)^{-1}\log n + \left(C_1 + \alpha^{-1}C_2\right)\beta^{-1}\Delta\right), \tag{3}$$

where β is as defined in (1), $\Delta = \mu_1 - \mu_2 > 0$, *$C_1$ is an absolute constant and C_2 is a constant that depends only on the free parameters of the algorithm, namely (m_0, γ).*

Comparison with the two-armed bandit problem. The expected cumulative regret of $\pi = \mathrm{UCB1}$ [5] after any number of plays n in a two-armed bandit problem with gap Δ is bounded as follows:

$$\mathbb{E}R_n^{\pi} \leqslant \min\left(\Delta n,\ 8\Delta^{-1}\log n + C_1\Delta\right). \tag{4}$$

Observe that the upper bounds in (3) and (4) differ in (α, β, C_2). The presence of the inflation factor β^{-1} in (3) is on account of the samples "wasted" due to false positives (rejecting the null, when it is in fact true) in the CAB problem. Specifically, $1 - \beta$ is an upper bound on the false positive rate of $\mathrm{ALG}(\mathrm{UCB1}, \Theta, 2)$ (Proposition 1). Furthermore, β is invariant w.r.t. the playing rule (UCB1, in this case) as long as it is sufficiently exploratory (This statement is formalized in Lemma 1,2 stated in Appendix F.). In that sense, β captures the added layer of complexity due to the countable-armed extension of the finite-armed problem. We believe this is not merely an artifact of our proof but in fact, reflecting a fundamentally different scaling of the best achievable regret in the CAB problem vis-à-vis its finite-armed counterpart. It is also noteworthy that β is independent of α; the implication is that (3) depends on the proportion of optimal arms only through the constant term, unlike Theorem 2.

Dependence of β on Δ. Since obtaining a closed-form expression for β as a function of Δ (see (1)) is hard, we compute it numerically on different reward configurations using Monte-Carlo simulations.

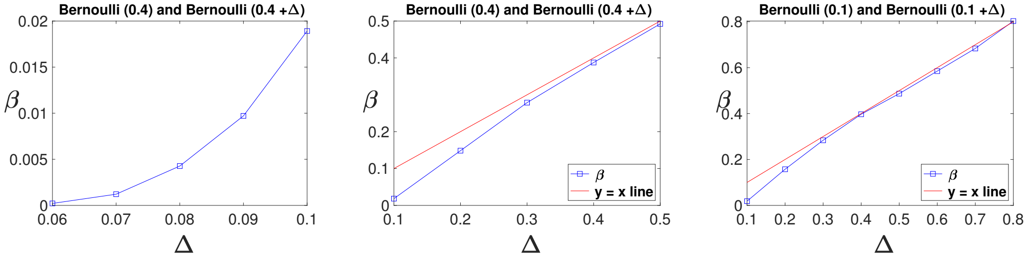

Figure 1: β vs. Δ: Monte-Carlo estimates of β plotted against Δ using (2) with $m_0 = 4000$ and $\gamma = 2.1$. Rewards associated with each type $i \in \mathcal{T}$ are modeled as $\mathrm{Bernoulli}(\mu_i)$.

An immediate observation from Figure 1 is that β scales approximately linearly with Δ when it is sufficiently large (see center and rightmost plots). This has the implication that the upper bound of Theorem 3 scales approximately as $\mathcal{O}\left(\Delta^{-2}\log n\right)$ on well-separated instances, which can be contrasted with the classical $\mathcal{O}\left(\Delta^{-1}\log n\right)$ scaling achievable in finite-armed problems. The extra Δ^{-1} term is reflective of the additional complexity of the CAB problem vis-à-vis the finite-armed problem. In addition, for small Δ (see leftmost plot), β seems to vanish very fast as $\Delta \to 0$. This suggests that the minimax regret of $\mathrm{ALG}(\mathrm{UCB1}, \Theta, 2)$ is orders of magnitude larger (in n) than $\mathcal{O}\left(\sqrt{n\log n}\right)$, which is UCB1's minimax regret in finite-armed problems. We conjecture that the minimax lower bound for the CAB problem is itself orders of magnitude larger than $\Omega\left(\sqrt{n}\right)$. Of course, characterizing the minimax statistical complexity of the CAB model remains an open problem.

Significance of UCB1's concentration in zero gap. That C_2 (appearing in (3)) is a constant is a highly non-trivial consequence of the concentration property of UCB1 à la part (i) of Theorem 4 stated in § 4. In the absence of this property, C_2 would scale with the horizon of play linearly and $\mathrm{ALG}(\mathrm{UCB1}, \Theta, 2)$ would effectively suffer linear regret. In what follows, we will demonstrate empirically that *Thompson Sampling most likely does not enjoy this concentration property*. To the best of our knowledge, this is the first example illustrating such a drastic performance disparity between algorithms based on UCB and Thompson Sampling in any stochastic bandit problem.

Proof sketch of Theorem 3. On homogeneous $\mathcal{A}$'s with arms of the optimal type (type 1), any playing rule incurs zero regret in expectation, whereas the expected regret is linear on homogeneous $\mathcal{A}$'s of type 2. On heterogeneous $\mathcal{A}$'s, the expected regret of UCB1 is logarithmic in the number of plays anytime. Since it is statistically impossible to distinguish between homogeneous $\mathcal{A}$'s of type 1 and type 2 in the absence of any distributional knowledge of the associated rewards, the decision maker must allocate all of her sampling effort (in expectation) to heterogeneous $\mathcal{A}$'s, to avoid high regret. This would ensure that UCB1 runs "uninterrupted" (in expectation) over the duration of play, thereby guaranteeing logarithmic regret. This argument precisely forms the backbone of our proof. The number of re-initializations of the algorithm needed for a heterogeneous $\mathcal{A}$ to get selected is a geometric random variable and furthermore, every time a homogeneous $\mathcal{A}$ gets selected, the algorithm re-initializes within a finite number of plays in expectation. Therefore, only finitely many plays (in expectation) are spent on homogeneous $\mathcal{A}$'s until a heterogeneous $\mathcal{A}$ gets selected. Subsequently, the algorithm (in expectation) allocates the residual sampling effort to $\mathcal{A}$ which allows UCB1 to run uninterrupted, thereby guaranteeing logarithmic regret. Full proof is relegated to Appendix D. □

Miscellaneous remarks. (i) Comparison with the state-of-the-art. The regret incurred by suitable adaptations of known algorithms for infinite-armed bandits, e.g., [22], etc., is provably worse by at least poly-logarithmic factors compared to the optimal $\mathcal{O}\left(\log n\right)$ rate achievable in the CAB problem. **(ii) Alternatives to UCB1 in ALG(UCB1, $\Theta, 2$).** The choice of UCB1 is entirely a consequence of our desire to keep the analysis simple, and does not preclude use of suitable alternatives satisfying a concentration property akin to part (i) of Theorem 4. **(iii) Improving sample-efficiency.** ALG(UCB1, $\Theta, 2$) indulges in wasteful exploration since it selects an entirely new consideration set of arms at the beginning of every epoch. This is done for the simplicity of analysis. Sample-efficiency can be improved by discarding only one arm at the end of an epoch and selecting only one new arm at the beginning of the next. Furthermore, sample history of the arm retained from the previous epoch can also be used in subsequent hypothesis testing iterations for faster identification of homogeneous consideration sets without forcing unnecessary additional plays. **(iv) Limitations.** In this paper, we assume that $|\mathcal{T}|$ is perfectly known to the decision maker. However, it remains unclear if sublinear regret would still be information-theoretically achievable on "well-separated" instances if said assumption is violated, ceteris paribus.

4 UCB1 and the zero gap problem

UCB1 [5] is a celebrated optimism-based algorithm for finite-armed bandits that adapts to the sub-optimality gap (separation) between the top two arms, and guarantees a worst-case regret of $\mathcal{O}\left(\sqrt{n\log n}\right)$ (ignoring dependence on the number of arms). This occurs when the separation scales with the horizon of play as $\mathcal{O}\left(\sqrt{n^{-1}\log n}\right)$. Our interest here, however, is in the scenario where this separation is exactly *zero*, as opposed to simply being vanishingly small in the limit $n \to \infty$. Without loss of generality, we restrict our focus to the special case of a stochastic two-armed bandit with *equal* mean rewards. Regret related questions are irrelevant in this setting since every policy incurs zero regret in expectation. However, questions concerning the asymptotic pathwise behavior under UCB1 and the sampling balance (or imbalance) between the arms in *zero gap*, remain unanswered in extant literature[6] to the best of our knowledge. In this paper, we provide the first analysis in this direction.

Theorem 4 (Concentration of UCB1 in zero gap) *Consider a stochastic two-armed bandit with rewards bounded in $[0,1]$ and arms having equal means. Let $N_i(n)$ denote the number of plays of arm i under UCB1 [5] up to and including time n. Then, the following results hold for any $i \in \{1,2\}$:*

(i) **Concentration.** *For any $n \in \mathbb{N}$, $\epsilon \in (0, 1/2)$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(\left|\frac{N_i(n)}{n} - \frac{1}{2}\right| \geqslant \epsilon\right) \leqslant \left(\frac{8}{\epsilon\delta}\right) n^{-\left(3-4\sqrt{1-4(1-\delta)^2\epsilon^2}\right)}.$$

(ii) **Convergence.** *$N_i(n)/n \to 1/2$ in probability as $n \to \infty$ (Convergence does not follow from concentration alone since the bound in (i) is vacuous for $\epsilon \leqslant \sqrt{7}/8$.).*

Result for generic UCB. Theorem 4 also extends to the generic UCB policy that uses $\sqrt{\rho n^{-1}\log n}$ as the optimistic bias, where $\rho > 1/2$ is called the exploration coefficient ($\rho = 2$ corresponds to UCB1). The concentration bound for said policy (informally called UCB(ρ)) is given by

$$\mathbb{P}\left(\left|\frac{N_i(n)}{n} - \frac{1}{2}\right| \geqslant \epsilon\right) \leqslant \left(\frac{2^{2\rho-1}}{\epsilon\delta}\right) n^{-\left(2\rho-1-2\rho\sqrt{1-4(1-\delta)^2\epsilon^2}\right)}. \tag{5}$$

While the tail progressively gets lighter as ρ increases, it is achieved at the expense of an inflated regret on instances with non-zero gap. Specifically, the authors in [4] showed that the expected regret of

[6] Extant work assumes a positive gap (cf. [4]); the resulting bounds are vacuous in the zero gap regime.

UCB(ρ) on well-separated instances scales as $\mathcal{O}\left(\rho \log n\right)$. They also showed that the tail of UCB(ρ)'s pseudo-regret on well-separated instances is bounded as $\mathbb{P}\left(R_n > z\right) = \mathcal{O}\left(z^{-(2\rho-1)}\right)$ for large enough z, implying a tail decay of $\mathcal{O}\left(z^{-(2\rho-1)}\right)$ for the fraction of *inferior* plays. Alternatively, (5) suggests for the fractional plays of *any* arm, a heavier tail decay of $\mathcal{O}\left(z^{-\left(2\rho-1-2\rho\sqrt{1-4(1-\delta)^2\epsilon^2}\right)}\right)$ in zero gap settings, which accounts for the slow convergence evident in Figure 2 (leftmost plot).

Miscellaneous remark. Theorem 4 (the convergence result in part (ii), in particular) is likely to have implications for inference problems involving adaptive data collected by UCB-inspired algorithms.

Parsing Theorem 4. To build some intuition, we pivot to the case of statistically identical arms. In this case, labels are exchangeable and therefore $\mathbb{E}\left(N_i(n)/n\right) = 1/2$ for $i \in \{1,2\}, n \in \mathbb{N}$. While symmetry between the arms is enough to guarantee convergence in expectation, it does not shed light on the pathwise behavior of UCB1. An immediate corollary of part (i) of Theorem 4 is that for any $\epsilon \in \left(\sqrt{3}/4, 1/2\right)$, there exists $\delta \in (0,1)$ that ensures $\sum_{n\in\mathbb{N}} \mathbb{P}\left(|N_i(n)/n - 1/2| \geqslant \epsilon\right) < \infty$. This implies that the arms are eventually sampled linearly in time, almost surely, at a rate that is at least $\left(1/2 - \sqrt{3}/4\right)$. That this rate cannot be pushed arbitrarily close to $1/2$ is not merely an artifact of our proof but also suggested by the extremely slow convergence of the empirical probability density of $N_1(n)/n$ to the Dirac delta at $1/2$ in Figure 2 (leftmost plot). This slow convergence likely led to the incorrect folk conjecture that optimism-based algorithms such as UCB1 and variants thereof do not converge à la part (ii) of Theorem 4 (e.g., see [14] and references therein). Instead, we believe the weaker conjecture that the convergence is not w.p. 1, is likely true. Full proof is given in Appendix E.

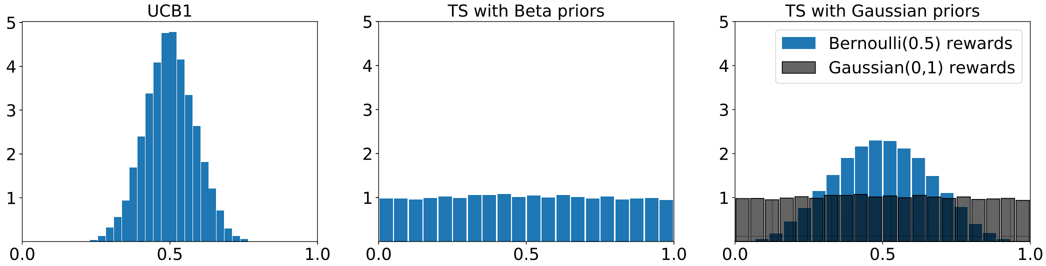

Figure 2: Two-armed bandit with Bernoulli(0.5) rewards: Histogram of the fraction of plays of arm 1 until time $n =$ 10,000 $\left(N_1\left(10^4\right)/10^4\right)$ under three different algorithms. Number of replications under each algorithm $\mathfrak{N} =$ 20,000. The algorithms are: UCB1 (leftmost), Thompson Sampling (TS) with Beta priors (center) and TS with Gaussian priors (rightmost) [3]. The last plot shows histograms for two instances: Bernoulli(0.5) rewards (in blue), and standard Gaussian rewards (dashed).

Empirical illustration. Figure 2 shows the histogram of the fraction of time a particular arm of a two-armed bandit having statistically identical arms with Bernoulli(0.5) rewards each was played under different algorithms. The leftmost plot corresponds to UCB1 and is evidently in consonance with the concentration property stated in part (i) of Theorem 4. The concentration phenomenon under UCB1 can be understood through the lens of reward stochasticity. Consider the simplest case where the rewards are deterministic. Then, we know from the structure of UCB1 that any arm is played at most twice before the algorithm switches over to the other arm. This results in $N_1(n)/n$ converging to $1/2$ pathwise, with an arm switch-over time that is at most 2. As the reward stochasticity increases, so does the arm switch-over time, which adversely affects this convergence. While it is a priori unclear whether $N_1(n)/n$ would still converge to $1/2$ in some mode if the rewards are stochastic, part (ii) of Theorem 4 states that the convergence indeed holds, albeit only in probability. A significant spread around $1/2$ in the leftmost plot despite $n = 10^4$ plays indicates that the convergence is rather slow. This forms the basis of our conjecture that $N_1(n)/n$ does not converge almost surely under optimism-based algorithms like UCB1 if at least one arm has a non-degenerate reward distribution.

A remark on Thompson Sampling. Concentration and convergence à la Theorem 4 should be contrasted with other popular gap-agnostic algorithms such as Thompson Sampling (TS). The center and righmost plots in Figure 2 correspond to TS under different choices of the prior distribution: Beta priors (center) and Gaussian priors (rightmost). These strongly disagree with the leftmost plot corresponding to UCB1. Explaining these zero gap phenomena under TS remains an open problem.

Broader Impact

The authors do not claim any immediate broader impact of this work as such.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous referees for their constructive feedback on the initial version of this paper. The authors also declare an absence of any competing interests.

References

[1] Agrawal, R. The continuum-armed bandit problem. *SIAM journal on control and optimization 33*, 6 (1995), 1926–1951.

[2] Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research 67*, 5 (2019), 1453–1485.

[3] Agrawal, S., and Goyal, N. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM) 64*, 5 (2017), 1–24.

[4] Audibert, J.-Y., Munos, R., and Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science 410*, 19 (2009), 1876–1902.

[5] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 2-3 (2002), 235–256.

[6] Auer, P., Ortner, R., and Szepesvári, C. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory* (2007), Springer, pp. 454–468.

[7] Berry, D. A., Chen, R. W., Zame, A., Heath, D. C., Shepp, L. A., et al. Bandit problems with infinitely many arms. *The Annals of Statistics 25*, 5 (1997), 2103–2116.

[8] Besson, L., and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971* (2018).

[9] Bonald, T., and Proutiere, A. Two-target algorithms for infinite-armed bandits with bernoulli rewards. In *Advances in Neural Information Processing Systems* (2013), pp. 2184–2192.

[10] Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems* (2009), pp. 201–208.

[11] Carpentier, A., and Valko, M. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning* (2015), pp. 1133–1141.

[12] Cesa-Bianchi, N., and Lugosi, G. *Prediction, learning, and games.* Cambridge university press, 2006.

[13] Chan, H. P., and Hu, S. Infinite arms bandit: Optimality via confidence bounds. *arXiv preprint arXiv:1805.11793* (2018).

[14] Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In *International Conference on Machine Learning* (2018), PMLR, pp. 1194–1203.

[15] Garivier, A., and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory* (2011), pp. 359–376.

[16] Garivier, A., Lattimore, T., and Kaufmann, E. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems* (2016), pp. 784–792.

[17] Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207* (2019).

[18] Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing* (2008), ACM, pp. 681–690.

[19] Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems* (2005), pp. 697–704.

[20] Lai, T. L., and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*, 1 (1985), 4–22.

[21] Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25*, 3/4 (1933), 285–294.

[22] Wang, Y., Audibert, J.-Y., and Munos, R. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems* (2009), pp. 1729–1736.

From Finite to Countable-Armed Bandits: Appendix

Anand Kalvit[1] and Assaf Zeevi[2]
Graduate School of Business
Columbia University
New York, USA
{[1]akalvit22,[2]assaf}@gsb.columbia.edu

A Proof of Theorem 1

Since the horizon of play is fixed at n, the decision maker may play at most n distinct arms. Therefore, it suffices to focus only on the sequence of the first n arms that may be played. A *realization* of an instance $\nu = (\mathcal{G}(\mu_1), \mathcal{G}(\mu_2))$ is defined as the n-tuple $r \equiv (r_i)_{1 \leqslant i \leqslant n}$, where $r_i \in \mathcal{G}(\mu_1) \cup \mathcal{G}(\mu_2)$ indicates the reward distribution of arm $i \in \{1, 2, ..., n\}$. It must be noted that the decision maker need not play every arm in r. The distribution over the possible realizations of $\nu = (\mathcal{G}(\mu_1), \mathcal{G}(\mu_2))$ in $\{r : r_i \in \mathcal{G}(\mu_1) \cup \mathcal{G}(\mu_2),\ 1 \leqslant i \leqslant n\}$ satisfies $\mathbb{P}(r_i \in \mathcal{G}(\max(\mu_1, \mu_2)) = \alpha$ for all $i \in \{1, 2, ..., n\}$.

Recall that the cumulative pseudo-regret after n plays of a policy π on $\nu = (\mathcal{G}(\mu_1), \mathcal{G}(\mu_2))$ is given by $R_n^\pi(\nu) = \sum_{m=1}^n \left(\max(\mu_1, \mu_2) - \mu_{t(\pi_m)}\right)$, where $t(\pi_m) \in \{1, 2\}$ indicates the type of the arm played by π at time m. Our goal is to lower bound $\mathbb{E}R_n^\pi(\nu)$, where the expectation is w.r.t. the randomness in π as well as the distribution over the possible realizations of ν. To this end, we define the notion of expected cumulative regret of π on a realization r of $\nu = (\mathcal{G}(\mu_1), \mathcal{G}(\mu_2))$ by

$$S_n^\pi(\nu, r) := \mathbb{E}^\pi \left[\sum_{m=1}^n \left(\max(\mu_1, \mu_2) - \mu_{t(\pi_m)} \right) \right],$$

where the expectation $\mathbb{E}^\pi$ is w.r.t. the randomness in π. Note that $\mathbb{E}R_n^\pi(\nu) = \mathbb{E}^\nu S_n^\pi(\nu, r)$, where the expectation $\mathbb{E}^\nu$ is w.r.t. the distribution over the possible realizations of ν. We define our problem class $\mathcal{N}_\Delta$ as the collection of Δ-separated instances given by

$$\mathcal{N}_\Delta := \left\{ (\mathcal{G}(\mu_1), \mathcal{G}(\mu_2)) : \mu_1 - \mu_2 = \Delta,\ (\mu_1, \mu_2) \in \mathbb{R}^2 \right\}.$$

Definition 1 (Consistent policy) *Let $\Lambda(r)$ denote the number of "optimal" arms in realization r. We call π, an asymptotically consistent policy for the problem class $\mathcal{N}_\Delta$ if for any instance $\nu \in \mathcal{N}_\Delta$ and any realization r thereof, it satisfies the following two conditions:*

$$\mathbb{E}R_n^\pi(\nu) = o\left(n^p\right) \qquad \textit{for every } p \in (0,1),\ \alpha \in (0,1]. \tag{1}$$

$$\mathbb{E}^\nu \left[S_n^\pi(\nu, r) \middle| \Lambda(r) = m \right] \geqslant \mathbb{E}^\nu \left[S_n^\pi(\nu, r) \middle| \Lambda(r) = k \right] \qquad \forall\, (m, n, k) : 0 \leqslant m \leqslant k \leqslant n. \tag{2}$$

The set of such policies is denoted by $\Pi_{\text{cons}}\left(\mathcal{N}_\Delta\right)$. Notice that (1), barring the condition on α, is the standard definition of asymptotic consistency first introduced in [6] and subsequently adopted by many other papers. The exclusion of $\alpha = 0$ is necessary since no policy can achieve sublinear regret in said case. We also remark that the additional condition in (2) is not restrictive since any reasonable policy is expected to incur a larger cumulative regret (in expectation) on realizations with fewer optimal arms.

Fix an arbitrary $\Delta > 0$ and consider an instance $\nu = (\{Q_1\}, \{Q_2\}) \in \mathcal{N}_\Delta$, where (Q_1, Q_2) are unit-variance Gaussian distributions with means (μ_1, μ_2) respectively. Consider an arbitrary realization $r \in \{Q_1, Q_2\}^n$ of ν and let $\mathcal{I} \subseteq \{1, 2, ..., n\}$ denote the set of inferior arms in r (arms with reward distribution Q_2). Consider another instance $\nu' \in \mathcal{N}_\Delta$ given by $\nu' = \left(\{\widetilde{Q}_1\}, \{Q_1\} \right)$, where $\widetilde{Q}_1$ is another unit variance Gaussian with mean $\mu_1 + \Delta$. Now consider a realization $r' \in \{\widetilde{Q}_1, Q_1\}^n$ of ν' that is such that the arms at positions in $\mathcal{I}$ have distribution $\widetilde{Q}_1$ while those at positions in $\{1, 2, ..., n\} \backslash \mathcal{I}$ have distribution Q_1. Notice that $\mathcal{I}$ is the set of optimal arms in r' (arms with reward distribution $\widetilde{Q}_1$), implying $\Lambda(r') = |\mathcal{I}|$. Then, the following always holds:

$$S_n^\pi(\nu, r) + S_n^\pi(\nu', r') \geqslant \left(\frac{\Delta n}{2}\right)\left(\mathbb{P}^\pi_{\nu,r}\left(\sum_{i\in\mathcal{I}} N_i(n) > \frac{n}{2}\right) + \mathbb{P}^\pi_{\nu',r'}\left(\sum_{i\in\mathcal{I}} N_i(n) \leqslant \frac{n}{2}\right)\right),$$

where $\mathbb{P}^\pi_{\nu,r}(\cdot)$ and $\mathbb{P}^\pi_{\nu',r'}(\cdot)$ denote the probability measures w.r.t. the instance-realization pairs (ν, r) and (ν', r') respectively, and $N_i(n)$ denotes the number of plays up to and including time n of arm $i \in \{1, 2, ..., n\}$. Using the Bretagnolle-Huber inequality (Theorem 14.2 of [7]), we obtain

$$S_n^\pi(\nu, r) + S_n^\pi(\nu', r') \geqslant \left(\frac{\Delta n}{4}\right)\exp\left(-\mathrm{D}\left(\mathbb{P}^\pi_{\nu,r}, \mathbb{P}^\pi_{\nu',r'}\right)\right),$$

where $\mathrm{D}\left(\mathbb{P}^\pi_{\nu,r}, \mathbb{P}^\pi_{\nu',r'}\right)$ denotes the KL-Divergence between $\mathbb{P}^\pi_{\nu,r}$ and $\mathbb{P}^\pi_{\nu',r'}$. Using Divergence decomposition (Lemma 15.1 of [7]), we further obtain

$$S_n^\pi(\nu, r) + S_n^\pi(\nu', r') \geqslant \left(\frac{\Delta n}{4}\right)\exp\left(-\left(\frac{\mathrm{D}\left(Q_2, \widetilde{Q}_1\right)}{\Delta}\right)S_n^\pi(\nu, r)\right) = \left(\frac{\Delta n}{4}\right)\exp\left(-2\Delta S_n^\pi(\nu, r)\right),$$

where the equality follows since $\widetilde{Q}_1$ and Q_2 are unit variance Gaussian distributions with means separated by 2Δ. Next, taking the expectation $\mathbb{E}^\nu$ on both the sides above and a direct application of Jensen's inequality thereafter yields

$$\mathbb{E}R_n^\pi(\nu) + \mathbb{E}^\nu S_n^\pi(\nu', r') \geqslant \left(\frac{\Delta n}{4}\right)\exp\left(-2\Delta\mathbb{E}R_n^\pi(\nu)\right). \tag{3}$$

Consider the $\mathbb{E}^\nu S_n^\pi(\nu', r')$ term in (3) and an arbitrary $\alpha \in (0, 1/2]$. Using a simple change-of-measure argument, we obtain

$$\begin{aligned}\mathbb{E}^\nu S_n^\pi(\nu', r') &= \mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\left(\frac{1-\alpha}{\alpha}\right)^{2\left(\Lambda(r') - n/2\right)}\right] \\ &\leqslant \mathbb{E}R_n^\pi(\nu') + \mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\left(\frac{1-\alpha}{\alpha}\right)^{2\left(\Lambda(r') - n/2\right)}\mathbb{1}\left\{\Lambda(r') > n/2\right\}\right], \end{aligned} \tag{4}$$

where the inequality follows since $\alpha \leqslant 1/2$. Now consider the second term on the RHS in (4). It follows that

$$\begin{aligned} &\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\left(\frac{1-\alpha}{\alpha}\right)^{2\left(\Lambda(r') - n/2\right)}\mathbb{1}\left\{\Lambda(r') > n/2\right\}\right] \\ &= \sum_{k>n/2}\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\left(\frac{1-\alpha}{\alpha}\right)^{2\left(\Lambda(r') - n/2\right)}\mathbb{1}\left\{\Lambda(r') = k\right\}\right] \\ &= \sum_{k>n/2}\left(\frac{1-\alpha}{\alpha}\right)^{(2k-n)}\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\mathbb{1}\left\{\Lambda(r') = k\right\}\right] \\ &= \sum_{k>n/2}\left(\frac{1-\alpha}{\alpha}\right)^{(2k-n)}\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\middle|\,\Lambda(r') = k\right]\mathbb{P}_{\nu'}\left(\Lambda(r') = k\right) \\ &= \sum_{k>n/2}\left(\frac{1-\alpha}{\alpha}\right)^{(2k-n)}\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\middle|\,\Lambda(r') = k\right]\binom{n}{k}\alpha^k(1-\alpha)^{(n-k)} \\ &= \alpha^n\sum_{k>n/2}\binom{n}{k}\left(\frac{1-\alpha}{\alpha}\right)^k\mathbb{E}^{\nu'}\left[S_n^\pi(\nu', r')\middle|\,\Lambda(r') = k\right]. \end{aligned} \tag{5}$$

Recall that $\nu' \in \mathcal{N}_\Delta$ and $\pi \in \Pi_{\text{cons}}\left(\mathcal{N}_\Delta\right)$. We have

$$\begin{aligned} \mathbb{E}R_n^\pi(\nu') &= \mathbb{E}^{\nu'} S_n^\pi(\nu', r') \\ &\geqslant \sum_{m=1}^{k} \mathbb{E}^{\nu'}\left[\left. S_n^\pi(\nu', r')\right| \Lambda(r') = m\right] \mathbb{P}_{\nu'}\left(\Lambda(r') = m\right) \qquad \text{(for any } k \leqslant n) \\ &\geqslant \mathbb{E}^{\nu'}\left[\left. S_n^\pi(\nu', r')\right| \Lambda(r') = k\right] \mathbb{P}_{\nu'}\left(\Lambda(r') \leqslant k\right). \qquad \text{(using (2))} \qquad (6) \end{aligned}$$

Since $\alpha \leqslant 1/2$, it follows that for any $k > n/2$, $\mathbb{P}_{\nu'}\left(\Lambda(r') \leqslant k\right) = \mathcal{O}_n(1)$ (the subscript n indicates that the asymptotic scaling is w.r.t. n). Using this observation together with (1) and (6), we conclude that

$$\forall\, k > n/2,\ \alpha \in (0, 1/2] \text{ and every } p \in (0,1),\ \mathbb{E}^{\nu'}\left[\left. S_n^\pi(\nu', r')\right| \Lambda(r') = k\right] = o\left(n^p\right). \qquad (7)$$

Combining (4), (5), (7) and using the fact that $\nu' \in \mathcal{N}_\Delta$ with $\pi \in \Pi_{\text{cons}}\left(\mathcal{N}_\Delta\right)$, we conclude

$$\forall\, k > n/2,\ \alpha \in (0, 1/2] \text{ and every } p \in (0,1),\ \mathbb{E}^{\nu} S_n^\pi(\nu', r') = o\left(n^p\right). \qquad (8)$$

Now consider (3). Taking the natural logarithm of both sides and rearranging, we obtain

$$\frac{\mathbb{E}R_n^\pi(\nu)}{\log n} \geqslant \left(\frac{1}{2\Delta}\right)\left(1 + \frac{\log\left(\frac{\Delta}{4}\right)}{\log n} - \frac{\log\left(\mathbb{E}R_n^\pi(\nu) + \mathbb{E}^{\nu} S_n^\pi(\nu', r')\right)}{\log n}\right).$$

Since $\nu, \nu' \in \mathcal{N}_\Delta$ and $\pi \in \Pi_{\text{cons}}\left(\mathcal{N}_\Delta\right)$, the assertion follows using (8) that for any $\alpha \in (0, 1/2]$,

$$\liminf_{n \to \infty} \frac{\mathbb{E}R_n^\pi(\nu)}{\log n} \geqslant \frac{1}{2\Delta}.$$

Therefore, for any $\Delta > 0$, $\exists\, \nu \in \mathcal{N}_\Delta$ and an absolute constant C s.t. the expected cumulative regret of any consistent policy π on ν satisfies $\forall\, \alpha \leqslant 1/2$ and n large enough, $\mathbb{E}R_n^\pi(\nu) \geqslant C\Delta^{-1}\log n$. □

B Proof of Theorem 2

We divide the horizon of play into epochs of length m each. For each $k \geqslant 0$, let S_k denote the cumulative pseudo-regret incurred by the algorithm when it is initialized at the beginning of epoch $(2k+1)$ and continued until the end of the horizon of play, i.e., the algorithm starts at time $2km+1$ and runs until time n. We are interested in an upper bound on $\mathbb{E}R_n^\pi = \mathbb{E}S_0$. To this end, suppose that the algorithm is initialized at time $2km+1$. Label the arms played in epochs $(2k+1)$ and $(2k+2)$ as '1' and '2' respectively. Let $\overline{X}_i$ denote the empirical mean reward from m plays of arm $i \in \{1, 2\}$. Recall that $t(i) \in \mathcal{T} = \{1, 2\}$ denotes the type of arm i, that type 1 is assumed optimal and lastly, that the probability of a new arm being of the optimal type is α. Suppose that $\mathbb{1}\{E\}$ denotes the indicator random variable associated with event E. Then, we have that S_k evolves according to the following stochastic recursive relation:

$$\begin{aligned} S_k = {} & \mathbb{1}\{t(1) = 1, t(2) = 2\}\left[\Delta m + \mathbb{1}\{\overline{X}_2 - \overline{X}_1 > \delta\}\left[n - (2k+2)m\right] + \mathbb{1}\{|\overline{X}_1 - \overline{X}_2| < \delta\}S_{k+1}\right] + \\ & \mathbb{1}\{t(1) = 2, t(2) = 1\}\left[\Delta m + \mathbb{1}\{\overline{X}_1 - \overline{X}_2 > \delta\}\left[n - (2k+2)m\right] + \mathbb{1}\{|\overline{X}_1 - \overline{X}_2| < \delta\}S_{k+1}\right] + \\ & \mathbb{1}\{t(1) = 2, t(2) = 2\}\left[2\Delta m + \mathbb{1}\{|\overline{X}_1 - \overline{X}_2| > \delta\}\Delta\left[n - (2k+2)m\right] + \mathbb{1}\{|\overline{X}_1 - \overline{X}_2| < \delta\}S_{k+1}\right] + \\ & \mathbb{1}\{t(1) = 1, t(2) = 1\}\mathbb{1}\{|\overline{X}_1 - \overline{X}_2| < \delta\}S_{k+1}. \end{aligned}$$

Collecting like terms together,

$$\begin{aligned} S_k = {} & \mathbb{1}\{t(1) = 1, t(2) = 2\}\mathbb{1}\{\overline{X}_2 - \overline{X}_1 > \delta\}\Delta\left[n - (2k+2)m\right] + \\ & \mathbb{1}\{t(1) = 2, t(2) = 1\}\mathbb{1}\{\overline{X}_1 - \overline{X}_2 > \delta\}\Delta\left[n - (2k+2)m\right] + \\ & \mathbb{1}\{t(1) = 2, t(2) = 2\}\mathbb{1}\{|\overline{X}_1 - \overline{X}_2| > \delta\}\Delta\left[n - (2k+2)m\right] + \\ & \left[\mathbb{1}\{t(1) \neq t(2)\} + 2\mathbb{1}\{t(1) = 2, t(2) = 2\}\right]\Delta m + \mathbb{1}\{|\overline{X}_1 - \overline{X}_2| < \delta\}S_{k+1}. \qquad (9) \end{aligned}$$

Define the following conditional events:

$$E_1 := \left\{\overline{X}_2 - \overline{X}_1 > \delta \;\middle|\; t(1) = 1,\ t(2) = 2\right\}, \qquad (10)$$

$$E_2 := \left\{\overline{X}_1 - \overline{X}_2 > \delta \;\middle|\; t(1) = 2,\ t(2) = 1\right\}, \qquad (11)$$

$$E_3 := \left\{\left|\overline{X}_1 - \overline{X}_2\right| > \delta \;\middle|\; t(1) = 2,\ t(2) = 2\right\}, \qquad (12)$$

$$E_4 := \left\{\left|\overline{X}_1 - \overline{X}_2\right| < \delta \;\middle|\; t(1) = t(2)\right\}, \qquad (13)$$

$$E_5 := \left\{\left|\overline{X}_1 - \overline{X}_2\right| < \delta \;\middle|\; t(1) \neq t(2)\right\}. \qquad (14)$$

Taking expectations on both sides in (9) and rearranging, one obtains the following using (10),(11),(12),(13),(14):

$$\begin{aligned}\mathbb{E}S_k = &\left[\alpha(1-\alpha)\left\{\mathbb{P}(E_1)+\mathbb{P}(E_2)\right\}+(1-\alpha)^2\mathbb{P}(E_3)\right]\Delta\left[n-(2k+2)m\right]\\ &+\left[2\alpha(1-\alpha)+2(1-\alpha)^2\right]\Delta m+\mathbb{P}\left(\left|\overline{X}_1-\overline{X}_2\right|<\delta\right)\mathbb{E}S_{k+1}.\end{aligned} \tag{15}$$

Notice that S_{k+1}, by definition, is independent of $(X_{i,j})_{i\in\{1,2\},1\leqslant j\leqslant m}$, and hence $\mathbb{E}\left[\mathbb{1}\{|\overline{X}_1-\overline{X}_2|<\delta\}S_{k+1}\right]=\mathbb{P}\left(\left|\overline{X}_1-\overline{X}_2\right|<\delta\right)\mathbb{E}S_{k+1}$ in (15). Further note that

$$\mathbb{P}\left(\left|\overline{X}_1-\overline{X}_2\right|<\delta\right)=\left[\alpha^2+(1-\alpha)^2\right]\mathbb{P}(E_4)+2\alpha(1-\alpha)\mathbb{P}(E_5). \tag{16}$$

From (15) and (16), we conclude after a little rearrangement the following:

$$\mathbb{E}S_k=\xi_1-\xi_2 k+\xi_3\mathbb{E}S_{k+1}, \tag{17}$$

where the ξ_i's do not depend on k and are given by

$$\xi_1 := \Delta\left[\alpha(1-\alpha)\left\{\mathbb{P}(E_1)+\mathbb{P}(E_2)\right\}+(1-\alpha)^2\mathbb{P}(E_3)\right](n-2m)+2\Delta(1-\alpha)m, \tag{18}$$

$$\xi_2 := 2\Delta\left[\alpha(1-\alpha)\left\{\mathbb{P}(E_1)+\mathbb{P}(E_2)\right\}+(1-\alpha)^2\mathbb{P}(E_3)\right]m, \tag{19}$$

$$\xi_3 := \left[\alpha^2+(1-\alpha)^2\right]\mathbb{P}(E_4)+2\alpha(1-\alpha)\mathbb{P}(E_5). \tag{20}$$

Observe that the recursion in (17) is solvable in closed-form and admits the following solution:

$$\mathbb{E}S_0=\xi_1\sum_{k=0}^{l-1}\xi_3^k-\xi_2\sum_{k=0}^{l-1}k\xi_3^k+\xi_3^l\mathbb{E}S_l, \tag{21}$$

where $l:=\lfloor n/(2m)\rfloor$. Since the ξ_i's are all non-negative for $n\geqslant 2m$ and $\mathbb{E}S_l\leqslant 2\Delta m$, we have for $n\geqslant 2m$,

$$\mathbb{E}R_n^\pi=\mathbb{E}S_0\leqslant\frac{\xi_1}{1-\xi_3}+2\Delta m. \tag{22}$$

Now using (10),(11),(12),(13),(14) and Hoeffding's inequality [4] along with the fact that the X_i's are bounded in $[0,1]$, we conclude

$$\left\{\mathbb{P}(E_1),\mathbb{P}(E_2)\right\}\leqslant\exp\left(-(\Delta+\delta)^2m/2\right), \tag{23}$$

$$\left\{\mathbb{P}(E_3),\mathbb{P}(E_4^c)\right\}\leqslant 2\exp\left(-\delta^2m/2\right), \tag{24}$$

$$\mathbb{P}(E_5)\leqslant\exp\left(-(\Delta-\delta)^2m/2\right). \tag{25}$$

From (18),(19),(20),(22),(23),(24) and (24), we conclude

$$\mathbb{E}R_n^\pi\leqslant\frac{2\Delta n\exp\left(-\delta^2m/2\right)+\Delta m}{\alpha\left(1-\exp\left(-(\Delta-\delta)^2m/2\right)\right)}+2\Delta m.$$

Finally since $m=\left\lceil\left(2/\delta^2\right)\log n\right\rceil$, the stated assertion follows, i.e., for all $n\geq 2m$,

$$\mathbb{E}R_n^\pi\leqslant 2\Delta\left(1+\frac{1}{2\alpha}\right)\left[\left(\frac{2}{\delta^2}\right)\log n+1\right]+\left(\frac{\Delta}{\alpha}\right)\left[2+f(n,\delta,\Delta)\right], \tag{26}$$

where $f(n,\delta,\Delta)=o(1)$ in n given by

$$f(n,\delta,\Delta):=\left(\frac{n^{-\left(\frac{\Delta-\delta}{\delta}\right)^2}}{1-n^{-\left(\frac{\Delta-\delta}{\delta}\right)^2}}\right)\left[\left(\frac{2}{\delta^2}\right)\log n+3\right]. \tag{27}$$

For $n<2m$, $\mathbb{E}R_n^\pi\leqslant 2\Delta m$ follows trivially. Therefore, the bound in (26) is valid for all $n\geqslant 1$. Of course, $\mathbb{E}R_n^\pi\leqslant\Delta n$ offers a sharper bound whenever Δ is very small, similar to finite-armed settings. Thus in conclusion, $\mathbb{E}R_n^\pi$ is bounded as follows for any n:

$$\mathbb{E}R_n^\pi\leqslant\min\left[\Delta n,\ 2\Delta\left(1+\frac{1}{2\alpha}\right)\left\{\left(\frac{2}{\delta^2}\right)\log n+1\right\}+\left(\frac{\Delta}{\alpha}\right)\left\{2+f(n,\delta,\Delta)\right\}\right].$$

□

C Proof of Proposition 1

The statement of the proposition assumes $|\mu_1 - \mu_2| = \Delta > 0$. However, we will only prove it for the case where $\mu_1 - \mu_2 = \Delta > 0$. The proof for the other case is symmetric and an identical bound will follow. Fix an arbitrary $(F_1, F_2) \in \mathcal{G}(\mu_1) \times \mathcal{G}(\mu_2)$ and consider the following stopping time:

$$\tau := \inf\left\{n \geqslant 1 : \sum_{k=1}^{n}(\Psi_k - \bar{\theta}_n) < 0\right\}, \tag{28}$$

where $\Psi_k := Y_{1,j}^{F_1} - Y_{2,j}^{F_2}$ and $\bar{\theta}_n := \theta_n/n$. Note that $\mathbb{E}\Psi_k = \Delta > 0$ (by assumption). Then, it follows that $\mathbb{P}\left(\bigcap_{m=1}^{\infty}\left|\sum_{j=1}^{m}\left(Y_{1,j}^{F_1} - Y_{2,j}^{F_2}\right)\right| \geqslant \theta_m\right) \geqslant \mathbb{P}(\tau = \infty)$. Therefore, it suffices to show that $\mathbb{P}(\tau = \infty)$ is bounded away from 0. To this end, fix an arbitrary $\lambda \in (0,1)$ and let $n_0 := \min\{k \in \mathbb{N} : \bar{\theta}_n \leqslant \lambda\Delta\}$. Since $\bar{\theta}_n \to 0$ as $n \to \infty$ and $\Delta > 0$, it follows that $n_0 < \infty$. Suppose that ω denotes an arbitrary sample-path and consider the following set:

$$E := \left\{\omega : \Psi_k(\omega) > \bar{\theta}_k;\ 1 \leqslant k \leqslant n_0\right\}. \tag{29}$$

Since Assumption 1 (main text) is satisfied, $n_0 < \infty$ and $\bar{\theta}_n$ is monotone decreasing in n with $\bar{\theta}_1 < 1$, it follows that $\mathbb{P}(E)$, as given below, is strictly positive.

$$\mathbb{P}(E) = \prod_{k=1}^{n_0}\mathbb{P}\left(\Psi_k > \bar{\theta}_k\right) > 0, \text{ where } n_0 = \min\{k \in \mathbb{N} : \bar{\theta}_n \leqslant \lambda\Delta\}. \tag{30}$$

Notice that $\tau > n_0$ on the event indicated by E. In particular,

$$\begin{aligned}
\tau | E &= \inf\left\{ n \geqslant n_0 + 1 : \sum_{k=n_0+1}^{n}(\Psi_k - \bar{\theta}_n) < -\sum_{k=1}^{n_0}(\Psi_k - \bar{\theta}_n)\ \middle|\ E\right\}\\
&\underset{(\dagger)}{\geqslant} \inf\left\{ n \geqslant n_0 + 1 : \sum_{k=n_0+1}^{n}(\Psi_k - \bar{\theta}_n) < -\sum_{k=1}^{n_0}(\bar{\theta}_k - \bar{\theta}_n)\ \middle|\ E\right\}\\
&\underset{(\ddagger)}{\geqslant} \inf\left\{ n \geqslant n_0 + 1 : \sum_{k=n_0+1}^{n}(\Psi_k - \bar{\theta}_n) < -\sum_{k=1}^{n_0}(\bar{\theta}_k - \bar{\theta}_{n_0})\ \middle|\ E\right\}\\
&\underset{(\bullet)}{\geqslant} \inf\left\{ n \geqslant n_0 + 1 : \sum_{k=n_0+1}^{n}(\Psi_k - \lambda\Delta) < -\sum_{k=1}^{n_0}(\bar{\theta}_k - \bar{\theta}_{n_0})\ \middle|\ E\right\}\\
&\underset{(\star)}{=} n_0 + \inf\left\{ n \geqslant 1 : \sum_{k=1}^{n}(\Psi'_k - \lambda\Delta) < -\eta\right\},
\end{aligned} \tag{31}$$

where ($\dagger$) follows from (29), ($\ddagger$) follows since $\bar{\theta}_n \leqslant \bar{\theta}_{n_0}$ for $n \geqslant n_0$, ($\bullet$) since $\bar{\theta}_n \leqslant \lambda\Delta$ for $n \geqslant n_0$, and ($\star$) holds with $\eta := \sum_{k=1}^{n_0}(\bar{\theta}_k - \bar{\theta}_{n_0})$ and $\Psi'_k := \Psi_{n_0+k}$ since $(\Psi'_k)_{k\in\mathbb{N}}$ is independent of E. Note that $\eta > 0$ since $\bar{\theta}_n$ is monotone decreasing in n. Now consider the following stopping time:

$$\tau' := \inf\left\{n \geqslant 1 : \sum_{k=1}^{n}(\Psi'_k - \lambda\Delta) < -\eta\right\}. \tag{32}$$

It follows from (31) and (32) that $\mathbb{P}(\tau = \infty | E) \geqslant \mathbb{P}(\tau' = \infty)$. We next show that $\mathbb{P}(\tau' = \infty)$ is bounded away from 0.

Let $S_n := \sum_{k=1}^{n}(\Psi'_k - \lambda\Delta)$, with $S_0 := 0$. Since the Ψ'_k's are i.i.d. with $\mathbb{E}\Psi'_1 = \Delta$ and $|\Psi'_k| \leqslant 1$, it follows that $W_n := \exp\left(aS_n\right)$ is a Martingale w.r.t. $(\Psi'_k)_{k\in\mathbb{N}}$, where '$a$' is the non-zero solution to $\mathbb{E}\left[\exp\left(a\left(\Psi'_1 - \lambda\Delta\right)\right)\right] = 1$ (Note that $\mathbb{E}\Psi'_1 = \Delta > 0$ and $\lambda \in (0,1)$ ensures $a < 0$.). Fix an arbitrary $b > 0$ and define $T_{\eta,b} := \inf\{n \geqslant 1 : S_n \notin [-\eta, b]\}$ (We already know that $\eta > 0$.). By Doob's Optional Stopping Theorem [3], it follows that $\mathbb{E}W_{\min(T_{\eta,b},n)} = \mathbb{E}W_0 = 1$. Furthermore, since the stopped Martingale $W_{\min(T_{\eta,b},n)}$ is uniformly integrable, we in fact have $\mathbb{E}W_{T_{\eta,b}} = 1$. Thereafter using Markov's inequality, we obtain $\mathbb{P}\left(S_{T_{\eta,b}} < -\eta\right) = \mathbb{P}\left(W_{T_{\eta,b}} > e^{-\eta a}\right) \leqslant \exp(\eta a)$.

Since $b > 0$ is arbitrary, taking $\lim_{b\to\infty}$ on both sides and invoking the Bounded Convergence Theorem, we finally conclude that $\mathbb{P}(\tau' = \infty) = \mathbb{P}\left(S_{T_{\eta,\infty}} \geqslant -\eta\right) \geqslant 1 - \exp(\eta a)$, and hence

$$\mathbb{P}(\tau = \infty | E) \geqslant 1 - \exp(\eta a) > 0. \tag{33}$$

In conclusion,

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \left|\sum_{j=1}^{m}\left(Y_{1,j}^{F_1} - Y_{2,j}^{F_2}\right)\right| \geqslant \theta_m\right) \geqslant \mathbb{P}(\tau = \infty) \geqslant \mathbb{P}(\tau = \infty|E)\mathbb{P}(E)$$

$$\underset{(*)}{\geqslant} (1 - \exp(\eta a)) \prod_{k=1}^{n_0} \mathbb{P}(\Psi_k > \bar{\theta}_k) > 0,$$

where $(*)$ follows from (30) and (33). Since $(F_1, F_2) \in \mathcal{G}(\mu_1) \times \mathcal{G}(\mu_2)$ is arbitrary, taking $\min_{F_1\in\mathcal{G}(\mu_1), F_2\in\mathcal{G}(\mu_2)}$ on both the sides above appealing to the fact that the $\mathcal{G}(\mu_i)$'s are finite, proves our assertion. □

D Proof of Theorem 3

Consider the first epoch and assign the labels $1, 2$ to the two arms picked to be played in this epoch. Suppose $N_i(n)$ denotes the number of times arm i is played up to and including time n. Let $M_n := \min\left(N_1(n), N_2(n)\right)$ and define the following stopping time:

$$\tau := \inf\left\{n \geqslant 2 : \left|\sum_{k=1}^{M_n}(X_{1,k} - X_{2,k})\right| < \theta_{M_n}\right\},$$

where the sequence $\Theta \equiv (\theta_m)_{m\in\mathbb{N}}$ is defined through (2) (main text). Then, τ denotes the time of the terminal play in the first epoch after which the algorithm starts over again. Recall that $t(i)$ denotes the type of arm i and define the following conditional stopping times:

$$\tau_I := \tau \mid \{t(1) = t(2) = 2\}, \tag{34}$$

$$\tau_D := \tau \mid \{t(1) \neq t(2)\}, \tag{35}$$

where the subscripts I and D above indicate "Identical" and "Distinct" types, respectively. Let S_n denote the cumulative pseudo-regret of UCB1 after n plays in a stochastic two-armed bandit problem with separation Δ. Recall that R_n^π denotes the cumulative pseudo-regret of $\pi = \text{ALG}\left(\text{UCB1}, \Theta, 2\right)$ after n plays; we shall suppress the superscript π for notational simplicity and write R_n for R_n^π. For any $n \in \mathbb{N}$, let R'_n be an i.i.d. copy of R_n. Then, R_n must satisfy the following stochastic recursive relation:

$$\begin{aligned}
R_n &= \mathbb{1}\left\{t(1) \neq t(2)\right\} S_{\min(\tau,n)} + \mathbb{1}\left\{t(1) = t(2) = 2\right\} \Delta \min(\tau, n) + R'_{n-\min(\tau,n)} \\
&\leqslant \mathbb{1}\left\{t(1) \neq t(2)\right\} S_n + \mathbb{1}\left\{t(1) = t(2) = 2\right\} \Delta\tau + R'_{n-\min(\tau,n)} \\
&= \mathbb{1}\left\{t(1) \neq t(2)\right\} S_n + \mathbb{1}\left\{t(1) = t(2) = 2\right\} \Delta\tau + \sum_{k=2}^{n} \mathbb{1}\{\tau = k\} R'_{n-k} \\
&\leqslant \mathbb{1}\left\{t(1) \neq t(2)\right\} S_n + \mathbb{1}\left\{t(1) = t(2) = 2\right\} \Delta\tau + \mathbb{1}\{\tau \leqslant n\} R'_n\ ,
\end{aligned} \tag{36}$$

where the last step holds since $R'_{n-k} \leqslant R'_n\ \forall\ k \leqslant n$ (this follows trivially since π is agnostic to the length of the horizon of play[1]). Taking expectations on both sides of (36), we obtain

$$\begin{aligned}
\mathbb{E}R_n &\underset{(\dagger)}{\leqslant} 2\alpha(1-\alpha)\mathbb{E}S_n + (1-\alpha)^2\Delta\mathbb{E}\tau_I + \left[2\alpha(1-\alpha)\mathbb{P}(\tau_D \leqslant n) + \alpha^2 + (1-\alpha)^2\right]\mathbb{E}R_n \\
&\underset{(\ddagger)}{\leqslant} 2\alpha(1-\alpha)\mathbb{E}S_n + (1-\alpha)^2\Delta\mathbb{E}\tau_I + \left[2\alpha(1-\alpha)(1-\beta) + \alpha^2 + (1-\alpha)^2\right]\mathbb{E}R_n \\
&= 2\alpha(1-\alpha)\mathbb{E}S_n + (1-\alpha)^2\Delta\mathbb{E}\tau_I + \left(1 - 2\beta\alpha(1-\alpha)\right)\mathbb{E}R_n \\
\implies\ \mathbb{E}R_n &\leqslant \left(\frac{1}{\beta}\right)\mathbb{E}S_n + \left(\frac{(1-\alpha)\Delta\mathbb{E}\tau_I}{2\beta\alpha}\right),
\end{aligned}$$

[1] We could not claim this directly for Algorithm 1 as it depended on ex ante knowledge of the length of play.

where (†) uses (34), (35) and the fact that $\mathcal{D}(\mathcal{T}) = (\alpha, 1-\alpha)$, and (‡) follows from part (i) of Lemma 2 (see Appendix F). We also know from part (ii) of Lemma 2 that $\mathbb{E}\tau_I < C_0$, where C_0 is a constant that depends on the user-defined parameters (m_0, γ). The proof now concludes by invoking Theorem 1 of [1] for an upper bound on $\mathbb{E}S_n$ in order to obtain the desired upper bound on $\mathbb{E}R_n$, i.e.,

$$\begin{aligned}\mathbb{E}R_n &\leqslant \left(\frac{8}{\beta\Delta}\right)\log n + \left(1 + \frac{\pi^2}{3} + \frac{(1-\alpha)C_0}{2\alpha}\right)\left(\frac{\Delta}{\beta}\right)\\ &\leqslant 8\left(\beta\Delta\right)^{-1}\log n + \left(C_1 + \alpha^{-1}C_2\right)\beta^{-1}\Delta,\end{aligned}$$

where $C_1 := 1 + \pi^2/3$ and $C_2 := C_0/2$. □

E Proof of Theorem 4

We begin by noting that the following is true for any integer $u > 1$ and $i \in \{1, 2\}$:

$$N_i(n) \leqslant u + \sum_{t=u+1}^{n} \mathbb{1}\left\{I_t = i,\ N_i(t) > u\right\},$$

where $I_t \in \{1, 2\}$ denotes the index of the arm played at time t. We set $u = (1/2 + \epsilon)\, n$ for an arbitrary $\epsilon \in (0, 1/2)$ and without loss of generality, carry out the rest of the analysis fixing $i = 1$. We have,

$$\begin{aligned}N_1(n) &\leqslant \left(\frac{1}{2}+\epsilon\right)n + \sum_{t=\left(\frac{1}{2}+\epsilon\right)n+1}^{n} \mathbb{1}\left\{I_t = 1,\ N_1(t) > \left(\frac{1}{2}+\epsilon\right)n\right\}\\ &\leqslant \left(\frac{1}{2}+\epsilon\right)n + \sum_{t=\left(\frac{1}{2}+\epsilon\right)n+1}^{n} \mathbb{1}\left\{I_t = 1,\ N_1(t) > \left(\frac{1}{2}+\epsilon\right)t\right\}\\ &= \left(\frac{1}{2}+\epsilon\right)n + \sum_{t=\left(\frac{1}{2}+\epsilon\right)n+1}^{n} \mathbb{1}\left\{B_{1,t-1} > B_{2,t-1},\ N_1(t-1) > \left(\frac{1}{2}+\epsilon\right)t - 1\right\},\end{aligned}$$

where $B_{i,t} := \overline{X}_i(t) + \sqrt{(2\log t)/N_i(t)}$ for $i \in \{1, 2\}$, with $\overline{X}_i(t)$ denoting the empirical mean reward from the first $N_i(t)$ plays of arm i. Therefore,

$$\begin{aligned}N_1(n) &\leqslant \left(\frac{1}{2}+\epsilon\right)n + \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{1}\left\{B_{1,t} > B_{2,t},\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right\}\\ &= \left(\frac{1}{2}+\epsilon\right)n + Z_n, \qquad (37)\end{aligned}$$

where $Z_n := \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{1}\left\{B_{1,t} > B_{2,t},\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right\}$. Then,

$$\begin{aligned}&\mathbb{E}Z_n\\ &= \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(B_{1,t} > B_{2,t},\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right)\\ &= \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(\frac{\sum_{j=1}^{N_1(t)} X_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} X_{2,j}}{N_2(t)} > \sqrt{2\log t}\left(\frac{1}{\sqrt{N_2(t)}} - \frac{1}{\sqrt{N_1(t)}}\right),\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right)\\ &= \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)} > \sqrt{2\log t}\left(\frac{1}{\sqrt{N_2(t)}} - \frac{1}{\sqrt{N_1(t)}}\right),\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right),\\ &\qquad (38)\end{aligned}$$

where $Y_{i,j} := X_{i,j} - \mathbb{E}X_{i,j}$ for $i \in \{1, 2\}$, $j \in \mathbb{N}$. Note that (38) follows since the mean rewards of both the arms are equal.

E.1 Proof of part (i)

Consider an arbitrary non-negative integer $m \leqslant \left(\frac{1}{2} - \epsilon\right) t - 1$. Let $n_1(m) := \left(\frac{1}{2} + \epsilon\right) t + m$ and $n_2(m) := t - n_1(m)$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)} > \sqrt{2\log t}\left(\frac{1}{\sqrt{N_2(t)}} - \frac{1}{\sqrt{N_1(t)}}\right),\ N_1(t) = n_1(m)\right)\\
&\leqslant \mathbb{P}\left(\frac{\sum_{j=1}^{n_1(m)} Y_{1,j}}{n_1(m)} - \frac{\sum_{j=1}^{n_2(m)} Y_{2,j}}{n_2(m)} > \sqrt{2\log t}\left(\frac{1}{\sqrt{n_2(m)}} - \frac{1}{\sqrt{n_1(m)}}\right)\right)\\
&\underset{(\dagger)}{\leqslant} \exp\left(-4\left(\frac{t - 2\sqrt{n_1(m)n_2(m)}}{t}\right)\log t\right)\\
&\underset{(\ddagger)}{\leqslant} \exp\left(-4\left(1 - \sqrt{1-4\epsilon^2}\right)\log t\right), \qquad (39)
\end{aligned}$$

where $(\dagger)$ follows using Hoeffding's inequality [4] and $(\ddagger)$, since the product $n_1(m)n_2(m)$ is maximized on the set $\{m : 0 \leqslant m \leqslant (1/2 - \epsilon)\, t - 1\}$ at $m = 0$. From (38), we have

$$\begin{aligned}
&\mathbb{E}Z_n\\
&= \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \sum_{m=0}^{\left(\frac{1}{2}-\epsilon\right)t-1} \mathbb{P}\left(\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)} > \sqrt{2\log t}\left(\frac{1}{\sqrt{N_2(t)}} - \frac{1}{\sqrt{N_1(t)}}\right),\ N_1(t) = n_1(m)\right)\\
&\underset{(\star)}{\leqslant} \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \sum_{m=0}^{\left(\frac{1}{2}-\epsilon\right)t-1} \exp\left(-4\left(1 - \sqrt{1-4\epsilon^2}\right)\log t\right)\\
&= \left(\frac{1}{2} - \epsilon\right) \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} t\exp\left(-4\left(1 - \sqrt{1-4\epsilon^2}\right)\log t\right)\\
&\underset{(*)}{<} 2^{\rho(\epsilon)} n^{-(\rho(\epsilon)-1)}, \qquad (40)
\end{aligned}$$

where $(\star)$ follows from (39) and $(*)$ holds with $\rho(\epsilon) := 3 - 4\sqrt{1-4\epsilon^2} > 0$ for $\epsilon > \sqrt{7}/8$. Now consider an arbitrary $\delta \in (0, 1)$. Then,

$$\begin{aligned}
\mathbb{P}\left(\frac{N_1(n)}{n} \geqslant \left(\frac{1}{2} + \epsilon + \delta\right)\right) &= \mathbb{P}\left(N_1(n) - \left(\frac{1}{2} + \epsilon\right) n \geqslant \delta n\right)\\
&\leqslant \mathbb{P}(Z_n \geqslant \delta n) && \text{(using (37))}\\
&\leqslant \frac{\mathbb{E}Z_n}{\delta n} && \text{(Markov's inequality)}\\
&\leqslant \left(\frac{2^{\rho(\epsilon)}}{\delta}\right) n^{-\rho(\epsilon)} && \text{(using (40))}\\
&\leqslant \left(\frac{8}{\delta}\right) n^{-\rho(\epsilon)}.
\end{aligned}$$

Note that $\rho(\epsilon) \leqslant 0$ for $\epsilon \leqslant \sqrt{7}/8$. Thus, the above result trivially holds for all $\epsilon \in (0, 1/2)$. An identical result holds also for $N_2(n)$ by the symmetry of our proof. Therefore for any $i \in \{1, 2\}$, we have

$$\mathbb{P}\left(\left|\frac{N_i(n)}{n} - \frac{1}{2}\right| \geqslant \epsilon + \delta\right) \leqslant \left(\frac{8}{\delta}\right) n^{-\left(3 - 4\sqrt{1-4\epsilon^2}\right)}.$$

The form of the result stated in the theorem can be obtained by making the following substitutions order-wise: $\delta \leftarrow \delta'\epsilon'$, $\epsilon \leftarrow (1 - \delta')\epsilon'$, $\delta' \leftarrow \delta$, $\epsilon' \leftarrow \epsilon$. □

E.2 Proof of part (ii)

From (38), we have

$$\begin{aligned}
&\mathbb{E}Z_n \\
&\leqslant \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)} > \sqrt{\frac{2\log t}{t}}\left(\frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right),\ N_1(t) \geqslant \left(\frac{1}{2}+\epsilon\right)t\right) \\
&\leqslant \sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(W_t > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right), \qquad (41)
\end{aligned}$$

where $W_t := \sqrt{\frac{t}{2\log t}}\left(\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)} - \frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)}\right)$. Now,

$$\begin{aligned}
&|W_t| \\
&\leqslant \sqrt{\frac{t}{2\log t}}\left(\left|\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{N_1(t)}\right| + \left|\frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{N_2(t)}\right|\right) \\
&= \sqrt{\frac{t}{\log t}}\left(\sqrt{\frac{\log\log N_1(t)}{N_1(t)}}\left|\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{\sqrt{2N_1(t)\log\log N_1(t)}}\right| + \sqrt{\frac{\log\log N_2(t)}{N_2(t)}}\left|\frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{\sqrt{2N_2(t)\log\log N_2(t)}}\right|\right) \\
&\leqslant \sqrt{\frac{t}{\log t}}\left(\sqrt{\frac{\log\log t}{N_1(t)}}\left|\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{\sqrt{2N_1(t)\log\log N_1(t)}}\right| + \sqrt{\frac{\log\log t}{N_2(t)}}\left|\frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{\sqrt{2N_2(t)\log\log N_2(t)}}\right|\right) \\
&= \sqrt{\frac{\log\log t}{\log t}}\left(\sqrt{\frac{t}{N_1(t)}}\left|\frac{\sum_{j=1}^{N_1(t)} Y_{1,j}}{\sqrt{2N_1(t)\log\log N_1(t)}}\right| + \sqrt{\frac{t}{N_2(t)}}\left|\frac{\sum_{j=1}^{N_2(t)} Y_{2,j}}{\sqrt{2N_2(t)\log\log N_2(t)}}\right|\right).
\end{aligned} \tag{42}$$

Notice that the following can be deduced from part (i) of Theorem 4 using the Borel-Cantelli Lemma:

$$\liminf_{t\to\infty} \frac{N_i(t)}{t} \geqslant \frac{1}{2} - \frac{\sqrt{3}}{4} \quad \text{w.p. } 1 \ \ \forall\, i \in \{1,2\}. \tag{43}$$

In addition to the result in (43) that holds *w.p.* 1, we also know that $N_i(t)$, for any $i \in \{1,2\}$ and $t \geqslant 0$, can be lower bounded *pathwise* by a deterministic non-decreasing function of time, say $\lambda(t)$, that grows to $+\infty$ as $t \to \infty$. This is a trivial consequence due to the structure of the UCB1 policy and the fact that the rewards are bounded. We therefore have for any $i \in \{1,2\}$,

$$\left|\frac{\sum_{j=1}^{N_i(t)} Y_{i,j}}{\sqrt{2N_i(t)\log\log N_i(t)}}\right| \leq \sup_{m\geq\lambda(t)}\left|\frac{\sum_{j=1}^{m} Y_{i,j}}{\sqrt{2m\log\log m}}\right|.$$

Now for any fixed $i \in \{1,2\}$, $\mathbb{E}Y_{i,j} \sim$ i.i.d. $\forall\, j$ with $\mathbb{E}Y_{i,1} = 0$ and $\text{Var}\left(Y_{i,1}\right) = \text{Var}\left(X_{i,1}\right) \leqslant 1$. Also, $\lambda(t)$ is non-decreasing and $\lambda(t) \uparrow \infty$. Therefore, the Law of the Iterated Logarithm [5] implies

$$\limsup_{t\to\infty}\left|\frac{\sum_{j=1}^{N_i(t)} Y_{i,j}}{\sqrt{2N_i(t)\log\log N_i(t)}}\right| \leqslant 1 \quad \text{w.p. } 1 \ \ \forall\, i \in \{1,2\}. \tag{44}$$

From (42), (43) and (44), we conclude that

$$\lim_{t\to\infty} W_t = 0 \quad \text{w.p. } 1. \tag{45}$$

Now consider an arbitrary $\delta > 0$. Then,

$$\begin{aligned}\mathbb{P}\left(\frac{N_1(n)}{n} \geqslant \left(\frac{1}{2}+\epsilon+\delta\right)\right) &= \mathbb{P}\left(N_1(n) - \left(\frac{1}{2}+\epsilon\right)n \geqslant \delta n\right) \\ &\underset{(\dagger)}{\leqslant} \mathbb{P}(Z_n \geqslant \delta n) \\ &\underset{(\ddagger)}{\leqslant} \frac{\mathbb{E}Z_n}{\delta n} \\ &\underset{(\star)}{\leqslant} \frac{1}{\delta n}\sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(W_t > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right),\end{aligned}$$

where $(\dagger)$ follows using (37), $(\ddagger)$ using Markov's inequality and $(\star)$ from (41). Now,

$$\begin{aligned}\mathbb{P}\left(\frac{N_1(n)}{n} \geqslant \left(\frac{1}{2}+\epsilon+\delta\right)\right) &\leqslant \frac{1}{\delta n}\sum_{t=\left(\frac{1}{2}+\epsilon\right)n}^{n-1} \mathbb{P}\left(W_t > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right) \\ &\leqslant \left(\frac{\frac{1}{2}-\epsilon}{\delta}\right) \sup_{\left(\frac{1}{2}+\epsilon\right)n \leqslant t \leqslant n-1} \mathbb{P}\left(W_t > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right) \\ &\leqslant \left(\frac{\frac{1}{2}-\epsilon}{\delta}\right) \sup_{t \geqslant n/2} \mathbb{P}\left(W_t > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right). \qquad (46)\end{aligned}$$

Using (45) and (46), we conclude that

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{N_1(n)}{n} \geqslant \left(\frac{1}{2}+\epsilon+\delta\right)\right) \leqslant \left(\frac{\frac{1}{2}-\epsilon}{\delta}\right) \limsup_{n\to\infty} \mathbb{P}\left(W_n > \frac{1}{\sqrt{\left(\frac{1}{2}-\epsilon\right)}} - \frac{1}{\sqrt{\left(\frac{1}{2}+\epsilon\right)}}\right) = 0.$$

Since $\delta > 0$ is arbitrary, it follows that $\lim_{n\to\infty} \mathbb{P}\left(\frac{N_1(n)}{n} \geqslant \frac{1}{2}+\epsilon\right) = 0$ for any $\epsilon > 0$. Since our proof is symmetric w.r.t. the arms, we also have $\lim_{n\to\infty} \mathbb{P}\left(\frac{N_2(n)}{n} \geqslant \frac{1}{2}+\epsilon\right) = 0 \implies \lim_{n\to\infty} \mathbb{P}\left(\frac{N_1(n)}{n} \leqslant \frac{1}{2}-\epsilon\right) = 0$. Therefore, $\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{N_i(n)}{n} - \frac{1}{2}\right| \geqslant \epsilon\right) = 0$ for $i \in \{1,2\}$ and any $\epsilon > 0$. □

F Ancillary results

Lemma 1 *Consider a stochastic two-armed bandit with rewards bounded in $[0,1]$. Suppose that the reward distributions of the two arms $(F_1, F_2) \in \mathcal{G}(\mu_1) \times \mathcal{G}(\mu_2)$ satisfy Assumption 1 (main text). Let $N_i(n)$ denote the number of times arm i is played by UCB1 [1] up to and including time n. At any time n^+, $(X_{i,k})_{k=1}^m$ denotes the sequence of rewards realized from the first $m \leqslant N_i(n)$ plays of arm i. For each $n \in \mathbb{N}$, let $M_n := \min(N_1(n), N_2(n))$ and consider the following stopping times:*

$$\tau := \inf\left\{n \geqslant 2 : \left|\sum_{k=1}^{M_n}(X_{1,k} - X_{2,k})\right| < \theta_{M_n}\right\}, \qquad (47)$$

$$\tau' := \inf\left\{n \geqslant 1 : \left|\sum_{k=1}^{n}(X_{1,k} - X_{2,k})\right| < \theta_n\right\}, \qquad (48)$$

where the sequence $\Theta \equiv \{\theta_n : n = 1, 2, ...\}$ is defined through (2) (main text). Then, $M_\tau = \tau'$ pathwise.

Lemma 2 *Consider the setting of Lemma 1. Recall that $\mathcal{T} = \{1,2\}$ and $t(i) \in \mathcal{T}$ denotes the type of arm i. Define the following conditional stopping times:*

$$\tau_D := \tau \mid t(1) \neq t(2), \qquad (49)$$

$$\tau_I := \tau \mid t(1) = t(2), \qquad (50)$$

where the subscripts D and I indicate "Distinct" and "Identical" types, respectively. Then, the following results hold:

(i) $\mathbb{P}(\tau_D = \infty) \geqslant \beta$*, where* β *is as defined in* (1) *(main text).*

(ii) $\mathbb{E}\tau_I < C_0$*, where* C_0 *is a constant that depends on the user-defined parameters* (m_0, γ) *featuring in (2) (main text) that ensure* Θ *satisfies the conditions of Proposition 1 (main text).*

F.1 Proof of Lemma 1

We begin by noting the following facts:

1. *Fact 1:* $(M_n)_{n\geqslant 2}$ is a non-decreasing sequence of natural numbers (starting from $M_2 = 1$), with $M_{n+1} \leqslant M_n + 1$.

2. *Fact 2:* For each $i \in \{1, 2\}$, $\liminf_{n\to\infty} N_i(n) = \infty$ pathwise[2] (consequence of UCB1 and bounded rewards). Consequently, $\liminf_{n\to\infty} M_n = \infty$ pathwise.

Define $\Psi_k := X_{1,k} - X_{2,k}$. Fix some $m \in \mathbb{N}$ and consider an arbitrary sample-path ω such that $M_\tau(\omega) = m$. Then on ω, we must also have $m = \inf\left\{l \geqslant 1 : \left|\sum_{k=1}^{l} \Psi_k(\omega)\right| < \theta_l\right\}$ (follows from the definitions of τ and τ'). Since the choice of m is arbitrary (due to *Fact 1* and *Fact 2*), it must be that on any arbitrary ω, $M_\tau(\omega) = \inf\left\{l \geqslant 1 : \left|\sum_{k=1}^{l} \Psi_k(\omega)\right| < \theta_l\right\}$. The assertion thus follows. □

F.2 Proof of Lemma 2 part (i)

We know from Lemma 1 that $M_\tau = \tau'$. In particular, this also implies $M_{\tau_D} = \tau' \mid t(1) \neq t(2)$. Notice that $\tau_D \geqslant 2M_{\tau_D}$ is always true. Thus, it follows that $\tau_D \geqslant 2\,\tau' \mid t(1) \neq t(2)$. Therefore, $\mathbb{P}\left(\tau_D = \infty\right) \geqslant \mathbb{P}\left(\tau' = \infty \mid t(1) \neq t(2)\right) = \mathbb{P}\left(\tau' = \infty \mid t(1) = 1,\ t(2) = 2\right) \geqslant \beta$ (Recall from (1) (main text) the definition of β.). The assertion thus follows. □

F.3 Proof of Lemma 2 part (ii)

Throughout this proof, the condition $t(1) = t(2)$ is implicit and we shall avoid writing it explicitly to simplify notation. Let $\Psi_k := X_{1,k} - X_{2,k}$. Consider the following:

$$\begin{aligned}
\mathbb{P}(\tau_I > n) &= \mathbb{P}\left(\bigcap_{l=2}^{n}\left\{\left|\sum_{k=1}^{M_l}\Psi_k\right| \geqslant \theta_{M_l}\right\}\right)\\
&\leqslant \mathbb{P}\left(\left|\sum_{k=1}^{M_n}\Psi_k\right| \geqslant \theta_{M_n}\right)\\
&= \sum_{m=1}^{n}\mathbb{P}\left(\left|\sum_{k=1}^{M_n}\Psi_k\right| \geqslant \theta_{M_n},\ N_1(n) = m\right)\\
&= \sum_{m=1}^{n}\mathbb{P}\left(\left|\sum_{k=1}^{\min(m,n-m)}\Psi_k\right| \geqslant \theta_{\min(m,n-m)},\ N_1(n) = m\right).
\end{aligned}$$

[2] For unbounded rewards, this would hold w.p. 1, not pathwise.

Consider an arbitrary $\kappa \in \left(0, 1/2 - \sqrt{3}/4\right)$. Splitting the above summation three-ways, we obtain

$$\begin{aligned}
\mathbb{P}(\tau_I > n) &\leqslant \sum_{m=1}^{\kappa n} \mathbb{P}\left(N_1(n) = m\right) + \sum_{m=\kappa n}^{(1-\kappa)n} \mathbb{P}\left(\left|\sum_{k=1}^{\min(m,n-m)} \Psi_k\right| \geqslant \theta_{\min(m,n-m)}\right) \\
&\qquad + \sum_{m=(1-\kappa)n}^{n} \mathbb{P}\left(N_1(n) = m\right) \\
&\leqslant \mathbb{P}\left(N_1(n) \leqslant \kappa n\right) + \mathbb{P}\left(N_2(n) \leqslant \kappa n\right) + \sum_{m=\kappa n}^{(1-\kappa)n} \mathbb{P}\left(\left|\sum_{k=1}^{\min(m,n-m)} \Psi_k\right| \geqslant \theta_{\min(m,n-m)}\right) \\
&\leqslant \mathbb{P}\left(N_1(n) \leqslant \kappa n\right) + \mathbb{P}\left(N_2(n) \leqslant \kappa n\right) + 2\sum_{m=\kappa n}^{(1-\kappa)n} \exp\left(\frac{-\theta^2_{\min(m,n-m)}}{2\min(m,n-m)}\right),
\end{aligned}$$

where the last step follows from Hoeffding's inequality [4] using the fact that Ψ_k's are i.i.d. with $\mathbb{E}\Psi_1 = 0$ and $|\Psi_1| \leqslant 1$. Recall that for any $\kappa \in \left(0, 1/2 - \sqrt{3}/4\right)$, part (i) of Theorem 4 guarantees that $\sum_{n=1}^{T} \left(\mathbb{P}\left(N_1(n) \leqslant \kappa n\right) + \mathbb{P}\left(N_2(n) \leqslant \kappa n\right)\right) = \mathcal{O}_T(1)$ (the subscript T is added to indicate that the asymptotic scaling is w.r.t. T), with the limit being a constant that depends on the user-defined parameters (m_0, γ) determining the sequence $(\theta_m)_{m\in\mathbb{N}}$ in (2) (main text). Therefore, we have

$$\sum_{n=1}^{T} \mathbb{P}(\tau_I > n) \leqslant \mathcal{O}_T(1) + 2\sum_{n=1}^{T} \sum_{m=\kappa n}^{(1-\kappa)n} \exp\left(\frac{-\theta^2_{\min(m,n-m)}}{2\min(m,n-m)}\right). \tag{51}$$

To analyze the double-summation term, consider the following:

$$\begin{aligned}
\sum_{m=\kappa n}^{(1-\kappa)n} \exp\left(\frac{-\theta^2_{\min(m,n-m)}}{2\min(m,n-m)}\right) &\leqslant \sum_{m=\kappa n}^{n/2} \exp\left(\frac{-\theta^2_m}{2m}\right) + \sum_{m=n/2}^{(1-\kappa)n} \exp\left(\frac{-\theta^2_{n-m}}{2(n-m)}\right) \\
&\leqslant 2\sum_{m=\kappa n}^{\infty} \exp\left(\frac{-\theta^2_m}{2m}\right) \\
&\leqslant 2\sum_{m=\kappa n}^{\infty} \exp\left(\frac{-\theta^2_{m-m_0}}{2(m-m_0)}\right),
\end{aligned} \tag{52}$$

Notice that

$$\frac{\theta^2_{m-m_0}}{2(m-m_0)} = \left(1 - \frac{m_0}{m}\right)\left(2\log m + (\gamma/2)\log\log m\right) = 2\log m + (\gamma/2)\log\log m + o_m(1), \tag{53}$$

where the last equality follows since m_0 and γ are finite user-defined parameters. Using (52) and (53), we obtain

$$\begin{aligned}
\sum_{m=\kappa n}^{(1-\kappa)n} \exp\left(\frac{-\theta^2_{\min(m,n-m)}}{2\min(m,n-m)}\right) &\leqslant 2\sum_{m=\kappa n}^{\infty} \exp\left(-\left(2\log m + (\gamma/2)\,\mathbf{loglog}\, m + o_m(1)\right)\right) \\
&= 2\sum_{m=\kappa n}^{\infty} \frac{\mathcal{O}_m(1)}{m^2 \left(\log m\right)^{\gamma/2}} \\
&\leqslant \frac{1}{\left(\log n + \log \kappa\right)^{\gamma/2}} \sum_{m=\kappa n}^{\infty} \frac{\mathcal{O}_m(1)}{m^2} \\
&= \mathcal{O}_n\left(\frac{1}{\left(\log n + \log\kappa\right)^{\gamma/2}}\left(\frac{1}{\kappa n} + \frac{1}{\kappa^2 n^2}\right)\right).
\end{aligned} \tag{54}$$

From (51) and (54), it follows that

$$\begin{aligned}
\sum_{n=1}^{T} \mathbb{P}(\tau_I > n) &\leqslant \mathcal{O}_T(1) + \sum_{n=1}^{T} \mathcal{O}_n\left(\frac{1}{\left(\log n + \log\kappa\right)^{\gamma/2}}\left(\frac{1}{\kappa n} + \frac{1}{\kappa^2 n^2}\right)\right) \\
&= \mathcal{O}_T(1),
\end{aligned}$$

where the conclusion in the last step follows since $\gamma > 2$ is a finite user-defined parameter and $\kappa \in \left(0, 1/2 - \sqrt{3}/4\right)$ is arbitrarily chosen. Therefore, the stated assertion that $\mathbb{E}\tau_I < C_0$, where C_0 is some finite constant that depends on (m_0, γ), follows. □

Remark. Part (i) of Theorem 4 has a significant bearing on this result. Specifically, if unlike UCB1, the playing rule does not satisfy a concentration property akin to the one stated in part (i) of Theorem 4, then the $\mathcal{O}_T(1)$ term on the RHS in (51) would instead be $\Omega(T)$.

G The CAB problem with $|\mathcal{T}| = K$

In this section, we extend our results to K-typed settings. Let $\mathcal{T} = \{1, 2, ..., K\}$ and $\mathcal{D}(\mathcal{T})$ denote the distribution over $\mathcal{T}$. The mean reward associated with type $i \in \mathcal{T}$ is denoted by μ_i. We assume that the mean rewards associated with each of the K types are *distinct*[3], and without loss of generality, assume that type 1 is optimal, i.e., $\mu_1 > \mu_i \ \forall \ i \in \mathcal{T}\backslash\{1\}$. The sub-optimality gap of type i is denoted by $\Delta_i := \mu_1 - \mu_i$, and the minimal separation between any pair of types, by $\Delta_0 := \min_{(i,j)\in\mathcal{T}^2 : i\neq j} |\mu_i - \mu_j|$. In § G.1 and § G.2, we propose gap-aware and gap-agnostic algorithms for the CAB problem with K types and state their performance guarantees (without proof).

G.1 A near-optimal gap-aware algorithm

Below, we present a simple fixed-design ETC (Explore-then-Commit) algorithm assuming ex ante knowledge of the duration of play[4] n and a separability parameter $\delta \in (0, \Delta_0]$. It is noteworthy that the informational requirement is significantly greater in the CAB problem compared to its finite-armed counterpart as it assumes knowledge of a lower bound on the minimal separation between any pair of types (Δ_0), instead of the minimal sub-optimality gap ($\min_{i>1}\Delta_i$) which is relatively coarser information ($\because \min_{i>1}\Delta_i > \Delta_0$).

Algorithm 1 ETC-∞(K): ETC for an infinite population of arms with $|\mathcal{T}| = K$.

```
Input: $(n, \delta)$, where $\delta \in (0, \Delta_0]$.
Set epoch length $L = \lceil 2\delta^{-2}\log n \rceil$. Set budget $T = n$.
Initialization: Select two new arms. Call it consideration set $\mathcal{A} = [K]$.
$m \leftarrow \min(L, T/K)$.
Play each arm in $\mathcal{A}$ $m$ times. Update budget: $T \leftarrow T - Km$.
if $|\sum_{k=1}^{m}(X_{i,k} - X_{j,k})| < \delta m$ for any distinct pair $(i,j) \in \mathcal{A}^2$ then
    Permanently discard $\mathcal{A}$ and go to Initialization.
else
    Commit the remaining budget of play to arm $i^* \in \arg\max_{i\in\mathcal{A}} \sum_{k=1}^{m} X_{i,k}$.
```

The stated version of the algorithm does not generalize well to K types. In order to appreciate this, consider the particular case of a uniform distribution over $\mathcal{T}$. In this case, a new arm is equally likely to be any one of the K possible types and therefore, it would take the algorithm K^K fresh draws in expectation of size K consideration sets in order to obtain one that is fully heterogeneous (contains one arm of each type). Thus, the expected cumulative regret would grow proportionally to K^K which is unacceptable. This can be improved to an $\mathcal{O}(K\log K)$ dependence on the number of types by suitable tweaks of the algorithm. Specifically, a natural modification would be to start with a consideration set containing a single arm and augmenting it sequentially by adding new arms (one at a time) that are sufficiently separated from each arm in the set. Instructively, the stopping point would occur when the algorithm accumulates K arms that are all different from each other, after which it would simply commit the residual sampling budget to the empirically best arm. One can show that the improvement due to said modification is significant in the sense that the regret of the modified algorithm scales as $\mathcal{O}\left((\log K)\left(\sum_{i=2}^{K}\Delta_i\right)\delta^{-2}\log n\right)$. However, the analysis of the modified algorithm, albeit similar in spirit to the regret analysis of Algorithm 1 (main text), is quite tedious and becomes further so if one were to consider generic K-point distributions over $\mathcal{T}$ (instead of the uniform distribution) and is therefore omitted from this text.

[3] This is a critical assumption, which if violated will cause our algorithms to incur linear regret.

[4] The standard exponential doubling trick can be employed to make the algorithm horizon-free, cf. [2].

G.2 A near-optimal gap-agnostic algorithm

Below, we present a generalization of our framework for the CAB problem with a binary $\mathcal{T}$ (ALG$(\pi, \Theta, 2)$, main text), adapted to an arbitrary finite cardinality $\mathcal{T}$.

Algorithm 2 ALG(Ξ, Θ, K): An algorithmic framework for countable-armed bandits with $|\mathcal{T}| = 2$.

```
**Input:** A $\Delta$-agnostic playing rule $\Xi$, a deterministic sequence $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ in $\mathbb{R}$.
**Initialization** (Starts a new epoch)**:** Select $K$ *new* arms. Call it consideration set $\mathcal{A} = [K]$.
For $s \in [K]$, play each arm in $\mathcal{A}$ once.
$m \leftarrow 1$.
**for** $s \in \{K+1, K+2, ...\}$ **do**
    **if** $|\sum_{k=1}^{m} (X_{i,k} - X_{j,k})| < \theta_m$ for any distinct pair $(i, j) \in \mathcal{A}^2$ **then**
        Permanently discard $\mathcal{A}$ and go to **Initialization**.
    **else**
        Play an arm from $\mathcal{A}$ according to $\Xi$.
        $m \leftarrow \min_{i \in \mathcal{A}} N_i(s)$.
```

Proposition 1 (Lower bound on the true negative rate) *For each $i \in \mathcal{T} = [K]$, let $\left(Y_{i,k}^{F_i}\right)_{k \in \mathbb{N}}$ denote an i.i.d. sequence of random variables with distribution $F_i \in \mathcal{G}(\mu_i)$ satisfying Assumption 1 (main text). Let $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ be a deterministic non-negative real-valued sequence such that $\{(\theta_m/m) : m = 1, 2, ...\}$ is monotone decreasing in m with $\theta_1 < 1$ and $\theta_m = o(m)$. For each $(i, j) \in [K]^2$ s.t. $i < j$, define the following stopping time:*

$$\tau_{i,j} := \inf \left\{ n \in \mathbb{N} : \left| \sum_{k=1}^{n} \left(Y_{i,k}^{F_i} - Y_{j,k}^{F_j} \right) \right| < \theta_n \right\}.$$

Then,

$$\widetilde{\beta} := \min_{F_1 \in \mathcal{G}(\mu_1), ..., F_K \in \mathcal{G}(\mu_K)} \mathbb{P} \left(\min_{(i,j) \in [K]^2 : i < j} \tau_{i,j} = \infty \right) > 0. \tag{55}$$

Remark. If $K = 2$, then $\widetilde{\beta} = \beta$, where β is as defined in (1) (main text).

Proposition 2 (Upper bound on the expected regret of ALG$(\mathbf{UCB1}, \Theta, K)$) *Consider the input sequence $\Theta \equiv \{\theta_m : m = 1, 2, ...\}$ given by*

$$\theta_m := \sqrt{m^2 (m + m_0)^{-1} \left(4 \log(m + m_0) + \gamma \log\log(m + m_0)\right)}, \tag{56}$$

where $m_0 \geqslant 0$ and $\gamma > 2$ are user-defined parameters that ensure Θ satisfies the conditions of Proposition 1 (for example, $m_0 = 11$ and $\gamma = 2.1$ is an acceptable configuration). Suppose that Assumption 1 (main text) is satisfied. Then, the expected cumulative regret of $\pi = ALG\ (UCB1, \Theta, K)$ after any number of plays n is bounded as follows:

$$\mathbb{E}R_n^{\pi} \leqslant \min \left[\left(\max_{i \in \{2, ..., K\}} \Delta_i \right) n,\ 8\widetilde{\beta}^{-1} \left(\sum_{i=2}^{K} \Delta_i^{-1} \right) \log n + \left(C_1 + \alpha^{-1} C_2 \right) \widetilde{\beta}^{-1} \left(\sum_{i=2}^{K} \Delta_i \right) \right], \tag{57}$$

where $\widetilde{\beta}$ is as defined in (55)*, $\Delta_i = \mu_1 - \mu_i > 0$ for $i \in \{2, ..., K\}$, C_1 is an absolute constant and C_2 is a constant that depends only on the free parameters of the algorithm, namely (m_0, γ).*

Remark. The constants C_1, C_2 above are the same as the ones appearing in Theorem 3 (main text).

References

[1] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 2-3 (2002), 235–256.

[2] BESSON, L., AND KAUFMANN, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971* (2018).

[3] GRIMMETT, G., GRIMMETT, G. R., STIRZAKER, D., ET AL. *Probability and random processes*. Oxford university press, 2001.

[4] HOEFFDING, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 301 (1963), 13–30.

[5] KHINTCHINE, A. über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae 6*, 1 (1924), 9–20.

[6] LAI, T. L., AND ROBBINS, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*, 1 (1985), 4–22.

[7] LATTIMORE, T., AND SZEPESVÁRI, C. Bandit algorithms.