Journal of Machine Learning Research 27 (2026) 1-25 Submitted 10/25; Revised 11/25; Published 6/26

Adaptive Algorithms for Infinitely Many-Armed Bandits: A Unified Framework

Emmanuel Pilliat EMMANUEL.PILLIAT@ENSAI.FR
Univ Rennes, Ensai, CNRS, CREST—UMR 9194
France

Editor: Kevin Jamieson

Abstract

We consider a bandit problem where the budget is smaller than the number of arms, which may be infinite. In this regime, the usual objective in the literature is to minimize simple regret. To analyze broad classes of distributions with potentially unbounded support, where simple regret may not be well-defined, we take a slightly different approach and seek to maximize the expected simple reward of the recommended arm, providing anytime guarantees. To that end, we introduce a distribution-free algorithm, OSE, that adapts to the distribution of arm means and matches the known minimax lower bounds, up to polylogarithmic factors, in the classical bounded settings, while applying to several broader distribution classes. We characterize the sample complexity through the rank-corrected inverse squared gap function. In particular, we recover known upper bounds and transition regimes for α less or greater than $1/2$ when the quantile function is $\lambda_\eta = 1 - \eta^\alpha$. We additionally identify new transition regimes depending on the noise level relative to α, which we conjecture to be nearly optimal. Additionally, we introduce an enhanced practical version, PROSE, that achieves state-of-the-art empirical performance for the main distribution classes considered in the literature.

Keywords: multi-armed bandits, infinitely many arms, simple regret, adaptive algorithms, distribution-free learning

1. Introduction

In this manuscript, we consider a problem where the learner has a very limited budget to pull arms from a reservoir of large size, potentially infinite. In contrast to classical bandit problems where the budget vastly exceeds the number of arms, there is no hope of finding the best arm as it is impossible to explore all arms. Nevertheless, when not all arms are equivalent and a non-negligible proportion of arms provide significantly better rewards on average, it is possible to detect some of them without exhaustive exploration. This setting is of considerable practical interest, as infinitely many-armed bandit problems arise naturally when available options vastly exceed feasible trials. Applications include content recommendation, where tens of thousands of new podcasts are released monthly (Aziz et al., 2022), hyperparameter optimization over large configuration spaces (Li et al., 2018), and crowdsourcing tasks like The New Yorker Cartoon Caption Contest with thousands of submissions (Jain et al., 2020). Additional domains include labor markets and resource exploration (Wang et al., 2008). The common challenge is identifying high-quality arms from a large set using a limited sampling budget.

Formally, we consider an infinite and countable set of arms $\mathcal{A}$ from which a learner iteratively chooses which arm to pull. We assume that each arm $a \in \mathcal{A}$ produces rewards $(X_{a,s})_{s\geq 1}$ distributed according to

$$X_{a,s} = \lambda_{\gamma(a)} + \varepsilon_{a,s} \quad , \tag{1}$$

where $(\gamma(a))_{a\geq 1}$ are independent and uniformly distributed in $[0,1]$, and $\eta \mapsto \lambda_\eta$ is a nonincreasing and right-continuous function from $(0,1]$ to $\mathbb{R}$, which corresponds to the quantile function of the arm means distribution. The noise variables $(\varepsilon_{a,s})_{s\geq 1}$ are assumed to be independent, centered, and ζ^2-sub-Gaussian, that is $\mathbb{E}[e^{\theta\varepsilon_{a,s}}] \leq \exp(\zeta^2\theta^2/2)$ for any $\theta \in \mathbb{R}$. We say that $\gamma(a)$ is the rank of arm a, and that a_1 is better than a_2 if $\gamma(a_1) \leq \gamma(a_2)$.

The model defined by (1) is equivalent to assuming that each arm has a random mean sampled from an unknown distribution $\mathcal{D}$, referred to as the arm reservoir distribution in Carpentier and Valko (2015). Indeed, the function $\eta \mapsto \lambda_\eta$ precisely characterizes the distribution $\mathcal{D}$ of the arm means in $\mathcal{A}$. Starting from any distribution $\mathcal{D}$ on $\mathbb{R}$, it is always possible to define a nonincreasing and right-continuous function λ such that $\lambda_{\gamma(a)}$ has distribution $\mathcal{D}$. If F is the cumulative distribution function (CDF) of $\mathcal{D}$, then the function $\lambda_\eta = \inf\{x \in \mathbb{R} : F(x) \geq 1-\eta\}$ is nonincreasing and right-continuous, and $\eta \mapsto \lambda_{1-\eta}$ is the inverse CDF of F. Since $\gamma(a)$ and $1-\gamma(a) \sim \mathcal{U}(0,1)$ are uniform random variables in $[0,1]$,

$$\mathbb{P}(\lambda_{\gamma(a)} \leq x) = \mathbb{P}(\lambda_{1-\gamma(a)} \leq x) = \mathbb{P}(1-\gamma(a) \leq F(x)) = F(x) \quad ,$$

which proves that $(\lambda_{\gamma(a)})_{a\in\mathcal{A}}$ are distributed according to $\mathcal{D}$. Notice that we do not assume $\mathcal{D}$ is bounded, since $\lambda_\eta \to +\infty$ as $\eta \to 0$ is allowed. From this observation, we say that arm a is a top-ρ arm if its rank satisfies $\gamma(a) \leq \rho$.

At each timestep t, the learner selects an arm $\hat{a}_t \in \mathcal{A}$ to pull, with this decision informed by previous observations. Upon pulling the chosen arm, the learner receives the observation

$$X_t = \lambda_{\gamma(\hat{a}_t)} + \varepsilon_t \quad . \tag{2}$$

Equivalently, we can express this using the counting variable $N_{a,t} = \sum_{t'=1}^{t} \mathbf{1}\{\hat{a}_{t'} = a\}$, which tracks how many times arm a has been pulled up to time t. With this notation, we can write the observation as $X_t = X_{a,N_{a,t}}$, connecting to the formulation in (1).

In addition to pulling an arm, the learner must provide a *recommendation* $\hat{r}_t \in \mathcal{A}$. Since $\mathcal{A}$ is infinite, identifying a best arm satisfying $\gamma(a) = 0$ is infeasible. Hence, one of the main objectives in this manuscript is to provide a strategy that maximizes the expected simple recommendation reward with high probability, defined as $\lambda_{\gamma(\hat{r}_t)}$. This quantity is well-defined even when $\mathcal{D}$ is unbounded. In the literature, $\mathcal{D}$ is usually assumed to be bounded, and the main objectives are either to minimize the simple regret $\lambda_0 - \lambda_{\gamma(\hat{r}_t)}$ or to maximize the probability of recommending an ϵ-good arm $\mathbb{P}(\lambda_{\gamma(\hat{r}_t)} \geq \lambda_0 - \epsilon)$ (Bubeck et al., 2009; Katz-Samuels and Jamieson, 2020; Carpentier and Valko, 2015; Zhao et al., 2023; Jamieson et al., 2013). These quantities are well-defined only if the distribution $\mathcal{D}$ is bounded, or equivalently if $\lambda_0 < \infty$. Note that we do not consider cumulative regret or cumulative reward here since De Heide et al. (2021) showed that it is impossible to adapt to $\mathcal{D}$, even for very simple classes of distributions.

Our goal is to design a strategy that maximizes the expected simple recommendation reward $\lambda_{\gamma(\hat{r}_t)}$ without prior knowledge of the distribution $\mathcal{D}$ or of $\eta \mapsto \lambda_\eta$. To handle this lack

of information, we require our strategy to be adaptive to the unknown distribution $\mathcal{D}$. We characterize this adaptivity through the ranks of the arms: recommendations $\hat{r}_t$ should have ranks $\gamma(\hat{r}_t)$ that are as small as possible. This leads us to the following formal definition of achievable rank sequences.

Definition 1 (δ-achievable rank sequence) *An algorithm producing recommendations $(\hat{r}_t)$ achieves a sequence of ranks (η_t) with confidence $1-\delta$ if*

$$\mathbb{P}(\gamma(\hat{r}_t) \le \eta_t \text{ for all } t \ge 1) \ge 1-\delta \ . \tag{3}$$

When such an algorithm exists, we say the rank sequence (η_t) is δ-achievable.

Thus, if recommendations $(\hat{r}_t)$ achieve rank sequence (η_t), then each recommended arm $\hat{r}_t$ has mean reward in the top η_t fraction of all arms. Since t is not predetermined, this corresponds to an anytime guarantee in the classical bandit literature (Lattimore, 2016; Jun and Nowak, 2016; Degenne and Perchet, 2016). A sequence $(\eta_t)_{t\ge1}$ is called a *best δ-achievable sequence* if it is δ-achievable and satisfies $\eta_t \le \eta'_t$ for all $t \ge 1$ whenever $(\eta'_t)_{t\ge1}$ is another δ-achievable sequence. Given a fixed confidence parameter $\delta \in (0,1)$, our primary goal is to derive an upper bound that holds for any best δ-achievable sequence of ranks. This upper bound will directly allow us to derive a lower bound on the expected simple recommendation reward $\lambda_{\gamma(\hat{r}_t)}$.

1.1 Related Work

Significant work has been done under assumptions on the distribution $\mathcal{D}$ of arm means, with a focus on the case where $\mathcal{D}$ is uniform (Berry et al., 1997; Bonald and Proutiere, 2013), or more generally when $\mathcal{D}$ is bounded and $\mathbb{P}(\lambda_0 - \lambda_{\gamma(a)} \ge \epsilon) \asymp \epsilon^\beta$ for some unknown $\beta > 0$ (Carpentier and Valko, 2015; Wang et al., 2008). In the latter case, Carpentier and Valko (2015) established that the minimax rate for simple regret is of order $\frac{1}{\sqrt{t}} \vee \frac{1}{t^{1/\beta}}$ (up to polylogarithmic factors) under standard Gaussian noise.

Recently, the focus has shifted toward distribution-free methods that make minimal assumptions about the underlying reward distribution $\mathcal{D}$. Rather than assuming parametric forms or specific tail behaviors, these approaches develop algorithms and theoretical guarantees that hold for broad classes of distributions (Katz-Samuels and Jamieson, 2020; Zhao et al., 2023; Li et al., 2018; Mason et al., 2020). A key technique is the bracketing trick introduced by Katz-Samuels and Jamieson (2020), which consists of progressively increasing the size of the arm set under consideration.

Most related to our work are the contributions of Zhao et al. (2023). They consider a fixed-budget T setting where the number of arms K may be significantly larger than the budget. To derive strong performance guarantees, they combine several existing techniques. Starting with the Sequential Halving (SH) algorithm (Karnin et al., 2013), they adapt it into an anytime algorithm called DSH using the doubling trick—for general discussions on this trick, see for instance Besson and Kaufmann (2018). This yields an algorithm with good guarantees when $K \le T$. To adapt to the data-poor regime where $K \gg T$, they incorporate the bracketing trick from Katz-Samuels and Jamieson (2020). This leads to their final method **BSH** and their main result (Theorem 6): an anytime upper bound on the probability of recommending a suboptimal arm, which holds uniformly over all error levels $\epsilon > 0$.

1.2 Contributions

First, we take a different approach from the literature by making no assumptions on the distribution of arm means—its support can be unbounded—or on the noise level. Our main criterion is therefore not simple regret but rather simple recommendation reward, where good performance means recommending arms a with small rank $\gamma(a)$. We characterize the achievable recommendation reward through the sample complexity of obtaining good rank sequences, formalized via a rank-corrected inverse squared gap function in (4).

Second, we connect our main result to the existing literature on many-armed bandits. We recover the distribution-dependent upper bound on simple regret from Zhao et al. (2023) up to polylogarithmic factors, which is essentially tight for Polynomial(α) discrete distributions (characterized by $\lambda_\eta = 1 - \sum_{i=1}^{K} (\frac{i}{K})^\alpha \mathbf{1}\{i < K\eta \leq i+1\}$), as shown in their Corollary 6.2. For the continuous counterpart—Beta$(1, 1/\alpha)$ distributions—we recover the minimax optimal bound (up to polylogarithmic factors) from Carpentier and Valko (2015).

Beyond existing rates, we identify new transition phases for Beta$(1, 1/\alpha)$ distributions that reveal the dependence on α and noise level ζ typically hidden in prior work. Specifically, our upper bounds exhibit a transition phase at timesteps around $t \asymp \left(\frac{\zeta^2}{\alpha^2}\right)^{\frac{1}{1-2\alpha}}$, which we interpret as a saturation effect in Section 4.

Third, we analyze Pareto-$1/\alpha$ distributions, a class of unbounded distributions for which simple regret is undefined. When the noise level satisfies $\zeta > 1$, our upper bound reveals transition regimes: the noise becomes negligible for identifying good arms once $t \gg \zeta^{1/\alpha}$. Before this threshold, when $\alpha < 1/2$, we establish that the expected simple recommendation reward scales at least as $\left(\frac{t}{\zeta^2}\right)^{\frac{\alpha}{1-2\alpha}}$.

Finally, we introduce **PROSE**, a practical version of our main algorithm **OSE**. Empirical comparisons with **BSH** from Zhao et al. (2023) on Polynomial(α) instances demonstrate that **PROSE** achieves strong simple regret performance with high computational efficiency. Moreover, unlike **BSH**, **PROSE** matches the long-term cumulative regret performance of the classical UCB algorithm.

1.3 Technical Overview

From a theoretical perspective, our main technical achievement is to provide a unified framework for many-armed bandits through the ranks of the arms, the quantile function, and the rank-corrected inverse squared gap function.

Specifically, our main result enables us to recover previous results (up to polylogarithmic factors) in terms of simple regret or probability of recommending an ϵ-good arm, and to analyze new settings where the distribution is potentially unbounded. From our main result, we characterize the order of magnitude of the simple regret and rank recommendation achieved by **OSE** for several classically studied distributions; in the classical bounded settings this matches the minimax-optimal rates (up to polylogarithmic factors) of Carpentier and Valko (2015) and Zhao et al. (2023). It also allows us to present upper bounds in new settings with unbounded distributions, where we conjecture the bounds to be nearly tight. Moreover, our upper bounds provide explicit dependencies on the noise level ζ and the distribution parameter.

The proof of our main theorem relies mainly on concentration bounds for, on the one hand, the number of top-ρ arms and, on the other hand, uniform concentration bounds on the noise of explored arms. To focus on the order of magnitude related to the distribution of arm means rather than on polylogarithmic factors—which could be of interest for future work—we do not use refined concentration bounds such as the iterated logarithm law for sums of sub-Gaussian variables, but rather rely on standard Hoeffding and Bernstein inequalities.

From a practical perspective, we provide a computationally efficient implementation of **PROSE**. This addresses a significant technical challenge: a naïve implementation requires quadratic time complexity in the number of timesteps (since we must compute a maximum at each step). To overcome this efficiency bottleneck, we develop optimized code that computes recommendations by iteratively updating a partition of sorted sets and performing cascading operations between these sets, achieving $O(\log^2 t)$ amortized complexity per iteration. Beyond statistical performance, we also compare the computational efficiency of **PROSE** with an optimized Julia implementation of the bracketing sequential halving with doubling trick (**BSH**) from Zhao et al. (2023). Implementations of **PROSE** and **BSH** are available on the author's GitHub webpage.

2. Main Result

Recommending a good arm amounts to identifying an arm with a small rank $\gamma(a)$. To quantify the difficulty of this task, we introduce the rank-corrected inverse squared gap function G for any pair of ranks $\rho < \nu$:

$$G(\rho, \nu) = \frac{\zeta^2 \nu}{\rho(\lambda_\rho - \lambda_\nu)^2} \vee \frac{1}{\rho} \,. \tag{4}$$

If $G(\rho, \nu) \le t/\psi$ for some $\psi > 0$, we say that rank ρ is ψ-significant at time t with respect to rank ν. $G(\rho, \nu)$ can be seen as a measure of the sample complexity for detecting a top-ρ arm among other bad arms a of rank of order ν, which happens for example when $\gamma(a) \in [\nu, 2\nu]$, and it captures three main effects. First, $\zeta^2/(\lambda_\rho - \lambda_\nu)^2$ represents the squared inverse gap between the top ρ-quantile and the top ν-quantile of $\mathcal{D}$. Consider two arms a_1 and a_2 satisfying $\gamma(a_1) \le \rho$ and $\gamma(a_2) \ge \nu$. The sample complexity for detecting with high probability that a_1 is better than a_2 is of order $\zeta^2/(\lambda_\rho - \lambda_\nu)^2$, up to logarithmic factors. Second, the scale factor $\frac{\nu}{\rho}$ accounts for potential needle-in-a-haystack effects. Even when the gap between top ρ and ν quantiles is significant, identifying a top-ρ arm with respect to arms of rank larger than ν becomes considerably more challenging when $\rho \ll \nu$. This increased difficulty stems from the relative scarcity of top-ρ arms: for every arm a satisfying $\gamma(a) \le \rho$, there exist on average ν/ρ arms satisfying $\gamma(a) \in [\nu, 2\nu]$. Finally, the term of order $1/\rho$ captures a fundamental lower bound: regardless of how large λ_ρ is relative to λ_ν, we require at least $\Omega(1/\rho)$ samples to observe at least one top-ρ arm.

Now that G characterizes the sample complexity order for detecting top-ρ arms among ν-bad arms, we can define the sample complexity for detecting a top-η arm among all other arms.

Definition 2 (ψ-significant rank at timestep t) *We define the sample complexity function for detecting a top-η arm as the nonincreasing function $S : (0,1] \mapsto \mathbb{R}$:*

$$S(\eta) = \inf_{\rho<\eta} \sup_{\nu\geq\eta} G(\rho,\nu) \ . \tag{5}$$

Moreover, for $\psi > 0$, we say that rank η is ψ-significant at time t if $S(\eta) \leq t/\psi$.

In other words, a rank η is ψ-significant at time t if there exists $\rho \leq \eta$ such that rank ρ is ψ-significant with respect to any other rank $\nu \geq \eta$. This leads us to define $\eta_t^*(\psi)$ as the smallest rank that is ψ-significant at time t:

$$\eta_t^*(\psi) = \inf\left\{\eta \in (0,1) : \ S(\eta) \leq \frac{t}{\psi}\right\} \ . \tag{6}$$

Hence, a rank η is ψ-significant at timestep t if and only if $\eta \geq \eta_t^*(\psi)$. The following theorem establishes that if we set $\psi := \psi_{t,\delta}$ as a large polylogarithmic factor in t/δ, then any sequence of $\psi_{t,\delta}$-significant ranks is achievable by some algorithm with probability at least $1-\delta$, in the sense of Definition 1.

Theorem 3 *Fix any $\delta \in (0,1)$ and $t \geq 1$. Assume that $\psi := \psi_{t,\delta} \geq 2^{30}\log^4(5t/\delta)$, and set the tuning parameter $\beta_t = 6\log(5t/\delta)$. With probability $1-\delta$, for all $t \geq 1$, the recommendation $\hat{r}_t$ of Algorithm 1 (***OSE***) defined in Section 3 satisfies $\gamma(\hat{r}_t) \leq \eta_t^*(\psi)$. This implies in particular that $\lambda_{\gamma(\hat{r}_t)} \geq \lambda_{\eta_t^*(\psi)}$.*

The proof of Theorem 3 is deferred to Appendix A. Hence, the recommendations of **OSE** are among the top $\eta_t^*(\psi)$ proportion of all arms with high probability at least $1-\delta$. To better understand the detectability condition characterized by $\eta_t^*(\psi)$, we now present a slightly relaxed version of Theorem 3.

Corollary 4 *For $\psi := \psi_{t,\delta} \geq 2^{30}\log^4(5t/\delta)$, we define $\tilde{S}$ and $\tilde{\eta}$ as follows:*

$$\tilde{S}(\eta) = \sup_{\nu\geq 2\eta} G(\eta,\nu) \quad \text{and} \quad \tilde{\eta}_t(\psi) = 2\inf\{\eta \in (0,\tfrac{1}{2}) : \ \tilde{S}(\eta) \leq \tfrac{t}{\psi}\} \ . \tag{7}$$

Then, **OSE** *also achieves the sequence $(\tilde{\eta}_t(\psi))_{t\geq 1}$ with probability at least $1-\delta$.*

Corollary 4 can be directly deduced from Theorem 3. Indeed, $S(2\eta) \leq \tilde{S}(\eta)$ for any $\eta \in (0,1/2)$, so that $\eta_t^* \leq \tilde{\eta}_t$. Hence, with probability at least $1-\delta$, $\gamma(\hat{r}_t) \leq \eta_t^* \leq \tilde{\eta}_t$. We refer the reader to Section 4 for direct consequences of Theorem 3. These include results for specific distribution classes $\mathcal{D}$: we recover the minimax simple regret upper bound of Carpentier and Valko (2015) (up to polylogarithmic factors) as a special case when $1-\lambda_\eta \asymp \eta^\alpha$, and match the more general result of Zhao et al. (2023) on ϵ-good arm identification up to a $\log(1/\delta)$ factor.

In Section 4, we show that Theorem 3 recovers the result of Zhao et al. (2023) up to a polylogarithmic factor in t/δ. Our polylogarithmic term is of order $\log^4(t/\delta)$ and is suboptimal: a more careful analysis could reduce these exponents. However, we accept this loss of polylogarithmic precision in exchange for clearer exposition and simpler analysis. This simplification allows us to highlight the main effect captured by the function G while maintaining practical relevance. Indeed, setting tuning parameters to constants instead of polylogarithmic expressions still achieves state-of-the-art performance in practice with our enhanced algorithm **PROSE**, described after **OSE** in Section 3.

3. Algorithms

Algorithm **OSE**, which achieves Theorem 3, proceeds as follows. At each time t, we define an exploration scope—a "bracket" in the terminology of Katz-Samuels and Jamieson (2020); Zhao et al. (2023)—consisting of arms eligible for pulling. We pull the arm maximizing the UCB within this scope and recommend the arm with the best LCB overall. **PROSE** enhances **OSE** by ranking arms according to their LCBs and restricting exploration scopes to high-LCB arms. **PROSE** yields significantly better empirical performance for a large class of distributions (Section 5), making it our recommended algorithm in practice.

Let $\mathcal{O}_t \subset \mathcal{A} = \{1, 2, \dots\}$ denote the set of arms observed before time t. At each timestep t, we either pull an arm from $\mathcal{O}_t$ or observe a new arm a. Without loss of generality, we assume $a \le t$ for any new arm. We denote by $\overline{X}_{a,t} = \frac{1}{N_{a,t}} \sum_{s=1}^{N_{a,t}} X_{a,s}$ the empirical mean of rewards obtained from arm a up to time t. For a given sequence of tuning parameters $\beta_t > 0$, we define the upper and lower confidence bounds as

$$\mathrm{UCB}_{a,t} = \overline{X}_{a,t} + \sqrt{\frac{\zeta^2 \beta_t}{N_{a,t}}} \quad \text{and} \quad \mathrm{LCB}_{a,t} = \overline{X}_{a,t} - \sqrt{\frac{\zeta^2 \beta_t}{N_{a,t}}} \ . \tag{8}$$

By convention, we set $\mathrm{UCB}_{a,t} = +\infty$ and $\mathrm{LCB}_{a,t} = -\infty$ for unsampled arms (that is when $N_{a,t} = 0$ or, equivalently, when $a > |\mathcal{O}_t|$). Algorithm 1 presents our **OSE** procedure. At each time step $t > 0$, we sample $U \sim \mathrm{Uniform}[0, 1]$ and set $Z = \lfloor t^U \rfloor$ to define the exploration scope size. We pull the arm $\hat{a}_t$ maximizing the UCB within $\{1, \dots, Z\}$. When $Z > |\mathcal{O}_t|$, this mechanism naturally explores a new arm. We recommend the arm with the highest LCB among all observed arms.

In Theorem 3, we set $\beta_t = 6 \log(5t/\delta)$ for a given failure probability $\delta > 0$. The law of the iterated logarithm for sums of sub-Gaussian random variables (see for instance, Lemma 5 of Verzelen et al. (2023)) would allow us to set β_t of order $\log(\log(t)/\delta)$, yielding slightly improved polylogarithmic factors as in the BUCB algorithm (Katz-Samuels and Jamieson, 2020). However, we opt for the simpler logarithmic form, which streamlines the proof and aligns with our practical implementation using a constant $\beta_t := \beta$ (see Section 5). Empirically, for **PROSE**, the results are not very sensitive to the choice of β.

Input: A set of arms $\mathcal{A}$, a sequence of tuning parameters $\beta_t > 0$
Output: A sequence of recommended arms $(\hat{r}_t)$
for $t = 1, 2, \dots$ **do**
  Generate an independent $U \sim \mathcal{U}(0, 1)$;
  Set $Z = \lfloor t^U \rfloor$; `// exploration scope`
  Pull arm $\hat{a}_t = \arg\max_{a \le Z} \mathrm{UCB}_{a,t-1}$; `// optimistic arm`
  Recommend $\hat{r}_t = \arg\max_{a \in \mathcal{A}} \mathrm{LCB}_{a,t}$; `// recommendation`
end

Algorithm 1: Optimistic Scope Exploration (**OSE**)

OSE shares strong similarities with Bracketing-UCB (**BUCB**) from Katz-Samuels and Jamieson (2020), which pulls arms maximizing the UCB over independently sampled subsets of exponentially growing sizes. The key difference is that **OSE** uses nested subsets $\{1, \dots, Z\}$ with random $Z = \lfloor t^U \rfloor$ for U uniform in $[0, 1]$, whereas **BUCB** independently samples new subsets at each bracket level with deterministic exponentially growing sizes. We adopt this nested structure because it simplifies the algorithm's presentation and naturally leads

to **PROSE**, which achieves substantially better empirical performance. Note that **BSH** (bracketing sequential halving) from Zhao et al. (2023) uses a similar bracketing trick as **BUCB** from Katz-Samuels and Jamieson (2020) by also iterating over disjoint subsets (brackets) of exponentially growing sizes. **OSE**, together with **BSH** and **BUCB**, has two main practical drawbacks. First, they continue pulling arms with small LCBs. Since $Z = 1$ occurs with probability at least $1/\log(t)$, approximately $t/\log(t)$ samples are wasted on arm 1 even when it is clearly suboptimal. Second, the random exploration scope Z has two negative consequences: it increases the variance of $\lambda_{\hat{r}_t}$ and prevents efficient computation. Computing the argmax of UCBs at each time step results in $O(t^2)$ computational complexity rather than linear.

To address the first drawback, we rerank arms at each step by their LCBs, prioritizing arms identified as promising over randomly chosen ones. For the second drawback, we use deterministic exploration scopes of exponentially growing sizes. In **PROSE**, we loop over sets of sizes $\lfloor t^{Q(j/\log(t))} \rfloor$, where $Q : [0,1] \to [0,1]$ is a quantile function. Setting $Q(x) = x$ (corresponding to $\mathcal{U}(0,1)$) yields scopes of size $\lfloor e^j \rfloor$, which we recommend for large arm counts. When more aggressive exploration strategies are desired, for example if the noise is small so that good arms are easily identifiable, we recommend choosing $Q(x) = x^{1/\gamma}$ with $\gamma > 1$ (the Beta$(\gamma, 1)$ quantile function). Large γ values make $Q(j/\log(t)) \approx 1$, approximating standard UCB by considering nearly all arms at each step.

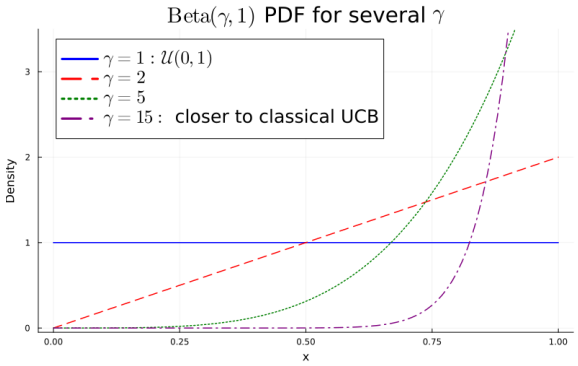

Figure 1: Probability density functions of Beta$(\gamma, 1)$ distributions for various values of γ. Higher values of γ concentrate the distribution toward 1, corresponding to more aggressive exploration strategies.

Input: A set of arms $\mathcal{A}$, a constant tuning parameter $\beta > 0$ to compute the UCBs and LCBs and a function $Q : [0,1] \to [0,1]$
Output: A sequence of recommended arms $(\hat{r}_t)$
Initialize $j = 1$ and permutation $\pi = \mathrm{id}$ representing the current ranking of the arms;
for $t = 1, 2, \ldots$ **do**
  if $j > \log(t)$ **then**
    $j = 1$;
  end
  Set $Z = t^{Q(j/\log(t))}$; // exploration scope
  Pull arm $\hat{a}_t = \arg\max_{\pi(a) \le Z} \mathrm{UCB}_{a,t-1}$; // optimistic arm
  Update π so that $\mathrm{LCB}_{\pi^{-1}(1),t} \ge \mathrm{LCB}_{\pi^{-1}(2),t} \ge \ldots$; // Ranking
  Recommend $\hat{r}_t = \arg\max_{a \in \mathcal{A}} \mathrm{LCB}_{a,t} = \mathrm{LCB}_{\pi^{-1}(1),t}$; // recommendation
  $j = j + 1$;
end

Algorithm 2: Progressive Ranking for Optimistic Scope Exploration (**PROSE**)

While Algorithm 2 has a naive computational complexity of $O(t \log t)$ per iteration due to re-sorting all arms by their LCBs in decreasing order and computing the maximum UCB within the exploration scope, we implement in our numerical study a significantly more efficient version with $O(\log^2 t)$ amortized complexity per iteration. The key optimization is to maintain arms in sorted order by LCB in decreasing order using incremental binary search insertions rather than full re-sorting. Additionally, we partition arms into $J \approx \log t$ differential sets: the j-th set contains arms whose LCB ranks fall in the interval $(Z_{j-1}, Z_j]$, where $Z_j = \lfloor t^{Q(j/\log t)} \rfloor$, with each set internally sorted by UCB in decreasing order. Finding the arm with maximum UCB within scope Z_j then requires only $O(J)$ comparisons among the top arms of each set. When pulling an arm changes its confidence bounds and shifts its LCB rank from one interval to another, we perform cascading updates by moving the boundary arms between adjacent differential sets to maintain consistency.

4. Implications of the Main Theorem

To fully appreciate the scope of Theorem 3, we first demonstrate how it generalizes the main result from Zhao et al. (2023) (up to polylogarithmic factors) before examining specific cases that make distributional tail assumptions, as in (Jamieson et al., 2013; Carpentier and Valko, 2015).

4.1 Uniform ϵ-error probability bound

To compare with Zhao et al. (2023), we assume that the distribution of arm means is bounded, which translates to $\lambda_0 < +\infty$. We define the gap between λ_0 and λ_η as $\Delta_\eta := \lambda_0 - \lambda_\eta$. We also define $g(\epsilon) = \sup\{\eta \in [0,1] : \Delta_\eta \le \epsilon\}$, which represents the largest rank η whose corresponding gap is at most ϵ. This leads us to define the sample complexity $H(\epsilon)$ as

$$H(\epsilon) = \frac{1}{g(\epsilon/2)} \vee \sup_{\eta \ge g(\epsilon)} \frac{\zeta^2 \eta}{g(\epsilon/2)\Delta_\eta^2} .$$

The supremum of $\frac{\zeta^2\eta}{g(\epsilon/2)\Delta_\eta^2}$ over all $\eta \geq g(\epsilon)$ is the continuous analogue of the function H_2 from Zhao et al. (2023). The additional term $1/g(\epsilon/2)$ extends our analysis to cases where $\lambda_0 > \zeta$, allowing for arm means that are potentially very large relative to the noise level. Note that Zhao et al. (2023) restricts their analysis to $\lambda_0 \leq \zeta = 1$, making the term $1/g(\epsilon/2)$ redundant in their setting. Indeed, when $\lambda_0 \leq \zeta = 1$, we have $\frac{\zeta^2}{\Delta_1^2} \geq 1$, which implies $\frac{1}{g(\epsilon/2)} \leq \frac{\zeta^2\eta}{g(\epsilon/2)\Delta_\eta^2}$ for $\eta = 1$. While our bound does not explicitly contain the $1/\epsilon^2$ term appearing in the trivial regime of Theorem 6 in Zhao et al. (2023), this term is implicitly captured by $H(\epsilon)$. Indeed, since $g(\epsilon/2) \leq g(\epsilon)$ and $\Delta_{g(\epsilon)} \leq \epsilon$, we have $H(\epsilon) \geq \frac{\zeta^2 g(\epsilon)}{g(\epsilon/2)\Delta_{g(\epsilon)}^2} \geq \zeta^2/\epsilon^2$.

The following corollary establishes that for any $\epsilon > 0$, Algorithm 1 returns ϵ-good recommendations $\hat{r}_t$ with high probability once $t \gg H(\epsilon)$.

Corollary 5 *Fix any $\epsilon > 0$ and $\delta \in (0,1)$. Assume that $\psi \geq 2^{30}\log^4(5t/\delta)$, and set the tuning parameter $\beta = 6\log(5t/\delta)$. With probability at least $1-\delta$, for any $t \geq 4\psi H(\epsilon)$, we have*

$$\lambda_{\gamma(\hat{r}_t)} \geq \lambda_0 - \epsilon \ .$$

Corollary 5 is analogous to Theorem 6 of Zhao et al. (2023), which analyzes the anytime version of Bracketing Doubling Sequential Halving (**BSH**). Our result requires a polylogarithmic factor of order $\log^4(t/\delta)$, while Zhao et al. (2023) achieves $\mathrm{polylog}(t)+\log(1/\delta)$. Our proof techniques could yield a δ-dependence of order $\log^2(1/\delta)$, but whether **OSE** can match the $\log(1/\delta)$ rate of **BSH** remains an open question. Nevertheless, as mentioned previously, we simplify all polylogarithmic factors throughout this manuscript for clearer exposition, and our numerical study demonstrates that these factors are negligible in practice.

The proof of Corollary 5 relies on a key observation: there exists a gap of at least $\epsilon/2$ between the quantiles corresponding to ranks $g(\epsilon/2)$ and $g(\epsilon)$. When $H(\epsilon) < t/\psi$, we can establish that $S(g(\epsilon)) \leq t/\psi$, which yields $\eta^* \leq g(\epsilon)$.

Proof [Proof of Corollary 5] Let $\nu \geq g(\epsilon)$ and $\rho < g(\epsilon/2)$. It holds that

$$G(\rho,\nu) = \frac{1}{\rho} \vee \frac{\zeta^2\nu}{\rho(\lambda_\rho - \lambda_\nu)^2} = \frac{1}{\rho} \vee \frac{\zeta^2\nu}{\rho(\Delta_\nu - \Delta_\rho)^2} \leq \frac{1}{\rho} \vee \frac{4\zeta^2\nu}{\rho\Delta_\nu^2} \ .$$

In the last inequality, we used that $\Delta_\rho \leq \epsilon/2 \leq \Delta_\nu/2$ from the definition of g. Hence, taking the limit as $\rho \to g(\epsilon/2)$ and the supremum over $\nu \geq g(\epsilon)$, we obtain that $S(g(\epsilon)) \leq 4H(\epsilon) \leq t/\psi$, which implies that $\eta_t^*(\psi) \leq g(\epsilon)$. From Theorem 3, we obtain that $\gamma(\hat{r}_t) \leq \eta_t^*(\psi) \leq g(\epsilon)$, and therefore $\lambda_0 - \lambda_{\gamma(\hat{r}_t)} \leq \epsilon$. ■

4.2 Applications to Various Distribution Types

Theorem 3 enables us to derive upper bounds on achievable ranks, lower bounds on simple rewards, and, when well-defined, bounds on simple regrets for several distribution classes. In what follows, we analyze three distribution types.

The first is the simplest we can consider: arms have mean u with probability η_0 or mean 0 with probability $1-\eta_0$. We refer to this as the Bernoulli-type distribution. This case was precisely analyzed in De Heide et al. (2021) when $\zeta = 1$ and $u \in [0,1]$, and we recover their result for identifying a good arm up to polylogarithmic factors in t/δ.

The second class corresponds to Beta$(1, 1/\alpha)$ distributions: bounded distributions on $[0, 1]$ satisfying $\lambda_\eta = 1 - \eta^\alpha$. This class has been extensively studied in the literature, both for bandits with infinitely many arms (Carpentier and Valko, 2015; Bonald and Proutiere, 2013) and in its discrete counterpart, the so-called Polynomial(α) instance problem (Zhao et al., 2023; Jamieson et al., 2013). In extreme value theory, these distributions fall in the Weibull domain of attraction (De Haan and Ferreira, 2006). A particularly interesting new result is the phase transition that occurs in our upper bound when $\alpha \geq 1/2$ for Beta distributions, for which we provide intuition below (10).

The third class consists of Pareto distributions with unbounded support in $[1, +\infty)$, which fall into the Fréchet domain of attraction in extreme value theory. These distributions are characterized by quantile functions $\lambda_\eta = \eta^{-\alpha}$ for $\alpha > 0$, and simple regret is not well-defined for them. To the best of our knowledge, no previous results exist for this class, as boundedness of the arm mean distribution is nearly always assumed in these settings. Nevertheless, we identify an interesting phase transition between the regimes $\alpha \leq 1/2$ and $\alpha > 1/2$ when the noise level ζ is large. The cases $\alpha \leq 1/2$ and $\alpha > 1/2$ respectively correspond to Pareto distributions with lighter and heavier tails.

For both regimes $\alpha \leq 1/2$ and $\alpha > 1/2$ in the Pareto class, we establish an upper bound of order $1/t$ (up to polylogarithmic factors in t/δ) for the rank $\gamma(\hat{r}_t)$ of the recommended arm when $t \gg \zeta^{1/\alpha}$ (up to a polylog factor). This bound is essentially tight: achieving a better rank than $1/t$ with high probability is impossible, as the probability that no arm $a \leq t$ satisfies $\gamma(a) \leq 1/t$ is at least $1 - (1 - 1/t)^t \geq 1 - e^{-1}$.

We call this the quasi-noiseless regime: once $t \gg \zeta^{1/\alpha}$, newly explored arms that are significantly better than the current recommendation can be detected in polylog(t/δ) steps. While Theorem 3 yields a trivial upper bound when $\alpha \geq 1/2$ and $t \lesssim \zeta^{1/\alpha}$, it provides a non-trivial bound of order $(\frac{\zeta^2}{t})^{1/(1-2\alpha)}$ when $\alpha < 1/2$ and $t \lesssim \zeta^{1/\alpha}$. Although we do not provide optimality results for the Pareto class in this manuscript, this suggests that the transition to the quasi-noiseless regime is much more abrupt when $\alpha \geq 1/2$. The intuition is that the learner must wait time of order $\zeta^{1/\alpha}$ before the distribution $\mathcal{D}$ outputs arms with mean $\gtrsim \zeta$. We summarize our findings for these three distribution classes in Table 1, with polylogarithmic factors suppressed through the rescaled time $\mathbf{t}$.

Type of the distribution $\mathcal{D}$	Quantile Function λ_η	UB on $\gamma(\hat{r}_t)$	LB on $\lambda_{\gamma(\hat{r}_t)}$
Bernoulli	$u\mathbf{1}\{\eta \leq \eta_0\}$	$\eta_0 \mathbf{1}\{\mathbf{t} \geq \frac{\zeta^2}{\eta_0 u^2}\}$	$u\mathbf{1}\{\mathbf{t} \geq \frac{\zeta^2}{\eta_0 u^2}\}$
Beta $(\alpha < 1/2)$	$1 - \eta^\alpha$	$\frac{1 \vee \zeta^2}{\mathbf{t}} \vee \left(\frac{\zeta^2}{\alpha^2 \mathbf{t}}\right)^{\frac{1}{2\alpha}}$	$1 - \left(\frac{1 \vee \zeta^2}{\mathbf{t}} \vee \left(\frac{\zeta^2}{\alpha^2 \mathbf{t}}\right)^{\frac{1}{2\alpha}}\right)^\alpha$
Beta $(\alpha \geq 1/2)$	$1 - \eta^\alpha$	$\frac{1}{\mathbf{t}} \vee \left(\frac{\zeta^2}{\alpha^2 \mathbf{t}}\right)^{\frac{1}{2\alpha}}$	$1 - \frac{1}{\mathbf{t}^\alpha} \vee \sqrt{\frac{\zeta^2}{\alpha^2 \mathbf{t}}}$
Pareto $(\alpha < 1/2)$	$\eta^{-\alpha}$	$\frac{1}{\mathbf{t}} \vee \left(\frac{\zeta^2}{\mathbf{t}}\right)^{\frac{1}{1-2\alpha}}$	$\mathbf{t}^\alpha \wedge \left(\frac{\mathbf{t}}{\zeta^2}\right)^{\frac{\alpha}{1-2\alpha}}$
Pareto $(\alpha \geq 1/2)$	$\eta^{-\alpha}$	$\frac{1}{\mathbf{t}} \vee \mathbf{1}\{\mathbf{t} < \zeta^{1/\alpha}\}$	$\mathbf{t}^\alpha \mathbf{1}\{\mathbf{t} \geq \zeta^{1/\alpha}\}$

Table 1: Orders of magnitude for the rank $\gamma(\hat{r}_t)$ and expected simple reward $\lambda_{\gamma(\hat{r}_t)}$ achieved by **OSE** with high probability across different distribution types. Here, $\mathbf{t}$ denotes $t/\tilde{\psi}$, where $\tilde{\psi}$ is a large polylogarithmic factor in t/δ that does not depend on α.

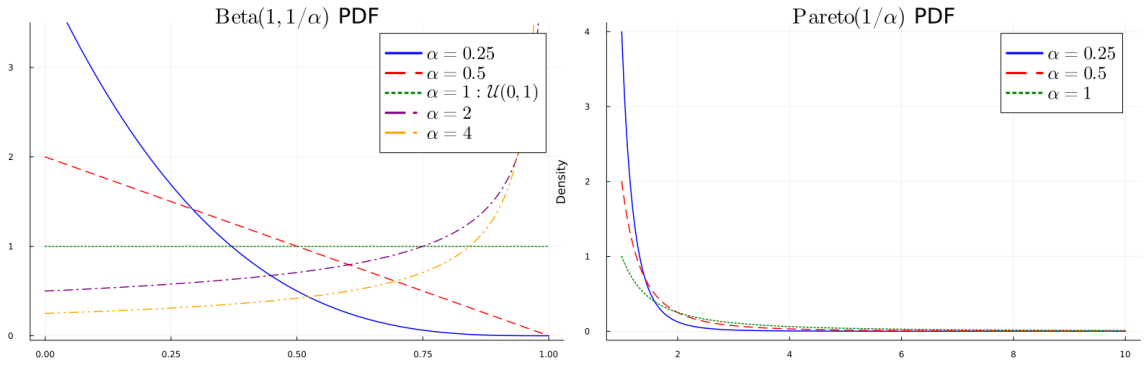

Figure 2: Probability density functions for Beta$(1, 1/\alpha)$ (left) and Pareto$(1/\alpha, 1)$ (right) distributions for various values of α. The Beta distribution is supported on $[0, 1]$ and becomes more concentrated near 1 as α increases. The Pareto distribution is supported on $[1, \infty)$ and has heavier tail as α increases.

Bernoulli. Assume that $\lambda_\eta = u$ if $\eta \le \eta_0$ and $\lambda_\eta = 0$ otherwise. This corresponds to assuming that $\lambda_{\gamma(a)}$ follows a Bernoulli distribution on $\{0, u\}$ with unknown parameter η_0. When the arm distribution is supported on $[0, 1]$ (implying $\zeta \le 1$), De Heide et al. (2021) established that the best arm can be identified after time t such that $t \ge \log(1/\delta)\log(t)\frac{1}{\eta_0 u^2}$. However, they also showed that adapting to the unknown η_0 is impossible for obtaining non-trivial upper bounds on the cumulative regret, even in this simplest setting.

Theorem 3 recovers the result of De Heide et al. (2021) up to a polylogarithmic factor. In this setting, $G(\rho, \nu) \le \frac{\zeta^2 \nu}{\rho u^2}$ if $\rho \le \eta_0 < \nu$, and $G(\rho, \nu) = +\infty$ otherwise. Consequently, $S(\eta) = +\infty$ if $\eta \le \eta_0$, and $S(\eta) = \frac{\zeta^2}{\eta_0 u^2}$ otherwise. Thus, by Theorem 3, as soon as $t \ge \frac{\psi \zeta^2}{\eta_0 u^2}$ – where $\psi := \psi_{t,\delta}$ denotes the polylogarithmic factor of order $\log^4(t/\delta)$ from Theorem 3 —the recommendation $\hat{r}_t$ satisfies $\gamma(\hat{r}_t) \le \eta_0$ with high probability. In other words, one of the good arms among the top η_0 proportion is identified.

Beta distribution $(1, \frac{1}{\alpha})$. Let us fix $\alpha > 0$ and assume that $\lambda_\eta = 1 - \eta^\alpha$ for $\eta \in [0, 1]$. This corresponds to distributions on $[0, 1]$ with CDF $F(\lambda) = 1 - (1 - \lambda)^{1/\alpha}$. Up to constant factors, these distributions correspond to those studied in Carpentier and Valko (2015) with $\alpha = 1/\beta$. The discrete counterpart is called Polynomial(α) instances in (Jamieson et al., 2013; Zhao et al., 2023). Then, for $\nu \ge 2\eta$, it holds that

$$G(\eta, \nu) = \frac{1}{\eta} \vee \frac{\zeta^2 \nu}{\eta\,(\nu^\alpha - \eta^\alpha)^2} \le \frac{1}{\eta} \vee \frac{\zeta^2 \nu^{1-2\alpha}}{\eta(1 - 2^{-\alpha})^2}$$

Since the map $\nu \mapsto \nu/(\nu^\alpha - \eta^\alpha)^2$ is decreasing then increasing on $[2\eta, 1]$ (its unique interior critical point is a minimum), the supremum $\tilde{S}(\eta) = \sup_{\nu \ge 2\eta} G(\eta, \nu)$ is attained at an endpoint, $\nu = 2\eta$ or $\nu = 1$. For $\alpha \ge 1/2$ the endpoint $\nu = 2\eta$ dominates; for $\alpha < 1/2$ both endpoints contribute, the one at $\nu = 1$ using the gap $\nu^\alpha - \eta^\alpha = 1 - \eta^\alpha \ge 1/2$ (valid for

$\eta \le 2^{-1/\alpha}$). Hence, for $\eta \le 2^{-1/\alpha}$,

$$\tilde{S}(\eta) \le \begin{cases} \frac{1\vee 4\zeta^2}{\eta} \vee \frac{8\zeta^2}{\alpha^2\eta^{2\alpha}} & \text{if } \alpha < 1/2 \\ \frac{1}{\eta} \vee \frac{8\zeta^2}{\alpha^2\eta^{2\alpha}} & \text{otherwise} \end{cases} \tag{9}$$

We used that $1-\eta^\alpha \ge 1/2$ for $\eta \le 2^{-1/\alpha}$ (when $\alpha < 1/2$), and that $2^\alpha - 1 \ge \alpha/2$ for all $\alpha \in (0,1)$. Hence, recalling that $\tilde{\eta}_t(\psi) = 2\inf\{\eta \in (0, \frac{1}{2}) : \ \tilde{S}(\eta) \le \frac{t}{\psi}\}$, we obtain that

$$\tilde{\eta}_t(\psi) \le \begin{cases} \dfrac{2\psi(1\vee 4\zeta^2)}{t} \vee \left(\dfrac{8\psi\zeta^2}{\alpha^2 t}\right)^{\frac{1}{2\alpha}} & \le \dfrac{1\vee\zeta^2}{\mathbf{t}} \vee \left(\dfrac{\zeta^2}{\alpha^2\mathbf{t}}\right)^{\frac{1}{2\alpha}} \quad \text{if } \alpha < 1/2 \\ \dfrac{2\psi}{t} \vee \left(\dfrac{8\psi\zeta^2}{\alpha^2 t}\right)^{\frac{1}{2\alpha}} & \le \dfrac{1}{\mathbf{t}} \vee \left(\dfrac{\zeta^2}{\alpha^2\mathbf{t}}\right)^{\frac{1}{2\alpha}} \quad \text{otherwise,} \end{cases} \tag{10}$$

where we set $\mathbf{t} = t/(24\psi)$. Applying the relaxed version Corollary 4 of Theorem 3, with probability larger than $1-\delta$, we have $\gamma(\hat{r}_t) \le \tilde{\eta}_t(\psi)$. The approximate inequalities are given up to polylogarithmic factors in t/δ, but the dependence on α captures the correct order of magnitude.

The interpretation is as follows. For $\alpha < 1/2$, the bound is the maximum of a noise-dependent term $\left(\frac{\zeta^2}{\alpha^2 t}\right)^{1/(2\alpha)}$ and a term $\frac{1\vee\zeta^2}{t}$ that depends on the noise only through a constant factor. Since the exponent satisfies $1/(2\alpha) > 1$, the noise-dependent term decreases faster in t: it dominates for small t and is overtaken by $\frac{1\vee\zeta^2}{t}$ once $t \gtrsim t^\star = \left(\frac{\zeta^2}{\alpha^2}\right)^{1/(1-2\alpha)}$. This is the mirror image of the case $\alpha \ge 1/2$ analyzed below, where $1/(2\alpha) \le 1$ and the noise-dependent term instead takes over for $t \gtrsim t^\star$, producing a saturation effect. The threshold $\alpha = 1/2$ thus acts as a symmetry axis separating an early-time noise penalty (that washes out) from a late-time one (that saturates).

When $\alpha > 1/2$ and $\zeta \gtrsim \alpha$, the dominant term is of order $\left(\frac{\zeta^2}{\alpha^2 t}\right)^{\frac{1}{2\alpha}}$, which corresponds to the regime $\beta < 2$ in Carpentier and Valko (2015). The interesting behavior occurs when $\zeta \lesssim \alpha$. In this case, there is a phase transition around time $t^\star = \left(\frac{\zeta^2}{\alpha^2}\right)^{\frac{1}{1-2\alpha}}$. For $t \lesssim t^\star$, we obtain the quasi-noiseless bound of order $1/t$, while for $t \gtrsim t^\star$, the dominant term becomes $t^{-\frac{1}{2\alpha}}$, as in the $\zeta \gtrsim \alpha$ case.

This phenomenon is somewhat surprising: initially (for $t \lesssim t^\star$), the algorithm achieves a very favorable rank guarantee of order $1/t$, but after time $t^\star$, the upper bound degrades significantly to $t^{-\frac{1}{2\alpha}}$. While we do not provide a matching lower bound, we conjecture that this is the correct order of magnitude for the rank of the recommendation, and we interpret this phenomenon as a saturation effect. For $\mathrm{Beta}(1, 1/\alpha)$ distributions with small noise levels ζ, the initial observations before $t^\star$ tend to be well separated near the boundary at 1, which facilitates easy detection of new good arms. However, as time progresses and t exceeds $t^\star$, the estimated means of good arms accumulate near 1, making it increasingly difficult to distinguish among them and identify arms with favorable rank.

To compare with existing results, which do not account for multiplicative factors that depend on α or ζ, let us analyze the expected simple recommendation reward and regret.

We derive from the upper bound on $\tilde{\eta}_t$ that

$$\lambda_{\gamma(\hat{r}_t)} \geq 1 - \tilde{\eta}_t^{\alpha} \geq \begin{cases} 1 - \left(\frac{2\psi(1\vee 4\zeta^2)}{t}\right)^{\alpha} \vee \sqrt{\frac{8\psi\zeta^2}{\alpha^2 t}} & \text{if } \alpha < 1/2 \\ 1 - \left(\frac{2\psi}{t}\right)^{\alpha} \vee \sqrt{\frac{8\psi\zeta^2}{\alpha^2 t}} & \text{otherwise.} \end{cases} \tag{11}$$

Hence, letting $\mathbf{t} = t/(24\psi)$, we obtain the following simple regret upper bound:

$$1 - \lambda_{\gamma(\hat{r}_t)} \lesssim \begin{cases} \left(\frac{1\vee\zeta^2}{\mathbf{t}}\right)^{\alpha} \vee \sqrt{\frac{\zeta^2}{\alpha^2\mathbf{t}}} & \text{if } \alpha < 1/2 \\ \frac{1}{\mathbf{t}^{\alpha}} \vee \sqrt{\frac{\zeta^2}{\alpha^2\mathbf{t}}} & \text{otherwise.} \end{cases} \tag{12}$$

Notably, when $\zeta = 1$, we recover the rate established in Carpentier and Valko (2015) for infinitely-armed bandits, which was shown to be minimax optimal up to polylogarithmic factors. Ignoring constant factors depending on α or ζ, these results also correspond to polynomial(α) instances in the finite-armed case, as studied in Jamieson et al. (2013) and Zhao et al. (2023).

Pareto distribution with parameter $\frac{1}{\alpha}$. Fix $\alpha > 0$ and assume that $\lambda_\eta = \eta^{-\alpha}$ for any $\eta \in (0,1)$. This corresponds to unbounded distributions with support on $[1, +\infty)$ and CDF $F(\lambda) = 1 - \lambda^{-1/\alpha}$. For $\nu \geq 2\eta$, we have

$$G(\eta,\nu) = \frac{1}{\eta} \vee \frac{\zeta^2\nu}{\eta\left(\eta^{-\alpha} - \nu^{-\alpha}\right)^2} \ .$$

The map $\nu \mapsto \nu/(\eta^{-\alpha} - \nu^{-\alpha})^2$ is decreasing then increasing on $[2\eta, 1]$ (its unique interior critical point $\nu = \eta(1+2\alpha)^{1/\alpha}$ is a minimum), so $\tilde{S}(\eta) = \sup_{\nu\geq 2\eta} G(\eta,\nu)$ is attained at the endpoint $\nu = 1$. Since $\eta^{-\alpha} - 1 \geq \eta^{-\alpha}/2$ for $\eta \leq 2^{-1/\alpha}$, this implies that $\tilde{S}(\eta) \leq \frac{1}{\eta} \vee 4\zeta^2\eta^{2\alpha-1}$. Hence, $\tilde{S}(\eta) \leq \frac{t}{\psi}$ if $\eta \geq \frac{\psi}{t}$ and $4\zeta^2\eta^{2\alpha-1} \leq \frac{t}{\psi}$. When $\alpha \geq 1/2$, a sufficient condition is $\eta \geq \frac{\psi}{t}$ and $\frac{t}{\psi} \geq 4\zeta^2(\frac{\psi}{t})^{2\alpha-1}$, which is equivalent to $t \geq \psi(4\zeta^2)^{\frac{1}{2\alpha}}$.

We obtain the following upper bound on $\tilde{\eta}_t$ depending on whether $\alpha < 1/2$ or $\alpha \geq 1/2$:

$$\tfrac{1}{2}\tilde{\eta}_t(\psi) \leq \begin{cases} \frac{\psi}{t} \vee \left(\frac{4\psi\zeta^2}{t}\right)^{\frac{1}{1-2\alpha}} & \leq \frac{1}{\mathbf{t}} \vee \left(\frac{\zeta^2}{\mathbf{t}}\right)^{\frac{1}{1-2\alpha}} & \text{if } \alpha < 1/2 \\ \frac{\psi}{t}\mathbf{1}\{t \geq \psi(4\zeta^2)^{\frac{1}{2\alpha}}\} & \leq \frac{1}{\mathbf{t}}\mathbf{1}\{\mathbf{t} \geq (4\zeta^2)^{\frac{1}{2\alpha}}\} & \text{if } \alpha \geq 1/2 \ , \end{cases} \tag{13}$$

where we recall that $\mathbf{t} = t/(16\psi)$.

When $\zeta > 1$ and $\alpha \geq 1/2$, the threshold of order $\zeta^{1/\alpha}$ represents the order of magnitude of the number of arms we must sample before observing an arm with mean at least ζ. Observing an arm with mean exceeding the noise level ζ before time t of order $\zeta^{1/\alpha}$ is unlikely, as a union bound yields

$$\mathbb{P}(\exists a \leq t : \ \lambda_{\gamma(a)} \geq \zeta) \leq t\zeta^{-1/\alpha} \ .$$

Moreover, before time of order $\zeta^{1/\alpha}$, we have only $\zeta^{1/\alpha} \leq \zeta^2$ samples, which is insufficient to detect a good arm among arms with means in $[0,\zeta]$ under noise level ζ.

However, when $\alpha < 1/2$, the inequality $\zeta^{1/\alpha} \leq \zeta^2$ no longer holds. In this regime, the bound of order $\left(\frac{\zeta^2}{\mathbf{t}}\right)^{1/(1-2\alpha)}$ reflects the fact that even before reaching the time when $\mathcal{D}$ outputs arms with mean exceeding ζ (entering a quasi-noiseless regime), it is possible to detect arms with good rank of order $\left(\frac{\zeta^2}{\mathbf{t}}\right)^{1/(1-2\alpha)}$.

Concerning the expected simple reward of the recommended arm, it holds with probability at least $1-\delta$ that

$$\lambda_{\gamma(\hat{r}_t)} \geq \begin{cases} \mathbf{t}^\alpha \wedge \left(\frac{\mathbf{t}}{\zeta^2}\right)^{\frac{\alpha}{1-2\alpha}} & \text{if } \alpha < 1/2 \\ \mathbf{t}^\alpha \mathbf{1}\{\mathbf{t} \geq (4\zeta^2)^{\frac{1}{2\alpha}}\} & \text{if } \alpha \geq 1/2 \end{cases}, \tag{14}$$

where $\mathbf{t} = t/(32\psi)$ (half the value used in the bound on $\tilde{\eta}_t(\psi)$), and ψ is as defined in Theorem 3.

5. Numerical Study

In this section, we analyze and compare the performance of four algorithms: **OSE** analyzed in Theorem 3, its enhanced version **PROSE**, **BSH** from Zhao et al. (2023), and a naïve implementation of **UCB** algorithm. We define **UCB** identically to **OSE**, except that the exploration scope Z is set to $+\infty$. At each step, it pulls the arm with maximum UCB and recommends the arm with maximum LCB. For practical simulations, we use constant parameters β independent of t and error probability. We recall that our polylogarithmic factors are not optimal and this simplifies tuning to a single parameter.

We compare these procedures on synthetic data where arm means follow a Beta$(1, 1/\alpha)$ distribution, corresponding to the quantile function $\lambda_\eta = 1 - \eta^\alpha$. For each arm's reward, we generate normal samples $\mathcal{N}(\lambda_{\gamma(a)}, 1)$. These choices align with standard distributions for arm means (Carpentier and Valko, 2015; Zhao et al., 2023; Jamieson et al., 2013) in the canonical case where $\zeta = 1$.

Figure 3 shows the empirical performance for $\alpha \in \{0.25, 0.5, 1, 2\}$. The results demonstrate that **PROSE**, with $\beta = 10$, consistently outperforms the three other procedures for $t \geq 10$. **BSH** exhibits a staircase pattern due to its doubling-trick mechanism. **OSE** performs slightly better than **BSH** except at large time steps, where **OSE** becomes more unstable. As for the naïve UCB algorithm, it only starts to outperform **OSE** and **BSH** after time $t \gg K$. This behavior is intuitive: before time $t = K$, UCB algorithm performs only pure exploration and starts exploiting better arms only after time $t = K$.

The instability of **OSE** at large time steps is due to using too small a tuning parameter $\beta = 10$. In theory, this parameter should increase as a polylog of t. Surprisingly, this instability is much less pronounced for **PROSE**, as shown in Figure 4 for $\alpha \in \{0.5, 1\}$. This figure shows that while **OSE** (right plots) is completely unstable for $\beta = 1$ and too conservative for $\beta = 50$, **PROSE** (left plots) exhibits only slight instability even for small β values and still outperforms **BSH** for $\beta = 50$.

While cumulative regret is not a quantity of interest in this paper, it is still instructive to examine how the procedures perform under this metric. Both **OSE** and **BSH** cannot achieve better than quasi-linear cumulative regret. Indeed, they both continue to use a proportion of order $t/\log(t)$ trials to pull arms that were chosen randomly at the beginning of the learning process. On the other hand, it is well known—see for instance Theorem 7.1

of Lattimore and Szepesvári (2020)—that in the worst case, the **UCB** algorithm achieves a cumulative regret bound of order $\sqrt{Kt}$ up to polylogarithmic factors, which is sublinear in t when $t \geq K$. We observe this sublinear trend for **UCB** in Figure 5. More interestingly, and perhaps surprisingly, **PROSE** follows the same long-term trend as **UCB**. This demonstrates that, at least for our simulated data, ranking arms at each iteration according to their LCB—the key modification that transforms **OSE** into **PROSE**—achieves best-of-both-worlds performance: strong simple regret when $t \lesssim K$ and recovery of the cumulative regret trend of the **UCB** algorithm when $t \gg K$.

All experiments were run on a machine with an Intel Core Ultra 7 165H CPU (16 cores, 32 threads, up to 5.0 GHz) and 31 GB of RAM, running Ubuntu. For a fair comparison of execution speed, simple regret, and cumulative regret between **PROSE** and **BSH**, we implemented the **BSH** algorithm in Julia 1.11 following the additional recommendations from the authors in Appendix D of Zhao et al. (2023). For **PROSE**, we use the optimized implementation described at the end of Section 5, which maintains arms sorted by LCB in decreasing order and partitions them into differential sets sorted by UCB in decreasing order, with all structures kept up to date incrementally. Code is available on the author's GitHub webpage.

Over 2000 trials ($T_{\max}$ = 50000, K = 5000 arms), **PROSE** averaged 76 ms per trial (range: 30–184 ms) compared to our implemented version of **BSH**: 89 ms (range: 33–439 ms), representing a 15% speedup with reduced variance. For comparison, the original **BSH** paper Zhao et al. (2023) reported execution times of approximately 2 hours per plot for 500 trials with similar problem sizes ($K \approx 6000$ arms, T = 10000 timesteps) on an AMD Ryzen 5 PRO 4650GE with 16GB RAM, corresponding to roughly 14 seconds per trial. Our implementation achieves approximately a 100× speedup, likely attributed to a combination of algorithmic optimizations for **BSH**, modern hardware, and the use of Julia rather than potentially unoptimized implementations in interpreted languages like Python or R.

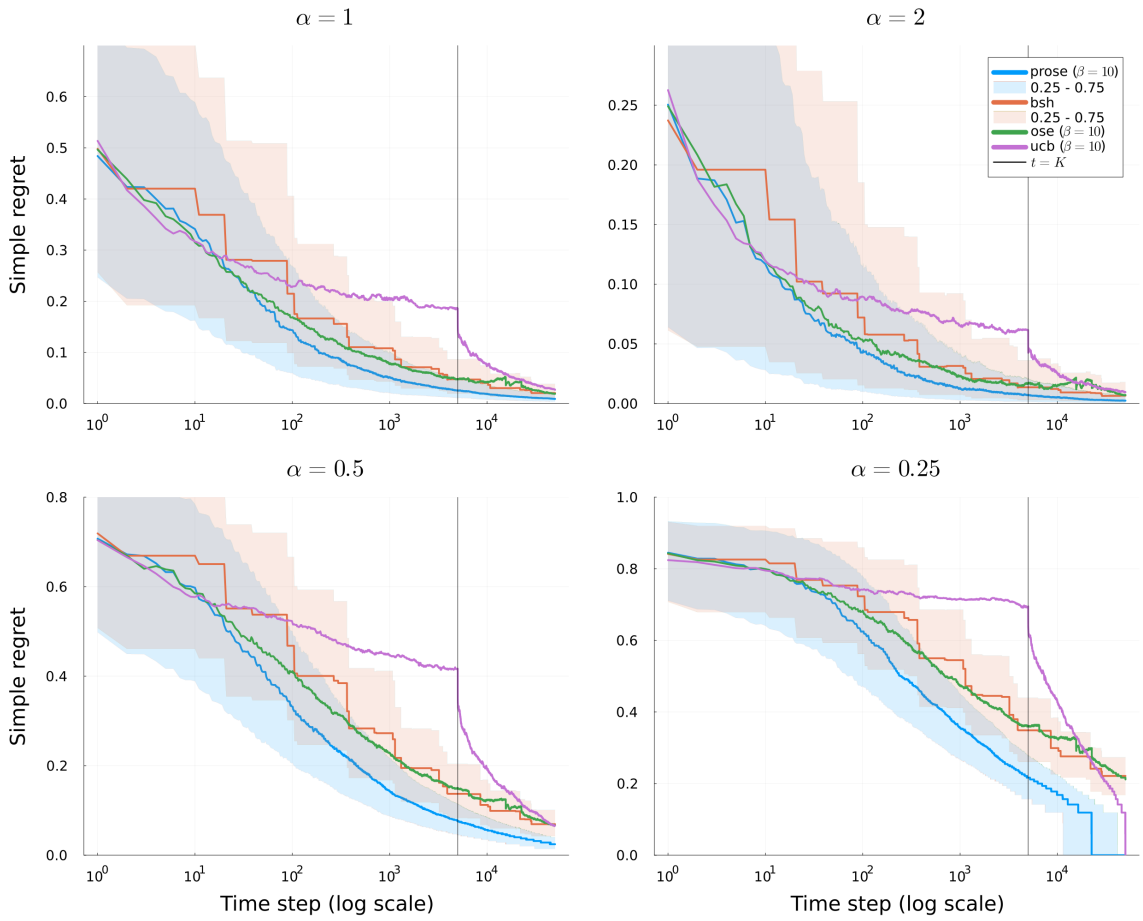

Figure 3: Median simple regrets for the four algorithms as a function of t on a log scale. Each plot shows results for quantile function $\lambda_\eta = 1-\eta^\alpha$ with $\alpha \in \{0.25, 0.5, 1, 2\}$, using $K = 5000$ arms and time horizon $T_{\max} = 50000$. Trajectories display median simple regret over $N = 2000$ trials. Shaded regions (blue for **PROSE**, orange for **BSH**) show the interquartile range.

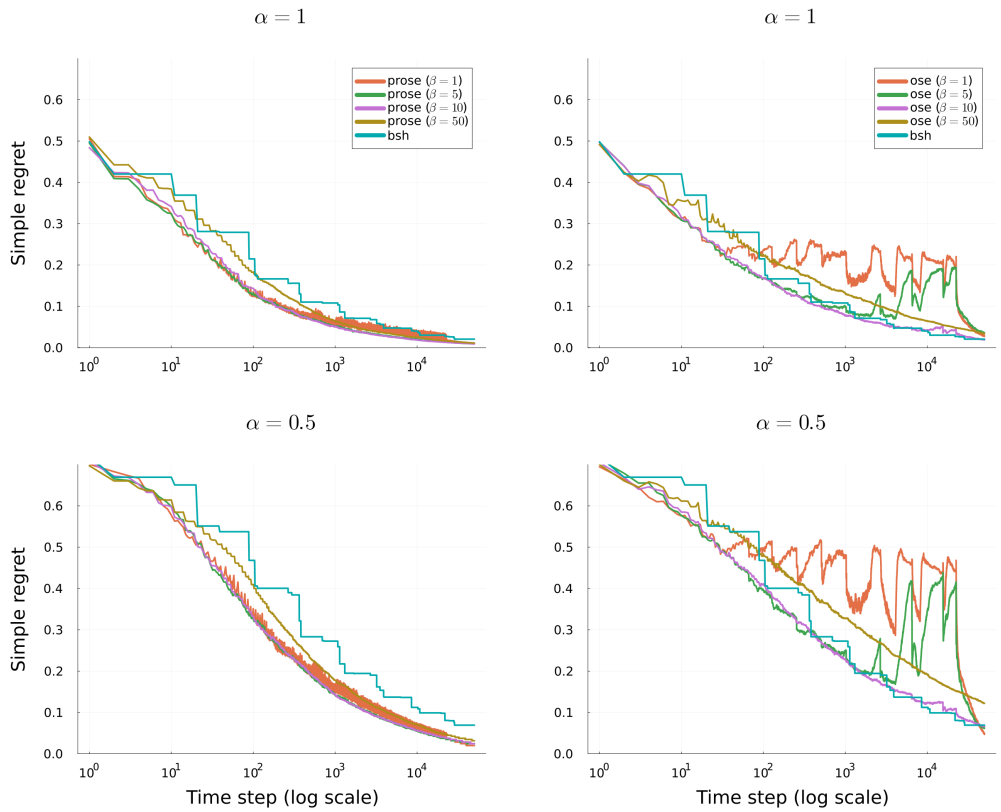

Figure 4: Median simple regrets for **BSH**, **PROSE** (left), and **OSE** (right) with tuning parameter $\beta \in \{1, 5, 10, 50\}$, for Beta$(1, 1/\alpha)$ instances with $\alpha \in \{0.5, 1\}$. Parameters: $K = 5000$ arms, time horizon $T_{\max} = 50000$, and $N = 2000$ trials per trajectory.

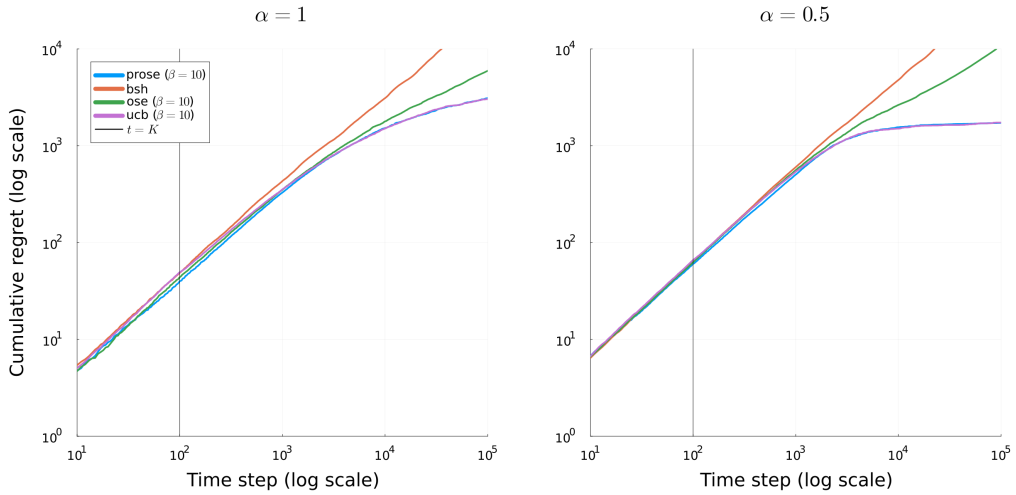

Figure 5: Cumulative regrets for **BSH**, **PROSE**, **OSE**, and **UCB** on Beta$(1, 1/\alpha)$ instances with $\alpha \in \{0.5, 1\}$. **PROSE** follows the same long-term sublinear trend as **UCB**, while **OSE** and **BSH** exhibit quasi-linear growth. Parameters: $K = 5000$ arms, $T_{\max} = 50000$, $N = 2000$ trials.

Appendix A. Proof of Theorem 3

A.1 High Probability Events

In the next three lemmas, we fix $\delta \in (0,1)$, which represents a small failure probability. Proofs are deferred to Appendix B. The first lemma establishes bounds on the noise sums that hold simultaneously for all $t \geq 1$ and all previously observed arms.

Lemma 6 *With probability at least $1-\delta/3$, for all $t \geq 1$ and all arm $a \in \mathcal{O}_t$ observed before time t,*

$$\left|\sum_{s=1}^{t} \varepsilon_{a,s}\right| \leq \sqrt{6\zeta^2 t \log(5t/\delta)} = \sqrt{\zeta^2 \beta t}\ .$$

Recall that at each step t of Algorithm 1, we generate a random variable $Z_t \in [1,t]$ representing the exploration scope, with probability density function $u \mapsto \frac{1}{\log(t)}\frac{1}{u}$. For $k \geq 1$, we define $\mathcal{T}(k,t) = \{t' \leq t : \ Z_{t'} \in [2^k, 2^{k+1})\}$, which denotes the set of time steps before t where the exploration scope falls in the interval $[2^k, 2^{k+1}]$.

Lemma 7 *With probability at least $1-\frac{\delta}{3}$, for all $t \geq 2^{12}\log^2(t)\log(5t/\delta)$ and all $k \in \{0,\dots,\lfloor \log_2(t)\rfloor - 1\}$,*

$$|\mathcal{T}(k,t)| \geq \frac{t}{64\log(t)}\ .$$

The following lemma formalizes the intuitive fact that among arms with indices in $[1,z]$ the number belonging to the top η fraction is of order ηz up to log terms and deviations of order $\sqrt{\eta z}$.

Lemma 8 *With probability at least $1-\frac{\delta}{3}$, for all $t \geq 1$, all $z \in \{1,\dots,t\}$ and all $\eta \in \{2^{-k}, k \in \{0,\dots,\lceil \log_2(t)\rceil\}\}$,*

$$\left|\sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta\} - \eta z\right| \leq \sqrt{8\eta z \log\left(\frac{5t}{\delta}\right)} + 4\log\left(\frac{5t}{\delta}\right)\ .$$

In particular, for any $\eta \in [\frac{1}{t}, 1]$, letting $\eta' = 2^{-\lfloor \log_2(1/\eta)\rfloor}$, we have that

$$\sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta\} \leq \sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta'\} \leq 8\left(\log\left(\frac{5t}{\delta}\right) \vee 2\eta z\right)\ . \tag{15}$$

For the second inequality, we used that $1 + 4 + \sqrt{8} \leq 8$ and $\eta' \leq 2\eta$. Moreover, if $\eta z \geq 64\log(5t/\delta)$, we have with $\eta'' = 2^{-\lceil \log_2(1/\eta)\rceil}$ that for all $z = 1,\dots,t$,

$$\sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta\} \geq \sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta''\} \geq 1\ . \tag{16}$$

A.2 Analysis of the Procedure, Proof of Theorem 3

We now proceed to the analysis of Algorithm 1. In what follows, we assume that the three events from Lemmas 6 to 8 simultaneously hold. This occurs with probability at least $1 - 3\delta/3 = 1-\delta$. Fix any step t of the procedure and η such that $S(\eta) \leq t/\psi$, where $S(\eta)$ is defined in (5). We aim to show that the recommendation $\hat{r}_t$ belongs to the top η fraction, meaning $\gamma(\hat{r}_t) \leq \eta$.

Step 1: Existence of a top-ρ arm with small exploration scope. The condition $S(\eta) \leq t/\psi$ implies the existence of $\rho < \eta$ such that $G(\rho, \nu) \leq t/\psi$ for all $\nu \geq \eta$. In other words, there exists $\rho < \eta$ such that for any $\nu \geq \eta$,

$$\lambda_\rho \geq \lambda_\nu + \zeta\sqrt{\frac{\nu\psi}{\rho t}} \quad \text{and } \rho \geq \frac{\psi}{t} \ . \tag{17}$$

Using (16) – which is derived from Lemma 6 – with $z = \lceil 64\log(5t/\delta)/\rho \rceil \leq t$, there exists an arm a_ρ such that

$$a_\rho \leq 64\log\left(\frac{5t}{\delta}\right)\frac{1}{\rho} \quad \text{and} \quad \gamma(a_\rho) \leq \rho \ . \tag{18}$$

The insight of (18) is that for an exploration scope of order significantly larger than $1/\rho$, there is at least one arm a_ρ in the top ρ fraction. From the assumption (17), since $\psi \geq 256\log(5t/\delta)$, we have $a_\rho \leq t/4$. Let $k_\rho = \lfloor \log_2(a_\rho) \rfloor + 1$, so that

$$2^{k_\rho - 1} \leq a_\rho < 2^{k_\rho} \ .$$

We recall that $\mathcal{T}(k_\rho, t) = \{t' \leq t : \ Z_{t'} \in [2^{k_\rho}, 2^{k_\rho+1})\}$. We define $N_{a,t}^{(k_\rho)} = \sum_{t' \in \mathcal{T}(k_\rho,t)} \mathbf{1}\{\hat{a}_{t'} = a\}$, which counts the number of times $t' \leq t$ arm a was pulled when $Z_{t'} \in [2^{k_\rho}, 2^{k_\rho+1})$. Crucially, whenever $Z_{t'} \in [2^{k_\rho}, 2^{k_\rho+1})$, the arm a_ρ lies within the exploration scope at step t' since $a_\rho < 2^{k_\rho} \leq Z_{t'}$, and it can thus be pulled if it has the largest UCB at time step t'.

Step 2: Lower and upper bounds on the UCB's for all arms. Fix any arm $a \geq 1$ and any time step $t' \leq t$. If arm a has not been pulled before time t', then its UCB at time t' is infinite. Otherwise, $N_{a,t'} \geq 1$ and From Lemma 6, we have $|\sum_{s=1}^{N_{a,t'}} \varepsilon_{a,s}| \leq \sqrt{\zeta^2 \beta N_{a,t'}}$. Hence,

$$\mathrm{UCB}_{a,t'} = \lambda_{\gamma(a)} + \frac{1}{N_{a,t'}}\sum_{s=1}^{N_{a,t'}} \varepsilon_{a,s} + \sqrt{\frac{\zeta^2\beta}{N_{a,t'}}} \geq \lambda_{\gamma(a)} \ .$$

In particular, since $\lambda_{\gamma(a_\rho)} \geq \lambda_\rho$,

$$\mathrm{UCB}_{a_\rho,t'} \geq \lambda_\rho \ . \tag{19}$$

Using the same argument, we also obtain that

$$\mathrm{UCB}_{a,t'} \leq \lambda_{\gamma(a)} + 2\sqrt{\frac{\zeta^2\beta}{N_{a,t'}}} \ . \tag{20}$$

Step 3: Non top-η arms have not been pulled too often. Let a be an arm that is not in the top η fraction, meaning $\gamma(a) \geq \eta$. Let $t' \leq t$ be the last step in $\mathcal{T}(k_\rho, t)$ at which a has been pulled. This means that $N_{a,t'-1}^{(k_\rho)} + 1 = N_{a,t'}^{(k_\rho)} = N_{a,t}^{(k_\rho)}$. Since $a_\rho \leq 2^{k_\rho} \leq Z_{t'}$, the UCB of a at step t' is necessarily larger than the UCB of a_ρ - Otherwise arm a_ρ would have been pulled instead:

$$\mathrm{UCB}_{a,t'-1} \geq \mathrm{UCB}_{a_\rho,t'-1} \ .$$

Using (19) for a_ρ and (20) for a we obtain that $\lambda_{\gamma(a)} + 2\sqrt{\frac{\zeta^2\beta}{N_{a,t'-1}}} \geq \lambda_\rho$, which implies that

$$N_{a,t}^{(k_\rho)} = N_{a,t'-1}^{(k_\rho)} + 1 \leq N_{a,t'-1} + 1 \leq \frac{4\zeta^2\beta}{(\lambda_\rho - \lambda_{\gamma(a)})^2} + 1 \leq \frac{4\beta}{\psi}\frac{\rho t}{\gamma(a)} + 1 \leq \frac{8\beta}{\psi}\frac{\rho t}{\gamma(a)} \ . \tag{21}$$

In the third inequality, we used assumption (17). In the last inequality, we used that $\frac{4\beta\rho t}{\psi\gamma(a)} \geq \frac{\rho t}{\psi} \geq 1$. Hence, an arm a such that $\gamma(a) \geq \eta$ has been pulled a number of times of order at most $\rho t/\gamma(a)$, within time steps t' in $\mathcal{T}(k_\rho, t)$.

Step 4: At least one top-η arm a^* has been pulled very often. Since $a_\rho \leq t/4$, we have that $k_\rho \leq \log_2(t) - 1$. Additionally, from (17), $t \geq \psi \geq 2^{12}\log^2(t)\log(5t/\delta)$ and we are in position to apply Lemma 7:

$$|\mathcal{T}(k_\rho, t)| \geq \frac{t}{64\log(t)} \ . \tag{22}$$

At any step $t' \in \mathcal{T}(k_\rho, t)$, the arm $\hat{a}_{t'}$ that was pulled satisfies $\hat{a}_{t'} \leq Z_{t'} < 2^{k_\rho+1}$. Hence,

$$|\mathcal{T}(k_\rho, t)| = \sum_{t' \in \mathcal{T}(k_\rho,t)} \sum_{a=1}^{2^{k_\rho+1}} \mathbf{1}\{\hat{a}_{t'} = a\} = \sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} \ .$$

Let $\nu_l = 2^{-l}$ for $l \in \mathcal{L} = \{0, 1, 2, \ldots, \lfloor \log_2(1/\eta) \rfloor\}$, and $\nu_{|\mathcal{L}|+1} = \eta$. We have

$$|\mathcal{T}(k_\rho, t)| = \sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} = \sum_{l \in \mathcal{L}} \sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} \mathbf{1}\{\nu_{l+1} < \gamma(a) \leq \nu_l\} + \sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} \mathbf{1}\{\gamma(a) \leq \eta\} \ . \tag{23}$$

If $\gamma(a) \in (\nu_{l+1}, \nu_l]$, then (21) gives that $N_{a,t}^{(k_\rho)} \leq \frac{8\beta}{\psi}\frac{\rho t}{\nu_{l+1}}$. Hence,

$$\begin{aligned}
\sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} \mathbf{1}\{\nu_{l+1} < \gamma(a) \leq \nu_l\} &\leq \frac{8\beta}{\psi}\frac{\rho t}{\nu_{l+1}} \sum_{a=1}^{2^{k_\rho+1}} \mathbf{1}\{\nu_{l+1} < \gamma(a) \leq \nu_l\} \\
&\leq \frac{64\beta}{\psi}(\log(5t/\delta) \vee 2\nu_l 2^{k_\rho+1})\frac{\rho t}{\nu_{l+1}} \\
&\leq \frac{64\beta}{\psi}(\log(5t/\delta) \vee 8\nu_l a_\rho)\frac{\rho t}{\nu_{l+1}} \\
&\leq \frac{2^{13}\beta}{\psi}\log(5t/\delta)t \leq \frac{|\mathcal{T}(k_\rho, t)|}{2|\mathcal{L}|} \ .
\end{aligned}$$

In the second inequality, we used (15) with ν_l and $z = 2^{k_\rho+1}$. For the fourth inequality, we used that $\nu_l/\nu_{l+1} \leq 2$ and $\rho a_\rho \leq 64\log(5t/\delta)$ and that $\rho/\nu_{l+1} \leq 1$. For the last inequality, we used (22) and the fact that $\psi \geq 2^{20}|\mathcal{L}|\log(t)\log(5t/\delta)\beta$. Summing this bound over the $|\mathcal{L}| \leq \log_2(1/\eta) + 1$ rank-shells and plugging into (23), we obtain

$$\sum_{a=1}^{2^{k_\rho+1}} N_{a,t}^{(k_\rho)} \mathbf{1}\{\gamma(a) \leq \eta\} \geq \frac{|\mathcal{T}(k_\rho, t)|}{2} \geq \frac{t}{128\log(t)} \ , \tag{24}$$

where we used Lemma 7 for the last inequality. Using (15) again, we obtain that

$$\begin{aligned}
\frac{t}{128\log(t)} &\leq \left(\max_{a:\gamma(a)\leq\eta} N_{a,t}^{(k_\rho)}\right) \sum_{a=1}^{2^{k_\rho+1}} \mathbf{1}\{\gamma(a) \leq \eta\} \\
&\leq \left(\max_{a:\gamma(a)\leq\eta} N_{a,t}^{(k_\rho)}\right) \cdot 8\left(\log\left(\frac{5t}{\delta}\right) \vee 8\eta a_\rho\right) \\
&\leq 2^{12}\log\left(\frac{5t}{\delta}\right)\left(\max_{a:\gamma(a)\leq\eta} N_{a,t}^{(k_\rho)}\right)\frac{\eta}{\rho} \ .
\end{aligned}$$

In the last inequality, we used that $a_\rho \le 64\log(\frac{5t}{\delta})\frac{1}{\rho}$. Hence, there exists a^* such that $\gamma(a^*) \le \eta$ and

$$N_{a^*,t} \ge N_{a^*,t}^{(k_\rho)} \ge \frac{1}{2^{19}\log^2(5t/\delta)}\frac{\rho}{\eta}t \ . \tag{25}$$

Step 5: This top-η arm a^* has an LCB of order λ_ρ. Let $t' \le t$ be the last step a^* was pulled, so that $\mathrm{UCB}_{a^*,t'-1} = \mathrm{UCB}_{a^*,t-1}$, $\mathrm{LCB}_{a^*,t'} = \mathrm{LCB}_{a^*,t}$ and $N_{a^*,t'-1} + 1 = N_{a^*,t}$. Then,

$$\lambda_{\gamma(a^*)} + 2\sqrt{\frac{\zeta^2\beta}{N_{a^*,t'-1}}} \ge \mathrm{UCB}_{a^*,t'-1} \ge \mathrm{UCB}_{a_\rho,t'-1} \ge \lambda_\rho \ .$$

Using (25), the inequality $\rho \ge \psi/t$, $\eta \le 1$ and $\psi \ge 2^{20}\log^2(5t/\delta)$, we have that $N_{a^*,t'-1} = N_{a^*,t} - 1 \ge \frac{1}{2^{20}\log^2(5t/\delta)}\frac{\rho}{\eta}t$. Hence,

$$\lambda_{\gamma(a^*)} \ge \lambda_\rho - \sqrt{2^{20}\log^2(5t/\delta)\zeta^2\beta\frac{\eta}{\rho t}} \ . \tag{26}$$

This implies that

$$\mathrm{LCB}_{a^*,t} \ge \lambda_{\gamma(a^*)} - 2\sqrt{\frac{\zeta^2\beta}{N_{a^*,t'-1}}} > \lambda_\rho - \sqrt{2^{24}\log^2(5t/\delta)\zeta^2\beta\frac{\eta}{\rho t}} \ . \tag{27}$$

Step 6: All non-top-η arms have smaller LCB than a^*. Let a be any arm in the $(1-\eta)$ bottom fraction, meaning $\gamma(a) \ge \eta$. From the same argument as in Step 2, it holds that $\mathrm{LCB}_{a,t} \le \lambda_\eta$. Hence,

$$\mathrm{LCB}_{a^*,t} - \mathrm{LCB}_{a,t} > \lambda_\rho - \lambda_\eta - \sqrt{2^{24}\log^2(5t/\delta)\zeta^2\beta\frac{\eta}{\rho t}} \ge \zeta\sqrt{\frac{\eta\psi}{\rho t}} - \sqrt{2^{24}\log^2(5t/\delta)\zeta^2\beta\frac{\eta}{\rho t}} > 0 \ . \tag{28}$$

The last inequality comes from the fact that we choose $\psi > 2^{24}\log^2(5t/\delta)\beta$ (the factor ζ cancels on both sides). Hence, arm a cannot be chosen as recommendation at step t by Algorithm 1, since a^* has a higher LCB. Therefore, the recommendation $\hat{r}_t$ is necessarily an η-good arm, that is $\gamma(\hat{r}_t) \le \eta$. As this is valid for any η such that $S(\eta) \le t/\psi$, we conclude that $\gamma(\hat{r}_t) \le \eta_t^*(\psi)$.

Appendix B. Proofs of Technical Lemmas

Proof [Proof of Lemma 6] From Hoeffding inequality applied to the independent ζ-subgaussian random variables $\varepsilon_{a,s}$, for any fixed $t \ge 1$, we have with probability at least $1 - \frac{2\delta}{\pi^2 t^3}$ that

$$\left|\sum_{s=1}^{t}\varepsilon_{a,s}\right| \le \sqrt{2\zeta^2 t\log(\tfrac{\pi^2 t^3}{2\delta})} \le \sqrt{6\zeta^2 t\log(5t/\delta)} \ .$$

Since $|\mathcal{O}_t| \le t$ and $\sum \frac{1}{t^2} = \pi^2/6$, we get the result with a union bound over all possible $a \in \mathcal{O}_t$ and $t \ge 1$. ■

Proof [Proof of Lemma 7] Let $s \geq 1 + 3t/4$. Since $2^{k+1} \leq t$, it holds that $s \geq \frac{3}{2}2^k$. Hence, Z_s satisfies

$$\mathbb{P}(Z_s \in [2^k, 2^{k+1})) = \int_{2^k}^{2^{k+1} \wedge s} \frac{1}{u \log(s)} du = \frac{\log(2^{k+1} \wedge s) - \log(2^k)}{\log(s)} \geq \frac{\log(3/2)}{\log(t)} \geq \frac{1}{4\log(t)} .$$

The $\lfloor t/4 \rfloor$ random variables $\mathbf{1}\{Z_s \in [2^k, 2^{k+1})\}$ for $s \in \{1 + \lceil 3t/4 \rceil, \dots, t\}$ are independent and take values in $\{0, 1\}$. Hence, using Hoeffding's inequality on the bounded variables $\mathbf{1}\{Z_s \in [2^k, 2^{k+1})\}$, we obtain that, with probability at least $1 - \frac{2\delta}{\pi^2 t^3}$,

$$|\mathcal{T}(k,t)| \geq \sum_{s \geq 3t/4+1} \mathbf{1}\{Z_s \in [2^k, 2^{k+1})\} \geq \frac{t}{32\log(t)} - \sqrt{\frac{t}{8}\log\left(\frac{\pi^2 t^3}{2\delta}\right)} \geq \frac{t}{64\log(t)} .$$

For the last inequality, we used the assumption which implies that $\sqrt{\frac{t}{8}\log(\frac{\pi^2 t^3}{2\delta})} \leq \frac{t}{64\log(t)}$. We obtain the result from a union bound over all $t \geq 1$ and all $k = 0, \dots, \lfloor \log_2(t) \rfloor - 1$ (there are at most $\log_2(t) \leq t$ such k's). ■

Proof [Proof of Lemma 8] Let us fix $z \in \{1, \dots, t\}$ and $\eta \in \{2^{-k}, k \in \{0, \dots, \lceil \log_2(t) \rceil\}\}$. The random variables $(\gamma(a))_{a=1,\dots,z}$ are iid uniform in $[0, 1]$. Hence $\mathbf{1}\{\gamma(a) \leq \eta\}$ are iid Bernoulli random variables with parameter η. From Bernstein's inequality—see Boucheron et al. (2003)—it holds that, with probability at least $1 - \frac{2\delta}{\pi^2 t^4}$,

$$\left|\sum_{a=1}^{z} \mathbf{1}\{\gamma(a) \leq \eta\} - \eta z\right| \leq \sqrt{2\eta z \log\left(\frac{\pi^2 t^4}{2\delta}\right)} + \log\left(\frac{\pi^2 t^4}{2\delta}\right) \leq \sqrt{8\eta z \log\left(\frac{5t}{\delta}\right)} + 4\log\left(\frac{5t}{\delta}\right) .$$

A union bound over all $z \in \{1, \dots, t\}$ and all $\eta \in \{2^{-k}, k \in \{0, \dots, \lceil \log_2(t) \rceil\}\}$ gives the same inequality, simultaneously for all z and all η, with probability $1 - \frac{2\delta}{\pi^2 t^2}$. A final union bound over all $t \geq 1$ gives the result. ■

References

M. Aziz, J. Anderton, K. Jamieson, A. Wang, H. Bouchard, and J. Aslam. Identifying new podcasts with high general appeal using a pure exploration infinitely-armed bandit strategy. In Proceedings of the 16th ACM Conference on Recommender Systems, pages 134–144, 2022.

D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp. Bandit problems with infinitely many arms. The Annals of Statistics, 25(5):2103 – 2116, 1997. doi: 10.1214/aos/1069362389. URL https://doi.org/10.1214/aos/1069362389.

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

T. Bonald and A. Proutiere. Two-target algorithms for infinite-armed bandits with bernoulli rewards. Advances in Neural Information Processing Systems, 26, 2013.

S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In Summer school on machine learning, pages 208–240. Springer, 2003.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In International conference on Algorithmic learning theory, pages 23–37. Springer, 2009.

A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In International Conference on Machine Learning, pages 1133–1141. PMLR, 2015.

L. De Haan and A. Ferreira. Extreme value theory: an introduction. Springer, 2006.

R. De Heide, J. Cheshire, P. Ménard, and A. Carpentier. Bandits with many optimal arms. Advances in Neural Information Processing Systems, 34:22457–22469, 2021.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In Proceedings of the 33rd International Conference on Machine Learning (ICML), pages 1587–1595, 2016.

L. Jain, K. Jamieson, R. Mankoff, R. Nowak, and S. Sievert. The new yorker cartoon caption contest dataset. https://nextml.github.io/caption-contest-data/, 2020.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. On finding the largest mean among many. stat, 1050:17, 2013.

K.-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In Proceedings of the 33rd International Conference on Machine Learning (ICML), pages 974–982, 2016.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In Proceedings of the 30th International Conference on Machine Learning (ICML), volume 28 of Proceedings of Machine Learning Research, pages 1238–1246. PMLR, 2013.

J. Katz-Samuels and K. Jamieson. The true sample complexity of identifying good arms. In International Conference on Artificial Intelligence and Statistics, pages 1781–1791. PMLR, 2020.

T. Lattimore. Regret analysis of the anytime optimally confident UCB algorithm. In Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT), pages 199–213, 2016.

T. Lattimore and C. Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research, 18(185):1–52, 2018.

B. Mason, R. Camilleri, S. Mukherjee, and K. Jamieson. Finding all ε-good arms in stochastic bandits. In Advances in Neural Information Processing Systems, volume 33, pages 20707–20718, 2020.

N. Verzelen, M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Optimal change-point detection and localization. The Annals of Statistics, 51(4):1586–1610, 2023.

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. Advances in Neural Information Processing Systems, 21, 2008.

Y. Zhao, C. Stephens, C. Szepesvári, and K.-S. Jun. Revisiting simple regret: Fast rates for returning a good arm. In International Conference on Machine Learning, pages 42110–42158. PMLR, 2023.