On Regret with Multiple Best Arms

Yinglun Zhu
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706
yinglun@cs.wisc.edu

Robert Nowak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706
rdnowak@wisc.edu

Abstract

We study a regret minimization problem with the existence of multiple best/near-optimal arms in the multi-armed bandit setting. We consider the case when the number of arms/actions is comparable or much larger than the time horizon, and make no assumptions about the structure of the bandit instance. Our goal is to design algorithms that can automatically adapt to the unknown hardness of the problem, i.e., the number of best arms. Our setting captures many modern applications of bandit algorithms where the action space is enormous and the information about the underlying instance/structure is unavailable. We first propose an adaptive algorithm that is agnostic to the hardness level and theoretically derive its regret bound. We then prove a lower bound for our problem setting, which indicates: (1) no algorithm can be minimax optimal simultaneously over all hardness levels; and (2) our algorithm achieves a rate function that is Pareto optimal. With additional knowledge of the expected reward of the best arm, we propose another adaptive algorithm that is minimax optimal, up to polylog factors, over all hardness levels. Experimental results confirm our theoretical guarantees and show advantages of our algorithms over the previous state-of-the-art.

1 Introduction

Multi-armed bandit problems describe exploration-exploitation trade-offs in sequential decision making. Most existing bandit algorithms tend to provide regret guarantees when the number of available arms/actions is smaller than the time horizon. In modern applications of bandit algorithm, however, the action space is usually comparable or even much larger than the allowed time horizon so that many existing bandit algorithms cannot even complete their initial exploration phases. Consider a problem of personalized recommendations, for example. For most users, the total number of movies, or even the amount of sub-categories, far exceeds the number of times they visit a recommendation site. Similarly, the enormous amount of user-generated content on YouTube and Twitter makes it increasingly challenging to make optimal recommendations. The tension between a very large action space and a limited time horizon poses a realistic problem in which deploying algorithms that converge to an optimal solution over an asymptotically long time horizon do not give satisfying results. There is a need to design algorithms that can exploit the highest possible reward within a limited time horizon. Past work has partially addressed this challenge. The quantile regret proposed in [12] to calculate regret with respect to an satisfactory action rather than the best one. The discounted regret analyzed in [25, 24] is used to emphasize short time horizon performance. Other existing works consider the extreme case when the number of actions is indeed infinite, and tackle such problems with one of two main assumptions: (1) the discovery of a near-optimal/best arm follows some probability measure with known parameters [6, 30, 4, 15]; (2) the existence of a smooth function represents the mean-payoff over a continuous subset [1, 20, 19, 8, 23, 17]. However, in many situations, neither assumption may be realistic. We make minimal assumptions in this paper. We study the regret minimization problem over a time horizon $T$, which might be unknown, with respect
to a bandit instance with \( n \) total arms, out of which \( m \) are best/near-optimal arms. We emphasize that the allowed time horizon and the given bandit instance should be viewed as features of one problem and together they indicate an intrinsic hardness level. We consider the case when the number of arms \( n \) is comparable or larger than the time horizon \( T \) so that no standard algorithm provides satisfying result. Our goal is to design algorithms that could adapt to the unknown \( m \) and achieve optimal regret.

1.1 Contributions and paper organization

We make the following contributions. In Section 2, we formally define the regret minimization problem that represents the tension between a very large action space and a limited time horizon; and capture the hardness level in terms of the number of best arms. We provide an adaptive algorithm that is agnostic to the unknown number of best arms in Section 3 and theoretically derive its regret bound. In Section 4, we prove a lower bound for our problem setting that indicates that there is no algorithm that can be optimal simultaneously over all hardness levels. Our lower bound also shows that our algorithm provided in Section 3 is Pareto optimal. With additional knowledge of the expected reward of the best arm, in Section 5, we provide an algorithm that achieves the non-adaptive minimax optimal regret, up to polylog factors, without the knowledge of the number of best arms. Experiments conducted in Section 6 confirm our theoretical guarantees and show advantages of our algorithms over previous state-of-the-art. We conclude our paper in Section 7. Most of the proofs are deferred to the Appendix due to lack of space.

1.2 Related work

Time sensitivity and large action space. As bandit models are getting much more complex, usually with large or infinite action spaces, researchers have begun to pay attention to tradeoffs between regret and time horizons when deploying such models. [13] study a linear bandit problem with ultra-high dimension, and provide algorithms that, under various assumptions, can achieve good reward within short time horizon. [24] also take time horizon into account and model time preference by analyzing a discounted regret. [12] consider a quantile regret minimization problem where they define their regret with respect to expected reward ranked at \((1 - \rho)\)-th quantile. One could easily transfer their problem to our setting; however, their regret guarantee is sub-optimal. [18, 4] also consider the problem with \( m \) best/near-optimal arms with no other assumptions, but they focus on the pure exploration setting; [4] additionally requires the knowledge of \( m \). Another line of research considers the extreme case when the number arms is infinite, but with some known regularities. [6] proposes an algorithm with a minimax optimality guarantee under the situation where the reward of each arm follows strictly Bernoulli distribution; [27] provides an anytime algorithm that works under the same assumption. [30] relaxes the assumption on Bernoulli reward distribution, however, some other parameters are assumed to be known in their setting.

Continuum-armed bandit. Many papers also study bandit problems with continuous action spaces, where they embed each arm \( x \) into a bounded subset \( X \subseteq \mathbb{R}^d \) and assume there exists a smooth function \( f \) governing the mean-payoff for each arm. This setting is firstly introduced by [11]. When the smoothness parameters are known to the learner or under various assumptions, there exists algorithms [20, 19, 8] with near-optimal regret guarantees. When the smoothness parameters are unknown, however, [23] proves a lower bound indicating no strategy can be optimal simultaneously over all smoothness classes; under extra information, they provide adaptive algorithms with near-optimal regret guarantees. Although achieving optimal regret for all settings is impossible, [17] design adaptive algorithms and prove that they are Pareto optimal. Our algorithms are mainly inspired by the ones in [17, 23]. A closely related line of work [28, 18, 5, 26] aims at minimizing simple regret in the continuum-armed bandit setting.

Adaptivity to unknown parameters. [9] argues the awareness of regularity is flawed and one should design algorithms that can adapt to the unknown environment. In situations where the goal is pure exploration or simple regret minimization, [18, 28, 16, 5, 26] achieve near-optimal guarantees with unknown regularity because their objectives trade-off exploitation in favor of exploration. In the case of cumulative regret minimization, however, [23] shows no strategy can be optimal simultaneously over all smoothness classes. In special situations or under extra information, [9, 10, 23] provide algorithms that adapt in different ways. [17] borrows the concept of Pareto optimality from economics and provide algorithms with rate functions that are Pareto optimal. Adaptivity is studied in statistics.
as well: in some cases, only additional logarithmic factors are required \cite{22,7}; in others, however, there exists an additional polynomial cost of adaptation \cite{1}.

2 Problem statement and notation

We consider the multi-armed bandit instance $\nu = (\nu_1, \ldots, \nu_n)$ with $n$ probability distributions with means $\mu_i = \mathbb{E}_{X \sim \nu_i} [X] \in [0, 1]$. Let $\mu_* = \max_{i \in [n]} \{\mu_i\}$ be the highest mean and $S_* = \{i \in [n] : \mu_i = \mu_*\}$ denote the subset of best arms. The cardinality $|S_*| = m$ is unknown to the learner. We could also generalize our setting to $S_* = \{i \in [n] : \mu_i \geq \mu_* - \epsilon(T)\}$ with unknown $|S'|$ (i.e., situations where there is an unknown number of near-optimal arms). Setting $\epsilon$ to be dependent on $T$ is to avoid an additive term linear in $T$, e.g., $\epsilon \leq 1/\sqrt{T} \Rightarrow \epsilon T \leq \sqrt{T}$. All theoretical results and algorithms presented in this paper are applicable to this generalized setting with minor modifications. For ease of exposition, we focus on the case with multiple best arms throughout the paper. At each time step $t \in [T]$, the algorithm/learner selects an action $A_t \in [n]$ and receives an independent reward $X_t \sim \nu_{A_t}$. We assume that $X_t - \mu_{A_t}$ is $(1/2)$-sub-Gaussian conditioned on $A_t$. We measure the success of an algorithm through the expected cumulative (pseudo) regret:

$$R_T = T \cdot \mu_* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right].$$

We use $\mathcal{R}(T, n, m)$ to denote the set of regret minimization problems with allowed time horizon $T$ and any bandit instance $\nu$ with $n$ total arms and $m$ best arms.\footnote{Throughout the paper, we denote by $[K]$ the set $\{1, \ldots, K\}$ for any positive integer $K$.} We emphasize that $T$ is part of the problem instance. We are particularly interested in the case when $n$ is comparable or even larger than $T$, which captures many modern applications where the available action space far exceeds the allowed time horizon. Although learning algorithms may not be able to pull each arm once, one should notice that the true/intrinsic hardness level of the problem could be viewed as $n/m$: selecting a subset uniformly at random with cardinality $\Theta(n/m)$ guarantees, with constant probability, the access to at least one best arm; but of course it is impossible to do this without knowing $m$. We quantify the intrinsic hardness level over a set of regret minimization problems $\mathcal{R}(T, n, m)$ as

$$\psi(\mathcal{R}(T, n, m)) = \inf\{\alpha \geq 0 : n/m \leq 2T^\alpha\},$$

where the constant $2$ in front of $T^\alpha$ is added to avoid otherwise the trivial case with all best arms when the infimum is $0$. $\psi(\mathcal{R}(T, n, m))$ is used here as it captures the minimax optimal regret over the set of regret minimization problem $\mathcal{R}(T, n, m)$, as explained later in our review of the MOSS algorithm and the lower bound. As smaller $\psi(\mathcal{R}(T, n, m))$ indicates easier problems, we then define the family of regret minimization problems with hardness level at most $\alpha$ as

$$\mathcal{H}_T(\alpha) = \{\cup \mathcal{R}(T, n, m) : \psi(\mathcal{R}(T, n, m)) \leq \alpha\},$$

with $\alpha \in [0, 1]$. Although $T$ is necessary to define a regret minimization problem, we actually encode the hardness level into a single parameter $\alpha$, which captures the tension between the complexity of bandit instance at hand and the allowed time horizon $T$: problems with different time horizons but the same $\alpha$ are equally difficult in terms of the achievable minimax regret (the exponent of $T$). We thus mainly study problems with $T$ large enough so that we could mainly focus on the polynomial terms of $T$. We are interested in designing algorithms with minimax guarantees over $\mathcal{H}_T(\alpha)$, but without the knowledge of $\alpha$.

MOSS and upper bound. In the classical setting, MOSS, designed by \cite{2} and further generalized to the sub-Gaussian case \cite{21} and improved in terms of constant factors \cite{14}, achieves the minimax optimal regret. In this paper, we will use MOSS as a subroutine with regret upper bound $O(\sqrt{nT})$ when $T \geq n$. For any problem in $\mathcal{H}_T(\alpha)$ with known $\alpha$, one could run MOSS on a subset selected uniformly at random with cardinality $O(T^\alpha)$ and achieve regret $O(T^{(1+\alpha)/2})$.

\footnote{We say a random variable $X$ is $\sigma$-sub-Gaussian if $\mathbb{E}[\exp(\lambda X)] \leq \exp(\sigma^2 \lambda^2/2)$ for all $\lambda \in \mathbb{R}$.}
Lower bound. The lower bound $\Omega(\sqrt{nT})$ in the classical setting does not work for our setting as its proof heavily relies on the existence of single best arm $[21]$. However, for problems in $\mathcal{H}_r(\alpha)$, we do have a matching lower bound $\Omega(T^{(1+\alpha)/2})$ as one could always apply the standard lower bound on an bandit instance with $n = T^\alpha$ and $m = 1$. For general value of $m$, a lower bound of the order $\Omega(\sqrt{T(n-m)/m}) = \Omega(T^{(1+\alpha)/2})$ for the $m$-best arms case could be obtained following similar analysis in Chapter 15 of $[21]$. Although $\log T$ may appear in our bounds, throughout the paper, we focus on problems with $T \geq 2$ as otherwise the bound is trivial.

3 An adaptive algorithm

Algorithm 1 takes time horizon $T$ and a user-specified $\beta \in [1/2, 1]$ as input, and it is mainly inspired by $[17]$. Algorithm 1 operates in iterations with geometrically-increasing length $\Delta T_i = 2^{p+i}$ with $p = \lceil \log_2 T^\beta \rceil$. At each iteration $i$, it restarts $\text{Moss}$ on a set $S_i$ consisting of $K_i = 2^{p+2-i}$ real arms selected uniformly at random plus a set of “virtual” mixture-arms (one from each of the $1 \leq j < i$ previous iterations, none if $i = 1$). The mixture-arms are constructed as follows. After each iteration $i$, let $\hat{p_i}$ denote the vector of empirical sampling frequencies of the arms in that iteration (i.e., the $k$-th element of $\hat{p_i}$ is the number of times arm $k$, including all previously constructed mixture-arms, was sampled in iteration $i$ divided by the total number of samples $\Delta T_i$). The mixture-arm for iteration $i$ is the $\hat{p_i}$-mixture of the arms, denoted by $\tilde{\nu_i}$. When $\text{Moss}$ samples from $\tilde{\nu_i}$, it first draws $i_* \sim \hat{p_i}$, then draws a sample from the corresponding arm $\nu_{i_*}$ (or $\nu_{i*}$). The mixture-arms provide a convenient summary of the information gained in the previous iterations, which is key to our theoretical analysis. Although our algorithm is working on fewer regular arms in later iterations, information summarized in mixture-arms is good enough to provide guarantees. We name our algorithm $\text{Moss}^{++}$ as it restarts $\text{Moss}$ at each iteration with past information summarized in mixture-arms. We provide an anytime version of Algorithm 1 in Appendix A.2 via the standard doubling trick.

Algorithm 1: $\text{Moss}^{++}$

**Input:** Time horizon $T$ and user-specified parameter $\beta \in [1/2, 1]$.

1. **Set:** $p = \lceil \log_2 T^\beta \rceil$, $K_i = 2^{p+2-i}$ and $\Delta T_i = \min\{2^{p+i}, T\}$.
2. **for** $i = 1, \ldots, p$ **do**
   3. Run $\text{Moss}$ on a subset of arms $S_i$ for $\Delta T_i$ rounds. $S_i$ contains $K_i$ real arms selected uniformly at random and the set of virtual mixture-arms from previous iterations, i.e., $\{\tilde{\nu}_j\}_{j<i}$.
   4. Construct a virtual mixture-arm $\tilde{\nu}_i$ based on empirical sampling frequencies of $\text{Moss}$ above.
5. **end for**

3.1 Analysis and discussion

We use $\mu_S = \max_{x \in S}\{E_{X \sim \nu}[X]\}$ to denote the highest expected reward over a set of distribution-s/arms $S$. For any algorithm that only works on $S$, we can decompose the regret into approximation error and learning error:

$$R_T = \underbrace{T \cdot (\mu_* - \mu_S)}_{\text{approximation error due to the selection of } S} + \underbrace{T \cdot \mu_S - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]}_{\text{learning error due to the sampling rule } \{A_t\}_{t=1}^{T}}.$$

This type of regret decomposition was previously used in $[20, 3, 17]$ to deal with the continuum-armed bandit problem. We consider here a probabilistic version, with randomness in the selection of $S$, for the classical setting.

The main idea behind providing guarantees for $\text{Moss}^{++}$ is to decompose its regret at each iteration, using Eq. (1), and then bound the expected approximation error and learning error separately. The expected learning error at each iteration could always be controlled as $\tilde{O}(T^\beta)$ thanks to regret guarantees for $\text{Moss}$ and specifically chosen parameters $p, K_i, \Delta T_i$. Let $i_*$ be the largest integer such that $K_i \geq 2T^\alpha \log \sqrt{T}$ still holds. The expected approximation error in iteration $i \leq i_*$ could be
upper bounded by $\sqrt{T}$ following an analysis on hypergeometric distribution. As a result, the expected regret in iteration $i \leq i^*$ is $O(T^\beta)$. Since the mixture-arm $\tilde{v}_{i^*}$ is included in all following iterations, we could further bound the expected approximation error in iteration $i > i^*$ by $O(T^{1+\alpha-\beta})$ after a careful analysis on $\Delta T_i/\Delta T_{i^*}$. This intuition is formally stated and proved in Theorem 1.

**Theorem 1.** Run MOSS++ with time horizon $T$ and an user-specified parameter $\beta \in [1/2, 1]$ leads to the following regret upper bound:

$$\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C (\log_2 T)^{5/2} \cdot T^{\min\{\beta, 1+\alpha-\beta\}} \cdot 1,$$

where $C$ is a universal constant.

**Remark 1.** We primarily focus on the polynomial terms in $T$ when deriving the bound, but put no effort in optimizing the polylog term. The 5/2 exponent of $\log_2 T$ might be tightened as well.

The theoretical guarantee is closely related to the user-specified parameter $\beta$: when $\beta > \alpha$, we suffer a multiplicative cost of adaptation $O(T^{(2\beta-\alpha-1)/2})$, with $\beta = (1+\alpha)/2$ hitting the sweet spot, comparing to non-adaptive minimax regret; when $\beta \leq \alpha$, there is essentially no guarantees. One may hope to improve this result. However, our analysis in Section 4 indicates: (1) achieving minimax optimal regret for all settings simultaneously is impossible; and (2) the rate function achieved by MOSS++ is already Pareto optimal.

### 4 Lower bound and Pareto optimality

#### 4.1 Lower bound

In this section, we show that designing algorithms with the non-adaptive minimax optimal guarantee over all values of $\alpha$ is impossible. We first state the result in the following general theorem.

**Theorem 2.** For any $0 \leq \alpha' < \alpha \leq 1$, assume $T^n \leq B$ and $[T^n] - 1 \geq \max\{T^n/4, 2\}$. If an algorithm is such that $\sup_{\omega \in \mathcal{H}_T(\alpha')} R_T \leq B$, then the regret of this algorithm is lower bounded on $\mathcal{H}_T(\alpha)$:

$$\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \geq 2^{-10} T^{1+\alpha} B^{-1}. \tag{2}$$

To give an interpretation of Theorem 2, we consider any algorithm/policy $\pi$ together with regret minimization problems $\mathcal{H}_T(\alpha')$ and $\mathcal{H}_T(\alpha)$ satisfying corresponding requirements. On one hand, if algorithm $\pi$ achieves a regret that is order-wise larger than $O(T^{(1+\alpha')/2})$ over $\mathcal{H}_T(\alpha')$, it is already not minimax optimal for $\mathcal{H}_T(\alpha')$. Now suppose $\pi$ achieves a near-optimal regret, i.e., $O(T^{(1+\alpha')/2})$, over $\mathcal{H}_T(\alpha')$; then, according to Eq. 2, $\pi$ must incur a regret of order at least $O(T^{1/2+\alpha-\alpha'/2})$ on one problem in $\mathcal{H}_T(\alpha')$. This, on the other hand, makes algorithm $\pi$ strictly sub-optimal over $\mathcal{H}_T(\alpha)$.

#### 4.2 Pareto optimality

We capture the performance of any algorithm by its dependence on polynomial terms of $T$ in the asymptotic sense. Note that the hardness level of a problem is encoded in $\alpha$.

**Definition 1.** Let $\theta : [0, 1] \to [0, 1]$ denote a non-decreasing function. An algorithm achieves the rate function $\theta$ if

$$\forall \epsilon > 0, \forall \alpha \in [0, 1], \lim_{T \to \infty} \sup_{\omega \in \mathcal{H}_T(\alpha)} \frac{R_T}{T^{\theta(\alpha)+\epsilon}} < +\infty.$$

Recall that a function $\theta'$ is strictly smaller than another function $\theta$ in pointwise order if $\theta'(\alpha) \leq \theta(\alpha)$ for all $\alpha$ and $\theta'(\alpha_0) < \theta(\alpha_0)$ for at least one value of $\alpha_0$. As there may not always exist a pointwise ordering over rate functions, following [17], we consider the notion of Pareto optimality over rate functions achieved by some algorithms.

**Definition 2.** A rate function $\theta$ is Pareto optimal if it is achieved by an algorithm, and there is no other algorithm achieving a strictly smaller rate function $\theta'$ in pointwise order. An algorithm is Pareto optimal if it achieves a Pareto optimal rate function.
Combining the results in Theorem 1 and Theorem 2 with above definitions, we could further obtain the following result in Theorem 3.

**Theorem 3.** The rate function achieved by MOSS++ with any $\beta \in [1/2, 1]$, i.e.,
\[
\theta_{\beta} : \alpha \mapsto \min\{\max\{\beta, 1 + \alpha - \beta\}, 1\},
\]

is Pareto optimal.

Fig. 1 provides an illustration of the rate functions achieved by MOSS++ with different $\beta$ as input, as well as the non-adaptive minimax optimal rate.

**Remark 2.** One should notice that the naive algorithm running MOSS on a subset selected uniformly at random with cardinality $O(T^{\beta'})$ is not Pareto optimal, since running MOSS++ with $\beta = (1 + \beta')/2$ leads to a strictly smaller rate function. The algorithm provided in [12], if transferred to our setting and allowing time horizon dependent quantile, is not Pareto optimal as well since it corresponds to the rate function $\theta(\alpha) = \max\{2.89\alpha, 0.674\}$.

![Figure 1: Pareto optimal rates](image)

## 5 Learning with extra information

Although previous Section 4 gives negative results on designing algorithms that could optimally adapt to all settings, one could actually design such an algorithm with extra information. In this section, we provide an algorithm that takes the expected reward of the best arm $\mu_*$ (or an estimated one with error up to $1/\sqrt{T}$) as extra information, and achieves near minimax optimal regret over all settings simultaneously. Our algorithm is mainly inspired by [23].

### 5.1 Algorithm

We name our Algorithm 3 Parallel as it maintains $\lceil \log T \rceil$ instances of subroutine, i.e., Algorithm 2 in parallel. Each subroutine $\text{SR}_i$ is initialized with time horizon $T$ and hardness level $\alpha_i = i/\lceil \log T \rceil$. We use $T_{i,t}$ to denote the number of samples allocated to $\text{SR}_i$ up to time $t$, and represent its empirical regret at time $t$ as $\hat{R}_{i,t} = T_{i,t} \cdot \mu_* - \sum_{t=1}^{T_{i,t}} X_{i,t}$ with $X_{i,t} \sim \nu_{A_{i,t}}$, being the $t$-th empirical reward obtained by $\text{SR}_i$ and $A_{i,t}$ being the index of the $t$-th arm pulled by $\text{SR}_i$.

**Algorithm 2: MOSS Subroutine (SR)**

**Input:** Time horizon $T$ and hardness level $\alpha$.

1. Select a subset of arms $S_{\alpha}$ uniformly at random with $|S_{\alpha}| = \lceil 2T^{\alpha} \log \sqrt{T} \rceil$ and run MOSS on $S_{\alpha}$.

Parallel operates in iterations of length $\lceil \sqrt{T} \rceil$. At the beginning of each iteration, i.e., at time $t = i \cdot \lceil \sqrt{T} \rceil$ for $i \in \{0\} \cup \lceil \sqrt{T} \rceil - 1$, Parallel first selects the subroutine with the lowest (breaking ties arbitrarily) empirical regret so far, i.e., $k = \arg \min_{i \in \lceil \log T \rceil} \hat{R}_{i,t}$; it then resumes the learning process of $\text{SR}_k$, from where it halted, for another $\lceil \sqrt{T} \rceil$ more pulls. All the information is updated at the end of that iteration. An anytime version of Algorithm 3 is provided in Appendix C.3.

### 5.2 Analysis

As Parallel discretizes the hardness parameter over a grid with interval $1/\lceil \log T \rceil$, we first show that running the best subroutine alone leads to regret $\tilde{O}(T^{(1+\alpha)/2})$.
We compare our algorithms with baselines on regret minimization problems with different hardness levels. For this experiment, we generate best arms with expected reward 0.9 and sub-optimal arms with time horizon $T$. In particular, all subroutines should achieve regret $\tilde{O}(T^{1+\alpha}/2)$, as the best subroutine does. Parallel then achieves the non-adaptive minimax optimal regret, up to polylog factors, without knowing the true hardness level $\alpha$.

Theorem 4. For any $\alpha \in [0, 1]$ unknown to the learner, run Parallel with time horizon $T$ and optimal expected reward $\mu_\star$ leads to the following regret upper bound:

$$\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C \cdot (\log T)^2 T^{(1+\alpha)/2},$$

where $C$ is a universal constant.

6 Experiments

We conduct three experiments to compare our algorithms with baselines. In Section 6.1, we compare the performance of each algorithm on problems with varying hardness levels. We examine how the regret curve of each algorithm increases on synthetic and real-world datasets in Section 6.2 and Section 6.3, respectively.

We first introduce the nomenclature of the algorithms. We use MOSS to denote the standard MOSS algorithm; and MOSS Oracle to denote Algorithm 2 with known $\alpha$. Quantile represents the algorithm (QRM2) proposed by [12] to minimize the regret with respect to the $(1-\rho)$-th quantile of means among arms, without the knowledge of $\rho$. One could easily transfer Quantile to our settings with top-$\rho$ fraction of arms treated as best arms. As suggested in [12], we reuse the statistics obtained in previous iterations of Quantile to improve its sample efficiency. We use MOSS++ to represent the vanilla version of Algorithm 1 and use empMOSS++ to represent an empirical version such that: (1) empMOSS++ reuse statistics obtained in previous round, as did in Quantile; and (2) instead of selecting $K_i$ real arms uniformly at random at the $i$-th iteration, empMOSS++ selects $K_i$ arms with the highest empirical mean for $i > 1$. We choose $\beta = 0.5$ for MOSS++ and empMOSS++ in all experiments. All results are averaged over 100 experiments. Shaded area represents 0.5 standard deviation for each algorithm.

6.1 Adaptivity to hardness level

We compare our algorithms with baselines on regret minimization problems with different hardness levels. For this experiment, we generate best arms with expected reward 0.9 and sub-optimal arms with time horizon $T$. In particular, all subroutines should achieve regret $\tilde{O}(T^{1+\alpha}/2)$, as the best subroutine does. Parallel then achieves the non-adaptive minimax optimal regret, up to polylog factors, without knowing the true hardness level $\alpha$.

Increasing $\beta$ generally leads to worse performance on problems with small $\alpha$ but better performance on problems with large $\alpha$. 

---

Algorithm 3: Parallel

Input: Time horizon $T$ and the optimal reward $\mu_\star$.

1: set: $p = \lceil \log T \rceil$, $\Delta = \lceil \sqrt{T} \rceil$ and $t = 0$.
2: for $i = 1, \ldots, p$ do
3: \hspace{1em} Set $\alpha_i = i/p$, initialize $\mathcal{SR}_i$ with $\alpha_i$, $T$; set $T_{i,t} = 0$, and $\hat{R}_{i,t} = 0$.
4: end for
5: for $i = 1, \ldots, \Delta - 1$ do
6: \hspace{1em} Select $k = \arg\min_{i \in [p]} \hat{R}_{i,t}$ and run $\mathcal{SR}_k$ for $\Delta$ rounds.
7: \hspace{1em} Update $T_{k,t} = T_{k,t} + \Delta$, $\hat{R}_{k,t} = \hat{R}_{k,t} + \sum_{t'=1}^{T_{k,t}} X_{k,t'}$, $t = t + \Delta$.
8: end for

Lemma 1. Suppose $\alpha$ is the true hardness parameter and $\alpha_i - 1/\lceil \log T \rceil < \alpha \leq \alpha_i$, run Algorithm 2 with time horizon $T$ and $\alpha_i$ leads to the following regret bound:

$$\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C \cdot \log T \cdot T^{(1+\alpha)/2},$$

where $C$ is a universal constant.

Since Parallel always allocates new samples to the subroutine with the lowest empirical regret so far, we know that the regret of every subroutine should be roughly of the same order at time $T$. In particular, all subroutines should achieve regret $\tilde{O}(T^{1+\alpha}/2)$, as the best subroutine does. Parallel then achieves the non-adaptive minimax optimal regret, up to polylog factors, without knowing the true hardness level $\alpha$.
with expected reward evenly distributed among \{0.1, 0.2, 0.3, 0.4, 0.5\}. All arms follow Bernoulli distribution. We set the time horizon to \( T = 50000 \) and consider the total number of arms \( n = 20000 \). We vary \( \alpha \) from 0.1 to 0.8 (with interval 0.1) to control the number of best arms \( m = \lceil n/2T^\alpha \rceil \) and thus the hardness level. In Fig. 2(a), the regret of any algorithm gets larger as \( \alpha \) increases, which is expected. MOSS does not provide satisfying performance due to the large action space and the relatively small time horizon. Although implemented in an anytime fashion, Quantile could be roughly viewed as an algorithm that runs MOSS on a subset selected uniformly at random with cardinality \( T^{0.347} \). Quantile displays good performance when \( \alpha = 0.1 \), but suffers regret much worse than MOSS++ and \( \text{empMOSS++} \) when \( \alpha \) gets larger. Note that the regret curve of Quantile gets flattened at \( 20000 \) is expected: it simply learns the best sub-optimal arm and suffers a regret \( 50000 \times (0.9 - 0.5) \). Although Parallel enjoys near minimax optimal regret, the regret it suffers from is the summation of 11 subroutines, which hurts its empirical performance. \( \text{empMOSS++} \) achieves performance comparable to MOSS Oracle when \( \alpha \) is small, and achieve the best empirical performance when \( \alpha \geq 0.3 \). When \( \alpha \geq 0.7 \), MOSS Oracle needs to explore most/all of the arms to statistically guarantee the finding of at least one best arm, which hurts its empirical performance.

### 6.2 Regret curve comparison

We compare how the regret curve of each algorithm increases in Fig. 2(b). We consider the same regret minimization configurations as described in Section 6.1 with \( \alpha = 0.25 \), \( \text{empMOSS++} \), MOSS++ and Parallel all outperform Quantile with \( \text{empMOSS++} \) achieving the performance closest to MOSS Oracle. MOSS Oracle, Parallel and \( \text{empMOSS++} \) have flattened their regret curve indicating they could confidently recommend the best arm. The regret curves of MOSS++ and Quantile do not flat as the random-sampling component in each of their iterations encourage them to explore new arms. Comparing to MOSS++, Quantile keeps increasing its regret at a much faster rate and with a much larger variance, which empirically confirms the sub-optimality of their regret guarantees.

### 6.3 Real-world dataset

We also compare all algorithms in a realistic setting of recommending funny captions to website visitors. We use a real-world dataset from the New Yorker Magazine Cartoon Caption Contest\footnote{https://www.newyorker.com/cartoons/contest} \footnote{Available online at https://nextml.github.io/caption-contest-data}. The dataset of 1-3 star caption ratings/rewards for Contest 652 consists of \( n = 10025 \) captions.\footnote{6} We use the ratings to compute Bernoulli reward distributions for each caption as follows. The mean of each caption/arm \( i \) is calculated as the percentage \( p_i \) of its ratings that were funny or somewhat funny (i.e., 2 or 3 stars). We normalize each \( p_i \) with the best one and then threshold each: if \( p_i \geq 0.8 \), then put \( p_i = 1 \); otherwise leave \( p_i \) unaltered. This produces a set of \( m = 54 \) best arms with rewards 1 and all...
other 9971 arms with rewards among $[0, 0.8]$. We set $T = 10^5$ and this results in a hardness level around $\alpha \approx 0.43$.

Using these Bernoulli reward models, we compare the performance of each algorithm, as shown in Fig. 3. MOSS, MOSS Oracle, Parallel and empMOSS++ have flattened their regret curve indicating they could confidently recommend the funny captions (i.e., best arms). Although MOSS could eventually identify a best arm in this problem, its cumulative regret is more than 7x of the regret achieved by empMOSS++ due to its initial exploration phase. The performance of Quantile is even worse, and its cumulative regret is more than 9x of the regret achieved by empMOSS++. One surprising phenomenon is that empMOSS++ outperforms MOSS Oracle in this realistic setting. Our hypothesis is that MOSS Oracle is a little bit conservative and selects an initial set with cardinality too large. This experiment demonstrates the effectiveness of empMOSS++ and MOSS++ in modern applications of bandit algorithm with large action space and limited time horizon.

7 Conclusion

We study a regret minimization problem with large action space but limited time horizon, which captures many modern applications of bandit algorithms. Depending on the number of best/near-optimal arms, we encode the hardness level, in terms of minimax regret achievable, of the given regret minimization problem into a single parameter $\alpha$, and we design algorithms that could adapt to this unknown hardness level. Our first algorithm MOSS++ takes a user-specified parameter $\beta$ as input and provides guarantees as long as $\alpha < \beta$; our lower bound further indicates the rate function achieved by MOSS++ is Pareto optimal. Although no algorithm can achieve near minimax optimal regret over all $\alpha$ simultaneously, as demonstrated by our lower bound, we overcome this limitation with an (often) easily-obtained extra information and propose Parallel that is near-optimal for all settings. Inspired by MOSS++. We also propose empMOSS++ with excellent empirical performance. Experiments on both synthetic and real-world datasets demonstrate the efficiency of our algorithms over the previous state-of-the-art.

Broader Impact

This paper provides efficient algorithms that work well in modern applications of bandit algorithms with large action space but limited time horizon. We make minimal assumption about the setting, and our algorithms can automatically adapt to unknown hardness levels. Worst-case regret guarantees are provided for our algorithms; we also show MOSS++ is Pareto optimal and Parallel is minimax optimal, up to polylog factors. empMOSS++ is provided as a practical version of MOSS++ with excellent empirical performance. Our algorithms are particularly useful in areas such as e-commerce and movie/content recommendation, where the action space is enormous but possibly contains multiple best/satisfactory actions. If deployed, our algorithms could automatically adapt to the hardness level of the recommendation task and benefit both service-providers and customers through efficiently delivering satisfactory content. One possible negative outcome is that items recommended to a specific user/customer might only come from a subset of the action space. However, this is unavoidable when the number of items/actions exceeds the allowed time horizon. In fact, one should notice that all items/actions will be selected with essentially the same probability, thanks to the incorporation of random selection processes in our algorithms. Our algorithms will not leverage/create biases due to the same reason. Overall, we believe this paper’s contribution will have a net positive impact.
Acknowledgments and Disclosure of Funding

The authors would like to thank anonymous reviewers for their comments and suggestions. This work was partially supported by NSF grant no. 1934612.

References


A\ Omitted proofs for Section 3

We introduce the notation $R_{T\mid\mathcal{F}} = T \cdot \mu_* - \mathbb{E}[\sum_{t=1}^T X_t \mid \mathcal{F}]$ for any $\sigma$-algebra. One should also notice that $\mathbb{E}[R_{T\mid\mathcal{F}}] = R_T$.

A.1 Proof of Theorem 1

**Lemma 2.** For an instance with $n$ total arms and $m$ best arms, and for a subset $S$ selected uniformly at random with cardinality $k$, the probability that none of the best arms are selected in $S$ is upper bounded by $\exp(-mk/n)$.

**Proof.** Consider selecting $k$ items out of $n$ items without replacement; and suppose there are $m$ target items. Let $\mathcal{E}$ denote the event where none of the target items are selected, we then have

$$\mathbb{P}(\mathcal{E}) = \frac{n-m}{n} \cdot \frac{(n-m)!}{n!} \cdot \prod_{i=0}^{k-1} \frac{n-m-i}{n-i}$$

$$\leq \exp\left(-\frac{m}{n} \cdot k\right),$$

where Eq. (4) comes from the fact that $\frac{n-m-i}{n-i}$ is decreasing in $i$; and Eq. (5) comes from the fact that $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$.

Selecting arms with replacement gives the same guarantee (which directly goes to Eq. (4)), and can be used in corner cases when $k > n$.

**Theorem 1.** Run MOSS++ with time horizon $T$ and an user-specified parameter $\beta \in [1/2, 1]$ leads to the following regret upper bound:

$$\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C \left(\log_2 T\right)^{5/2} \cdot T^{\min\{\beta, 1 + \alpha - \beta\}} \cdot 1,$$

where $C$ is a universal constant.

**Proof.** Let $T_i = \sum_{j=1}^i \Delta T_j$. We first notice that Algorithm 1 is a valid algorithm in the sense that it selects an arm $A_t$ for any $t \in [T]$, i.e., it does not terminate before time $T$: the argument is clearly true if there exists $i \in [p]$ such that $\Delta T_i = T$; otherwise, we can show that

$$T_p = \sum_{i=1}^p \Delta T_i = 2(2^p - 1) \geq 2^p \geq T,$$

for all $\beta \in [1/2, 1]$.

We will only consider the case when $\alpha < \beta$ in the following since otherwise Theorem 1 trivially holds due to $T^{1+\alpha - \beta} \geq T$.

Let $\mathcal{F}_{i-1}$ represents the information available at the beginning of iteration $i$, including the random selection process of generating $S_i$. We denote $R_{\Delta T_i} = \Delta T_i \cdot \mu_* - \mathbb{E}[\sum_{t=T_i+1}^T X_t]$ the expected cumulative regret at iteration $i$. Recall that we use $R_{\Delta T_i \mid \mathcal{F}_{i-1}}$ to represent the expected regret conditioned on $\mathcal{F}_{i-1}$ and have $\mathbb{E}[R_{\Delta T_i \mid \mathcal{F}_{i-1}}] = R_{\Delta T_i}$; we also use $\mu_{S_i} = \max_{p \in S_i} \{\mathbb{E}_{X \sim p}[X]\}$. Applying Eq. (1) on $R_{\Delta T_i \mid \mathcal{F}_{i-1}}$ leads to

$$R_{\Delta T_i \mid \mathcal{F}_{i-1}} = \Delta T_i \cdot (\mu_* - \mu_{S_i}) + \left(\Delta T_i \cdot \mu_{S_i} - \mathbb{E}\left[\sum_{t=T_i+1}^T \mu_{A_t} \mid \mathcal{F}_{i-1}\right]\right),$$

for all $i \in [T]$.\]
where, by a slightly abuse of notations, we use $\mu_{A_t}$ to refer to the mean of arm $A_t \in S_i$, which could also be the mean of a virtual arm constructed in one of the previous iterations.

Note that we will be working on a bandit instance with $(\sqrt{2/2})$-sub-Gaussian noise when the extra virtual arms are included: let $X$ be a sample from a virtual mixture-arm $\tilde{\nu}_i$, which is realized by first sampling an index $i_j$ (of a real arm) from the empirical measure, and then draw $X$ from the real arm $\nu_{i_j}$. We have $X - E[X] = (X - \mu_{i_j}) + (\mu_{i_j} - E[X])$ and thus for any $\lambda \in \mathbb{R}$,

$$
E[\exp (\lambda (X - E[X]))] = E \left[ E \left[ \exp \left( (X - \mu_{i_j}) \right) \right] | i_j \right] \\
= E \left[ \exp \left( \lambda (\mu_{i_j} - E[X]) \right) \right] \\
\leq \exp \left( \frac{\lambda^2/4}{2} \right) E \left[ \exp \left( \lambda (\mu_{i_j} - E[X]) \right) \right] \\
\leq \exp \left( \frac{\lambda^2/4 + \lambda^2/4}{2} \right) \\
= \exp \left( \frac{\lambda^2}{2} \right)
$$

where Eq. (7) comes from the fact that $\mu_{i_j} \in [0, 1]$ and $E[\mu_{i_j}] = E[X]$. In the following, we’ll directly plug in the regret bound of MOSS for the $1$-sub-Gaussian case.

We first consider the learning error for any iteration $i \in [p]$. Although $\mu_{S_i}$ is random, it is fixed at time $T_{i-1} + 1$ [17]. Since MOSS restarts at each iteration, conditioning on the information available at the beginning of the $i$-th iteration, i.e., $\mathcal{F}_{i-1}$, and apply the regret bound for MOSS, we have:

$$
\Delta T_i \cdot \mu_{S_i} - E \left[ \sum_{t=T_{i-1}+1}^{T_i} \mu_{A_t} | \mathcal{F}_{i-1} \right] \leq 39 \sqrt{|S_i| \Delta T_i + |S_i|}
$$

$$
= 39 \sqrt{(K_i + i - 1) \Delta T_i + (K_i + i - 1)} \\
\leq 39 \sqrt{K_i \Delta T_i + (p - 1) \Delta T_i + (K_i + p - 1)} \\
\leq 39 \sqrt{22^{p+2} + (p - 1)T + 2^{p+1} + (p - 1)} \\
\leq 39 \sqrt{16T^{2\beta} + \log_2(T^{\beta})T + 4T^{\beta} + \log_2 T^{\beta}} \\
\leq 166 (\log_2 T)^{1/2} \cdot T^{\beta}
$$

where Eq. (8) comes from the guarantee of MOSS [21]; Eq. (9) comes from $i \leq p$; Eq. (10) comes from the definition of $K_i$ and $\Delta T_i$; Eq. (11) comes from the fact that $p = [\log_2 T^{\beta}] \leq \log_2 T^{\beta} + 1$; Eq. (12) comes from some trivial boundings on the constants.

Taking expectation over all randomness on Eq. (6), we obtain

$$
R_{\Delta T_i} \leq \Delta T_i \cdot E \left[ (\mu_* - \mu_{S_i}) \right] + 166 (\log_2 T)^{1/2} \cdot T^{\beta}.
$$

Now, we only need to consider the first term, i.e., the expected approximation error over the $i$-th iteration. Let $\mathcal{E}_i$ denote the event that none of the best arms, among regular arms, is selected in $S_i$, according to Lemma 2 we further have

$$
\Delta T_i \cdot E \left[ (\mu_* - \mu_{S_i}) \right] \leq \Delta T_i \cdot E \left[ (0 \cdot \mathbb{P}(-\mathcal{E}_i) + 1 \cdot \mathbb{P}(\mathcal{E}_i)) \right] \\
\leq \Delta T_i \cdot \exp(-K_i/(2T^\alpha))
$$

where we use the fact the $\mu_i \in [0, 1]$ in Eq. (14); and directly plug $n/m \leq 2T^\alpha$ into Eq. (5) to get Eq. (15).

Let $i_* \in [p]$ be the largest integer, if exists, such that $K_{i_*} \geq 2T^\alpha \log \sqrt{T}$, we then have that, for any $i \leq i*$,

$$
\Delta T_i \cdot E \left[ (\mu_* - \mu_{S_i}) \right] \leq \Delta T_i \sqrt{T} \leq T \sqrt{T} \leq \sqrt{T}.
$$

\footnote{One can remove the $(\log_2 T)^{1/2}$ term in many cases, e.g., when $\beta > 1/2$ and $T$ is large enough (with respect to $\beta$). However, we mainly focus on the polynomial terms here.}
We now analyze the expected approximation error for iteration $i_*$.

If we have $K_1 < 2T^\alpha \log \sqrt{T}$, we then set $i_* = 1$. Notice that $K_1 = 2^{p+1} = 2^{\lceil \log_2 T^\beta \rceil + 1} \geq 2T^\beta > 2T^\alpha$, we then have

$$\Delta T_1 \cdot E[(\mu_* - \mu_S)] \leq \Delta T_1 \exp(-1) \leq 2^{p+1} \exp(-1) < 2T^\beta.$$  \hfill (17)

Combining Eq. (13) with Eq. (16) or Eq. (17), we have for any $i \leq i_*$, and in particular for $i = i_*$, we have

$$R_{\Delta T_i} \leq \max\{\sqrt{T}, 2T^\beta\} + 166 \log_2 T)^{1/2} \cdot T^\beta$$

$$\leq 168 \log_2 T)^{1/2} \cdot T^\beta.$$  \hfill (18)

In the case when $i_* = p$ or when $\Delta T_i = \min\{2^{p+i}, T\} = T$, we know that MOSS++ will in fact stop at a time step no larger than $T_{i_*}$ (since the allowed time horizon is $T$), and incur no regret in iterations $i > i_*$. In the following, we only consider the case when $i_* < p$ and $\Delta T_i = 2^{p+i}$. As a result, we have $K_1 \Delta T_i = 2^{p+2}$ and thus

$$\Delta T_i = \frac{2^{p+2}}{K_i} > \frac{2^{p+1}}{T^\alpha \log T},$$  \hfill (19)

where Eq. (19) comes from the fact that $K_i < \max\{2T^\alpha \log T, 2T^\alpha \log \sqrt{T}\} = 2T^\alpha \log T$ by definition of $i_*$. We now analysis the expected approximation error for iteration $i > i_*$. Since the sampling information during the $i_*$-th iteration is summarized in the virtual mixture-arm $\nu_i$, and being added to all $S_i$ for all $i > i_*$. Let $\tilde{\mu}_i = E_{X \sim \tilde{S}_i}[X]$ denote the expected reward of sampling according to the virtual mixture-arm $\tilde{\nu}_i$. For any $i > i_*$, we then have

$$\Delta T_i \cdot E[(\mu_* - \mu_S)] \leq \Delta T_i \cdot (\mu_* - E[\tilde{\mu}_i])$$

$$= \frac{\Delta T_i}{\Delta T_i} \cdot (\Delta T_i \cdot (\mu_* - E[\tilde{\mu}_i]))$$

$$= \frac{\Delta T_i}{\Delta T_i} \cdot \left(\Delta T_i \cdot \mu_* - \sum_{t=T_{i-1}+1}^{T_i} E[\mu_{A_t}]\right)$$

$$= \frac{\Delta T_i}{\Delta T_i} \cdot R_{\Delta T_i}$$

$$< \frac{2^{p+1}}{T^\alpha \log T} \cdot 168 \log_2 T)^{1/2} \cdot T^\beta$$

$$\leq \frac{T^{1+\alpha+\beta}}{2^{2p}} \cdot 84 \log_2 T)^{3/2}$$

$$\leq 84 \log_2 T)^{3/2} \cdot T^{1+\alpha-\beta},$$  \hfill (20)

where Eq. (20) comes from the fact that $\Delta T_i \leq T$ and some rewriting; Eq. (21) comes from the fact that $p = \lceil \log_2 T^\beta \rceil \geq \log_2 T^\beta$.

Combining Eq. (21) and Eq. (13) gives the following regret bound for iterations $i > i_*$:

$$R_{\Delta T_i} \leq 250 \log_2 T^{3/2} \cdot T^{\max\{\beta, 1+\alpha-\beta\}},$$

where the constant 250 simply comes from $84 + 166$.

Since the cumulative regret is non-decreasing in $t$, we have

$$R_T \leq \sum_{i=1}^{p} R_{\Delta T_i}$$

$$\leq 250 \cdot \log_2 T^{3/2} \cdot T^{\max\{\beta, 1+\alpha-\beta\}}$$

$$\leq 250 \cdot (\log_2 T + 1) \cdot (\log_2 T)^{3/2} \cdot T^{\max\{\beta, 1+\alpha-\beta\}}$$

$$\leq 251 \cdot (\log_2 T)^{5/2} \cdot T^{\max\{\beta, 1+\alpha-\beta\}},$$  \hfill (22)

where Eq. (22) comes from the fact that $p = \lceil \log_2 (T^\beta) \rceil \leq \log_2 (T^\beta) + 1 \leq \log_2 T + 1$. Our results follows after noticing $R_T \leq T$ is a trivial upper bound. \hfill □
A.2 Anytime version

**Algorithm 4:** Anytime version of MOSS++

**Input:** User specified parameter $\beta \in [1/2, 1]$.

1. **for** $i = 0, 1, \ldots$ **do**
   2. Run Algorithm 1 with parameter $\beta$ for $2^i$ rounds (note that we will set $p = \lceil \log_2 2^\beta \rceil = \lceil i \beta \rceil$).
   3. **end for**

**Corollary 1.** For any unknown time horizon $T$, run Algorithm 4 with an user-specified parameter $\beta \in [1/2, 1]$ leads to the following regret upper bound:

$$
\sup_{\omega \in H_T(\alpha)} R_T \leq C (\log_2 T)^{5/2} \cdot T^{\min\{\max\{\beta, 1+\alpha-\beta\}, 1\}},
$$

where $C$ is a universal constant.

**Proof.** Let $t_*$ be the smallest integer such that

$$
\sum_{i=0}^{t_*} 2^i = 2^{t_*+1} - 1 \geq T.
$$

We then only need to run Algorithm 1 for at most $t_*$ times. By the definition of $t_*$, we also know that $2^{t_*} \leq T$, which leads to $t_* \leq \log_2 T$.

Let $\gamma = \min\{\max\{\beta, 1+\alpha-\beta\}, 1\}$. From Theorem 1 we know that the regret at $i \in [t_*]$-th round, denoted as $R_{2^i}$, could be upper bounded by

$$
R_{2^i} \leq 251 (\log_2 2^{(2^i)^{5/2}} \cdot (2^i)^\gamma = 251 i^{5/2} \cdot (2^i)^\gamma \leq 251 i^{5/2} \cdot (2^i)^\gamma 
$$

For $i = 0$, we have $R_{2^0} \leq 1 \leq 251 (\log_2 T)^{5/2} \cdot (2^i)^0$ as well as long as $T \geq 2$.

Now for the unknown time horizon $T$, we could upper bound the regret by

$$
R_T \leq \sum_{i=0}^{t_*} R_{2^i}
\leq 251 (\log_2 T)^{5/2} \cdot \left( \sum_{i=0}^{t_*} (2^i)^\gamma \right)
\leq 251 (\log_2 T)^{5/2} \cdot \int_{x=0}^{t_*+1} (2^x)^\gamma dx
\leq 251 (\log_2 T)^{5/2} \cdot \frac{1}{\log 2^{\gamma}} \cdot (2^{t_*+1} - 1)
\leq \frac{2\gamma}{\log 2^\gamma} 251 (\log_2 T)^{5/2} \cdot T^\gamma
\leq 1449 (\log_2 T)^{5/2} \cdot T^\gamma,
$$

where Eq. (23) comes from upper bounding summation by integral; and Eq. (24) comes from a trivial bound on the constant when $1/2 \leq \gamma \leq 1$. \hfill \Box

B Omitted proofs for Section 4

B.1 Proof of Theorem 2

**Theorem 2.** For any $0 \leq \alpha' < \alpha \leq 1$, assume $T^\alpha \leq B$ and $|T^\alpha| - 1 \geq \max\{T^\alpha/4, 2\}$. If an algorithm is such that $\sup_{\omega \in H_T(\alpha')} R_T \leq B$, then the regret of this algorithm is lower bounded on $H_T(\alpha)$:

$$
\sup_{\omega \in H_T(\alpha)} R_T \geq 2^{-10} T^{1+\alpha} B^{-1}.
$$

(2)
The proof of Theorem 2 is mainly inspired by the proofs of lower bounds in [23, 17]. Before the start of the proof, we first state a generalized version of Pinsker’s inequality developed in [17] (Lemma 3 therein).

**Lemma 3.** Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures. For any random variable $Z \in [0, 1]$, we have

$$|\mathbb{E}_\mathbb{P}[Z] - \mathbb{E}_\mathbb{Q}[Z]| \leq \sqrt{\text{KL}(\mathbb{P}, \mathbb{Q})/2}.$$  

We consider $K + 1$ bandit instances $\{\nu_i\}_{i=0}^K$ such that each bandit instance is a collection of $n$ distributions $\nu_i = (\nu_{i1}, \nu_{i2}, \ldots, \nu_{in})$ where each $\nu_{ij}$ represents a Gaussian distribution $N(\mu_{ij}, 1/4)$ with $\mu_{ij} = \mathbb{E}[\nu_{ij}]$. For any given $0 < \alpha' < \alpha < 1$ and time horizon $T$ large enough, we choose $n, m_0, m, K \in \mathbb{N}$ such that the following three conditions are satisfied:

1. $n = m_0 + Km$;
2. $n/m_0 \leq 2T^\alpha$;
3. $n/m \leq 2T^\alpha$.

**Proposition 1.** Integers satisfying the above three conditions exist. For instance, we could first fix $m \in \mathbb{N}_+$ and set $K = \lceil T^\alpha \rceil - 1 \geq 2$. One could then set $m_0 = \lceil T^\alpha - \alpha' \rceil$ and $n = m_0 + Km$.

**Proof.** We notice that the first condition holds by construction. We now show that the second and the third conditions hold.

For the second condition, we have

$$\frac{n}{m_0} = \frac{m_0 + Km}{m_0} = 1 + \frac{m(\lceil T^\alpha \rceil - 1)}{m \lceil T^\alpha - \alpha' \rceil} \leq 1 + \frac{T^\alpha}{T^\alpha - \alpha'} \leq 2T^\alpha.$$

For the third condition, we have

$$\frac{n}{m} = \frac{m_0 + Km}{m} = \frac{m(\lceil T^\alpha - \alpha' \rceil + (\lceil T^\alpha \rceil - 1)m)}{m} = \lceil T^\alpha - \alpha' \rceil + \lceil T^\alpha \rceil - 1 \leq T^\alpha - \alpha' + T^\alpha \leq 2T^\alpha.$$

Now we group $n$ distribution into $K + 1$ different groups based on their indices: $S_0 = [m_0]$ and $S_i = [m_0 + i \cdot m] \setminus [m_0 + (i - 1) \cdot m]$. Let $\Delta \in (0, 1]$ be a parameter to be tuned later, we then define $K + 1$ bandit instances $\nu_i$ for $i \in \{0\} \cup [K]$ by assigning different values to their means $\mu_{ij}$:

$$\mu_{ij} = \begin{cases} \Delta/2 & \text{if } j \in S_0, \\ \Delta & \text{if } j \in S_i \text{ and } i \neq 0, \\ 0 & \text{otherwise}. \end{cases} \quad (25)$$

We could clearly see there are $m_0$ best arms in instance $\nu_0$ and $m$ best arms in instances $\nu_i, \forall i \in [K]$.

Based on our construction in Proposition 1, we could then conclude that, with time horizon $T$, the

---

* $K \geq 2$ holds for $T$ large enough.
regret minimization problem with respect to \( \nu_i \) is in \( \mathcal{H}_T(\alpha') \); and similarly the regret minimization problem with respect to \( \nu_i \) is in \( \mathcal{H}_T(\alpha) \), \( \forall i \in [K] \).

For any \( t \in [T] \), the tuple of random variables \( H_t = (A_1, X_1, \ldots, A_t, X_t) \) is the outcome of an algorithm interacting with an bandit instance up to time \( t \). Let \( \Omega_t = ([n] \times \mathbb{R})^t \subseteq \mathbb{R}^{2t} \) and \( \mathcal{F}_t = \mathcal{B}(\Omega_t) \); one could then define a measurable space \((\Omega_t, \mathcal{F}_t)\) for \( H_t \). The random variables \( A_1, X_1, \ldots, A_t, X_t \) that make up the outcome are defined by their coordinate projections:

\[
A_t(a_1, x_1, \ldots, a_t, x_t) = a_t \quad \text{and} \quad X_t(a_1, x_1, \ldots, a_t, x_t) = x_t.
\]

For any fixed algorithm/policy \( \pi \) and bandit instance \( \nu_i \), \( \forall i \in \{0\} \cup [K] \), we are now constructing a probability measure \( \mathbb{P}_{i,t} \) over \((\Omega_t, \mathcal{F}_t)\). Note that a policy \( \pi \) is a sequence \((\pi_t)_{t=1}^T \), where \( \pi_t \) is a probability kernel from \((\Omega_{t-1}, \mathcal{F}_{t-1})\) to \(([n], 2^n)\). For each \( i \), we define another probability kernel \( p_{i,t} \) from \((\Omega_{t-1} \times [n], \mathcal{F}_{t-1} \otimes 2^n)\) to \( (\mathbb{R}, \mathcal{B}(\mathbb{R})) \) that models the reward. Assuming the reward is distributed according to \( \mathcal{N}(\mu_{ia_t}, 1/4) \), we give its explicit expression for any \( B \in \mathcal{B}(\mathbb{R}) \) as:

\[
p_{i,t}(\{a_1, x_1, \ldots, a_t\}, B) = \int_B \sqrt{\frac{2}{\pi}} \exp \left( -2(x - \mu_{ia_t}) \right) dx.
\]

The probability measure over \( \mathbb{P}_{i,t} \) over \((\Omega_t, \mathcal{F}_t)\) could then be define recursively as \( \mathbb{P}_{i,t} = p_{i,t}(\mathbb{P}_{i,t-1}) \). We use \( E_0 \) to denote the expectation taken with respect to \( \mathbb{P}_{i,t} \). Apply the same analysis as on page 21 of [17], we obtain the following proposition on KL decomposition.

**Proposition 2.**

\[
\text{KL} (\mathbb{P}_{0,T}, \mathbb{P}_{i,t}) = E_0 \left[ \sum_{t=1}^T \text{KL} (\mathcal{N}(\mu_{0A_t}, 1/4), \mathcal{N}(\mu_{iA_t}, 1/4)) \right].
\]

With respect to notations and constructions described above, we now prove Theorem 2.

**Proof.** (Theorem 2) Let \( N_{S_i}(T) = \sum_{t=1}^T \mathbbm{1}(A_t \in S_i) \) denote the number of times the algorithm \( \pi \) selects an arm in \( S_i \) up to time \( T \). Let \( R_{i,T} \) denote the expected (pseudo) regret achieved by the algorithm \( \pi \) interacting with the bandit instance \( \nu_i \). Based on the construction of bandit instance in Eq. (25), we have

\[
R_{0,T} \geq \frac{\Delta}{2} \sum_{i=1}^K E_0 [N_{S_i}(T)], \quad (26)
\]

and \( \forall i \in [K] \),

\[
R_{i,T} \geq \frac{\Delta}{2} (T - E_i [N_{S_i}(T)]) = \frac{T\Delta}{2} \left( 1 - \frac{E_i [N_{S_i}(T)]}{T} \right). \quad (27)
\]

According to Proposition 2 and the calculation of KL-divergence between two Gaussian distributions, we further have

\[
\text{KL}(\mathbb{P}_{0,T}, \mathbb{P}_{i,T}) = E_0 \left[ \sum_{t=1}^T \text{KL} (\mathcal{N}(\mu_{0A_t}, 1/4), \mathcal{N}(\mu_{iA_t}, 1/4)) \right]
\]

\[
= E_0 \left[ \sum_{t=1}^T 2 (\mu_{0A_t} - \mu_{iA_t})^2 \right]
\]

\[
= 2 E_0 [N_{S_i}(T)] \Delta^2, \quad (28)
\]

where Eq. (28) comes from the fact that \( \mu_{0j} \) and \( \mu_{ij} \) only differs for \( j \in S_i \) and the difference is exactly \( \Delta \).
We now consider the average regret over $i \in [K]$:

$$
\frac{1}{K} \sum_{i=1}^{K} R_{i,T} \geq \frac{T}{2} \left( 1 - \frac{1}{K} \sum_{i=1}^{K} \frac{E_i[N_{S_i}(T)]}{T} \right) \\
\geq \frac{T}{2} \left( 1 - \frac{1}{K} \sum_{i=1}^{K} \left( \frac{E_0[N_{S_i}(T)]}{T} + \frac{\sqrt{KL(P_{0,T}, P_{i,T})}}{2} \right) \right) \\
= \frac{T}{2} \left( 1 - \frac{1}{K} \sum_{i=1}^{K} \frac{E_0[N_{S_i}(T)]}{T} - \frac{1}{K} \sum_{i=1}^{K} \sqrt{E_0[N_{S_i}(T)]} \Delta^2 \right) \\
\geq \frac{T}{2} \left( 1 - \frac{1}{K} \sum_{i=1}^{K} \frac{E_0[N_{S_i}(T)]}{T} - \frac{\sqrt{\sum_{i=1}^{K} E_0[N_{S_i}(T)]} \Delta^2}{K} \right) \\
\geq \frac{T}{2} \left( 1 - \frac{1}{K} - \frac{2\Delta R_{0,T}}{K} \right) \\
\geq \frac{T}{2} \left( 1 - \frac{2\Delta B}{K} \right),
$$

where Eq. (29) comes from applying Lemma 3 with $Z = N_{S_i}(T)/T$ and $P = P_{0,T}$ and $Q = P_{i,T}$; Eq. (30) comes from applying Eq. (28); Eq. (31) comes from concavity of $\sqrt{\cdot}$ and the fact that $\sum_{i=1}^{K} E_0[N_{S_i}(T)] \leq T$; Eq. (32) comes from applying Eq. (26); and finally Eq. (33) comes from the fact that $K \geq 2$ by construction and the assumption that $R_{0,T} \leq B$.

To obtain a large value for Eq. (33), one could maximize $\Delta$ while still make $\sqrt{2\Delta B/K} \leq 1/4$. Set $\Delta = 2^{−5}KB^{−1}$, following Eq. (35), we obtain

$$
\frac{1}{K} \sum_{i=1}^{K} R_{i,T} \geq 2^{−8}TKB^{−1} \\
= 2^{−8}T ([T^\alpha] - 1) B^{−1} \geq 2^{−10}T^{1+\alpha}B^{−1},
$$

where Eq. (34) comes from the construction of $K$; and Eq. (35) comes from the assumption that $[T^\alpha] - 1 \geq T^\alpha/4$.

Now we only need to make sure $\Delta = 2^{−5}KB^{−1} \leq 1$. Since we have $K = [T^\alpha] - 1 \leq T^\alpha$ by construction and $T^\alpha \leq B$ by assumption, we obtain $\Delta = 2^{−5}KB^{−1} \leq 2^{−5} < 1$ as desired.

### B.2 Proof of Theorem 3

**Lemma 4.** Suppose an algorithm achieves rate function $\theta$, then for any $0 < \alpha \leq \theta(0)$, we have

$$
\theta(\alpha) \geq 1 + \alpha - \theta(0).
$$

**Proof.** Fix $0 < \alpha \leq \theta(0)$. For any $\epsilon > 0$, there exists constant $c_1$ and $c_2$ such that for sufficiently large $T$,

$$
\sup_{\omega \in \mathcal{H}_T(0)} R_T \leq c_1T^{\theta(0)+\epsilon} \quad \text{and} \quad \sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq c_2T^{\theta(\alpha)+\epsilon}.
$$

Let $B = \max\{c_1, 1\} \cdot T^{\theta(0)+\epsilon}$, we could see that $T^\alpha \leq T^{\theta(0)} \leq B$ holds by assumption. For $T$ large enough, the condition $[T^\alpha] - 1 \geq \max\{T^{\alpha}/4, 2\}$ of Theorem 2 holds. We then have

$$
c_2T^{\theta(\alpha)+\epsilon} \geq 2^{−10}T^{1+\alpha} \left( \max\{c_1, 1\} \cdot T^{\theta(0)+\epsilon} \right)^{-1} = 2^{−10}T^{1+\alpha-\theta(0)-\epsilon}/\max\{c_1, 1\}.
$$

For $T$ sufficiently large, we then must have

$$
\theta(\alpha) + \epsilon \geq 1 + \alpha - \theta(0) - \epsilon.
$$

Let $\epsilon \to 0$ leads to the desired result. □
We consider the following two exclusive cases.

We need to prove that no other algorithms achieve strictly smaller rates in pointwise order.

\[ \text{MOSS} \]

where Eq. (39) comes from the regret bound of Theorem 3.

The rate function achieved by \[ \alpha \]

Lemma 1.

\[ \theta(\alpha) \geq \min \{ \max \{ \theta(0), 1 + \alpha - \theta(0) \}, 1 \}, \quad (37) \]

with \( \theta(0) \in [1/2, 1] \).

**Proof.** For any rate function \( \theta \) achieved by an algorithm, we first notice that \( \theta(\alpha) \geq \theta(\alpha') \) for any \( 0 \leq \alpha' < \alpha \leq 1 \) since \( \mathcal{H}_T(\alpha') \subseteq \mathcal{H}_T(\alpha) \); this also implies \( \theta(\alpha) \geq \theta(0) \). From Lemma 4 we further obtain \( \theta(\alpha) \geq 1 + \alpha - \theta(0) \) if \( \alpha \leq \theta(0) \). Thus, for any \( \alpha \in (0, \theta(0)) \), we have \( \theta(\alpha) \geq \max \{ \theta(0), 1 + \alpha - \theta(0) \} \).

(38)

Note that this indicates \( \theta(\theta(0)) = 1 \), as we trivially have \( R_T \leq T \). For any \( \alpha \in (\theta(0), 1) \), we have \( \theta(\alpha) \geq \theta(\theta(0)) = 1 \), which leads to \( \theta(\alpha) = 1 \) for \( \alpha \in (\theta(0), 1] \). To summarize, we obtain the desired result in Eq. 37. We have \( \theta(0) \in [1/2, 1] \) since the minimax optimal rate among problems in \( \mathcal{H}_T(0) \) is 1/2.

**Theorem 3.** The rate function achieved by \( \text{MOSS}^+ \) with any \( \beta \in [1/2, 1] \), i.e.,

\[ \theta_\beta : \alpha \mapsto \min \{ \max \{ \beta, 1 + \alpha - \beta \}, 1 \}, \quad (3) \]

is Pareto optimal.

**Proof.** From Theorem 1 we know that the rate in Eq. (3) is achieved by Algorithm 1 with input \( \beta \). We only need to prove that no other algorithms achieve strictly smaller rates in pointwise order.

Suppose, by contradiction, we have \( \theta' \) achieved by an algorithm such that \( \theta'(\alpha) \leq \theta_\beta(\alpha) \) for all \( \alpha \in [0, 1] \) and \( \theta'(\alpha_0) < \theta(\alpha_0) \) for at least one \( \alpha_0 \in [0, 1] \). We then must have \( \theta'(0) \leq \theta_\beta(0) \).

We consider the following two exclusive cases.

**Case 1** \( \theta'(0) = \beta \). According to Lemma 5 we must have \( \theta' \geq \theta_\beta \), which leads to a contradiction.

**Case 2** \( \theta'(0) = \beta' < \beta \). According Lemma 5 we must have \( \theta' \geq \theta_\beta \). However, \( \theta_\beta \) is not strictly better than \( \theta_\beta \), e.g., \( \theta_\beta (2\beta - 1) = 2\beta - \beta' > \beta = \theta_\beta (2\beta - 1) \), which also leads to a contradiction.

### C Omitted proofs for Section 5

#### C.1 Proof of Lemma 1

**Lemma 1.** Suppose \( \alpha \) is the true hardness parameter and \( \alpha - 1/\lceil \log T \rceil < \alpha \leq \alpha_i \), run Algorithm 2 with time horizon \( T \) and \( \alpha_i \) leads to the following regret bound:

\[ \sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C \log T \cdot T^{(1+\alpha)/2}, \]

where \( C \) is a universal constant.

**Proof.** According to Lemma 2 the definition of \( \alpha \) and the assumption that \( \alpha \leq \alpha_i \), we know that \( \mathbb{P}(\mathcal{E}) \leq 1/\sqrt{T} \). We now upper bound the regret:

\[ R_T \leq \left( 39 \sqrt{|S_{\alpha_i}|T + |S_{\alpha_i}|} \right) \cdot \mathbb{P}(\neg \mathcal{E}) + T \cdot \mathbb{P}(\mathcal{E}) \]

\[ \leq \left( 39 \sqrt{|S_{\alpha_i}|T + |S_{\alpha_i}|} \right) \cdot 1 + T \cdot \frac{1}{\sqrt{T}} \]

\[ \leq 56 \log T \cdot T^{(1+\alpha_i)/2} + 2 \log T \cdot T^{\alpha_i} + \sqrt{T} \]

\[ \leq 59 \log T \cdot T^{(1+\alpha_i)/2} \]

\[ \leq 59 \log T \cdot T^{(1+\alpha_i)/2} \cdot T^{1/2 \lceil \log T \rceil} \]

\[ \leq 59 \sqrt{T} \log T \cdot T^{(1+\alpha_i)/2} \]

where Eq. (39) comes from the regret bound of MOSS; Eq. (40) comes from the assumption that \( \alpha_i < \alpha + 1/\lceil \log T \rceil \); and Eq. (41) comes from the fact that \( T^{1/2 \lceil \log T \rceil} = e^{(\log T) / (2 \lceil \log T \rceil)} \leq e^{\sqrt{T}} \). □

---

*One can sharpen the \( \log T \) term to \( \left( \log T \right)^{1/2} \) in many cases, e.g., when \( \alpha < 1 \) and \( T \) is large enough (with respect to \( \alpha \)). Again, we mainly focus on the polynomial terms here.*
C.2 Proof of Theorem 4

We first provide a martingale (difference) concentration result from [29] (a rewrite of Theorem 2.19).

**Lemma 6.** Let \( \{D_t\}_{t=1}^{\infty} \) be a martingale difference sequence adapted to filtration \( \{\mathcal{F}_t\}_{t=1}^{\infty} \). If \( \mathbb{E}[\exp(\lambda D_t)|\mathcal{F}_{t-1}] \leq \exp(\lambda^2\sigma^2/2) \) almost surely for any \( \lambda \in \mathbb{R} \), we then have

\[
P\left( \left| \sum_{i=1}^{t} D_i \right| \geq \epsilon \right) \leq 2 \exp\left(-\frac{\epsilon^2}{2t\sigma^2}\right).
\]

**Theorem 4.** For any \( \alpha \in [0, 1] \) unknown to the learner, run Parallel with time horizon \( T \) and optimal expected reward \( \mu^* \) leads to the following regret upper bound:

\[
\sup_{\omega \in \mathcal{H}_T(\alpha)} \hat{R}_T \leq C \left( \log T \right)^{2} T^{(1+\alpha)/2},
\]

where \( C \) is a universal constant.

**Proof.** This proof largely follows the proof of Theorem 4 in [23]. For any \( T \in \mathbb{N}_+ \) and \( i \in [[\log T]] \), recall \( S_{R_i} \) is the subroutine initialized with \( T \) and \( \alpha_i = i/[[\log T]] \). We use \( T_{i,t} \) to denote the number of samples allocated to \( S_{R_i} \) up to time \( t \), and represent its empirical regret at time \( t \) as \( \hat{R}_{i,t} = T_{i,t} \cdot \mu^* - \sum_{t'=1}^{t} X_{i,t'} \), where \( X_{i,t'} \sim \nu_{A_{i,t'}} \) is the \( t \)-th empirical reward obtained by \( S_{R_i} \) and \( A_{i,t} \) is the index of the \( t \)-th arm pulled by \( S_{R_i} \). We consider the corresponding regret \( R_{i,t} = T_{i,t} \cdot \mu^* - \sum_{t'=1}^{T_{i,t}} \mathbb{E}[\mu_{A_{i,t'}}] \) (which is random in \( T_{i,t} \)). We choose \( \delta = 1/\sqrt{T} \) as the confidence parameter and provide \( \delta = \delta/\log T \) failure probability to each subroutine.

Notice that \( R_{i,t} - \hat{R}_{i,t} = \sum_{t'=1}^{T_{i,t}} (X_{i,t'} - \mathbb{E}[\mu_{A_{i,t'}}]) \) is a martingale with respect to filtration \( \mathcal{F}_t = \sigma\left( \bigcup_{t \in [[\log T]]} \{T_{1,1}, A_{1,1}, X_{1,1}, \ldots, T_{i,t}, A_{i,t}, X_{i,t}, T_{i,t+1}, A_{i,t+1}, X_{i,t+1}, \ldots\} \right) \); and \( (R_{i,t} - \hat{R}_{i,t}) - (R_{i,t-1} - \hat{R}_{i,t-1}) \) defines a martingale difference sequence. Since, no matter what value \( T_{i,t} \) takes, \( X_{i,T_{i,t}} - \mathbb{E}[\mu_{A_{i,T_{i,t}}}] = (X_{i,T_{i,t}} - \mu_{A_{i,T_{i,t}}}) + (\mu_{A_{i,T_{i,t}}} - \mathbb{E}[\mu_{A_{i,T_{i,t}}}] \right) \) is \( (2/2) \)-sub-Gaussian (following a similar analysis as in Eq. (40)), applying Lemma 6 together with a union bound gives:

\[
P\left( \forall i \in [[\log T]], \forall t \in [T] : |\hat{R}_{i,t} - R_{i,t}| \geq \sqrt{T_{i,t} \cdot \log (2T/|\log T|)/\delta} \right) \leq \delta. \tag{42}
\]

We use \( \mathcal{E} = \left\{ \forall i \in [[\log T]], \forall t \in [T] : |\hat{R}_{i,t} - R_{i,t}| < \sqrt{T_{i,t} \cdot \log (2T/|\log T|)/\delta} \right\} \) to denote the good event that holds true with probability at least \( 1 - \delta \). Since the regret could be trivially upper bounded by \( T \cdot \delta = \sqrt{T} \) when \( \mathcal{E} \) doesn’t hold, we only focus on the case when event \( \mathcal{E} \) holds in the following.

Fix any subroutine \( k \in [[\log T]] \) and consider its empirical regret \( \hat{R}_{k,T} \) up to time \( T \). For any \( j \neq k \), let \( T_j \leq T \) be the last time that the subroutine \( S_{R_j} \) was invoked, we have

\[
\hat{R}_{j,T_j} \leq \hat{R}_{k,T_j} \\
\leq R_{k,T_j} + \sqrt{T_{k,T_j} \cdot \log (2T/|\log T|)/\delta} \\
\leq R_{k,T} + \sqrt{T \cdot \log (2T/|\log T|)/\delta}, \tag{43}
\]

where Eq. (43) comes from the fact that the cumulative regret \( R_{k,t} \) in non-decreasing in \( t \). Since \( S_{R_j} \) will only run additional \( \lceil \sqrt{T} \rceil \) rounds after it was selected at time \( T_j \), we further have

\[
\hat{R}_{j,T_j} \leq \hat{R}_{j,T_j} + \sqrt{T_j} \\
\leq R_{k,T} + \sqrt{5T \cdot \log (2T/|\log T|)/\delta}, \tag{44}
\]

where Eq. (44) comes from the combining Eq. (43) with a trivial bounding \( \lceil \sqrt{T} \rceil \leq \sqrt{4T} \) for all \( T \in \mathbb{N}_+ \). Combining Eq. (44) with the fact that \( R_{j,T_j} \leq \hat{R}_{j,T_j} + \sqrt{T \cdot \log (2T/|\log T|)/\delta} \) leads to

\[
R_{j,T} \leq R_{k,T} + 4\sqrt{T \cdot \log (2T/|\log T|)/\delta}. \tag{45}
\]
Let \( i_* \in \lceil \log T \rceil \) denote the index such that \( \alpha_{i_* - 1} < \alpha \leq \alpha_{i_*} \). As the total regret is the sum of all subroutines, we have that, for some universal constant \( C \),

\[
\sum_{i=1}^{\lceil \log T \rceil} R_{i,T} \leq \lceil \log T \rceil \cdot \left( R_{i_*,T} + 4 \sqrt{T \cdot \log (2T \lceil \log T \rceil / \delta)} \right) \leq \lceil \log T \rceil \cdot \left( 59 \sqrt{e \log T} \cdot T^{(1+\alpha)/2} + 4 \sqrt{T \cdot \log (2T^{3/2} \lceil \log T \rceil)} \right)
\]

\[
\leq C \cdot (\log T)^2 T^{(1+\alpha)/2},
\]

where Eq. (46) comes from setting \( k = i_* \) in Eq. (45); Eq. (47) comes from applying Lemma 1 with the non-decreasing nature of cumulative regret and taking \( \delta = 1/\sqrt{T} \). Integrate once more leads to the desired result.

\[\Box\]

### C.3 Anytime version

The anytime version of Algorithm 3 could be constructed as following.

**Algorithm 5: Anytime version of Parallel**

1: for \( i = 0, 1, \ldots \) do
2: Run Algorithm 3 with the optimal expected reward \( \mu_* \), for \( 2^i \) rounds.
3: end for

**Corollary 2.** For any time horizon \( T \) and \( \alpha \in [0, 1] \) unknown to the learner, run Algorithm 5 with optimal expected reward \( \mu_* \) leads to the following anytime regret upper:

\[
\sup_{\omega \in \mathcal{H}_T(\alpha)} R_T \leq C \cdot (\log T)^2 T^{(1+\alpha)/2},
\]

where \( C \) is a universal constant.

**Proof.** The proof is similar to the one for Corollary 1. \[\square\]