Refined Algorithms for Infinitely Many-Armed Bandits with Deterministic Rewards

Yahel David(✉) and Nahum Shimkin

Department of Electrical Engineering,
Technion—Israel Institute of Technology, 32000 Haifa, Israel
yahel83@gmail.com

Abstract. We consider a variant of the Multi-Armed Bandit problem which involves a large pool of a priori identical arms (or items). Each arm is associated with a deterministic value, which is sampled from a probability distribution with unknown maximal value, and is revealed once that arm is chosen. At each time instant the agent may choose a new arm (with unknown value), or a previously-chosen arm whose value is already revealed. The goal is to minimize the cumulative regret relative to the best arm in the pool. Previous work has established a lower bound on the regret for this model, depending on the functional form of the tail of the sample distribution, as well as algorithms that attain this bound up to logarithmic terms. Here, we present a more refined algorithm that attains the same order as the lower bound. We further consider several variants of the basic model, involving an anytime algorithm and the case of non-retainable arms. Numerical experiments demonstrate the superior performance of the suggested algorithms.

Keywords: Many-armed bandits · Regret minimization

1 Introduction

We consider a statistical learning problem in which the learning agent faces a large pool of possible items, or *arms*, each associated with a numerical *value* which is unknown a priori. At each time step the agent chooses an arm, whose exact value is then revealed and considered as the agent's reward at this time step. The goal of the learning agent is to maximize the cumulative reward, or, more specifically, to minimize the cumulative n-step regret (relative to the largest value available in the pool). At every time step, the agent should decide between sampling a new arm (with unknown value) from the pool, or sampling a previously sampled arm with a known value. Clearly, this decision represents the *exploration vs. exploitation* trade-off in the classic multi-armed bandit model. Our model assumes that the number of available arms in the pool is unlimited, and that the value of each newly observed arm is an independent sample from a common probability distribution. We study two variants of the basic model: the *retainable arms* case, in which the learning agent can return to any of the

A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 464–479, 2015.
DOI: 10.1007/978-3-319-23528-8_29

previously sampled arms (with known value), and the case of non-retainable arms, where previously sampled arms are lost if not immediately reused.

This model falls within the so-call infinitely-many armed framework, studied in [3,4,6,7,10,11]. In most of these works, which are further elaborated on below, the observed rewards are stochastic and the arms are retainable. Here, we continue the work in [7] that assumes that the potential reward of each arm is fixed and precisely observed once that arm is chosen. This simpler framework allows to obtain sharper bounds which focus on the basic issue of the sample size required to estimate the maximal value in the pool. At the same time, the assumption that the reward is deterministic may be relevant in various applications, such as parts inspection, worker selection, and communication channel selection. For this model, a lower bound on the regret and fixed time horizon algorithms that attain this lower bound up to logarithmic terms were presented in [7]. In the present paper, we propose algorithms that attain the same order as the lower bound (with no additional logarithmic terms) under a fairly general assumption on the tail of the probability distribution of the value. We further demonstrate that these bounds may not be achieved without this assumption. Furthermore, for the case where the time horizon is not specified, we provide an anytime algorithm that also attains the lower bound under similar conditions.

As mentioned above, several papers have studied a similar model with stochastic rewards. A lower bound on the regret was first provided in [3], for the case of Bernoulli arms, with the arm values (namely the expected rewards) distributed uniformly on the interval $[0, 1]$. For a known value distribution, algorithms that attain the same regret order as that lower bound are provided in [3,6,10], and an algorithm which attains that bound exactly under certain conditions is provided in [4]. In [11], the model was analyzed under weaker conditions that involve the form of the tail of the value distribution which is assumed known; however, significantly, the maximal value need not be known a priori. A lower bound and algorithms that achieve it up to logarithmic terms were developed for this case. The assumptions in the present paper are milder, in the sense that the tail distribution is not restricted in its form and only an upper bound on this tail is assumed rather than exact match. Our work also addresses the case of non-retainable arms, which has not been considered in the above-mentioned papers.

In a broader perspective, the present model may be compared to the continuum-armed bandit problem studied in [1,5,9]. In this model the arms are chosen from a continuous set, and the arm values satisfy some continuity properties over this set. In the model discussed here, we do not assume any regularity conditions across arms. The non-retainable arms version of our model is reminiscent of the classical *secretary problem*, see for example [8] and [2] for extensive surveys. In the secretary problem, the learning agent interviews job candidates sequentially, and wishes to maximize the probability of hiring the best candidate in the group. Our model considers the cumulative reward (or regret) as the performance measure.

The paper proceeds as follows. In the next section we present our model and the associated lower bound developed in [7]. Section 3 presents our algorithms and regret bounds for the basic model (with known time horizon and retainable arms). The extensions to anytime algorithms and the case of non-retainable arms are presented in Section 4. Some numerical experiments which compare the performance of the proposed algorithms to previous ones are described in Section 5, followed by concluding remarks.

2 Model Formulation and Lower Bound

We consider an infinite pool of arms, with values that are drawn independently from a common (but unknown) probability distribution with a cumulative distribution function $F(\mu)$, $\mu \in \mathbb{R}$. Let μ^* denote the supremal value, namely, the maximal value in the support of the measure defined by $F(\mu)$. As mentioned, once an arm is sampled its value is revealed, and at each time step $t = 1, ..., n$, a new or a previously sampled arm may be chosen. Our performance measure is the following cumulative regret.

Definition 1. *The regret at time step n is defined as:*

$$regret(n) = E\left[\sum_{t=1}^{n} (\mu^* - r(t))\right], \tag{1}$$

where $r(t)$ is the reward obtained at time t, namely, the value of the arm chosen at time t.

The following notations will be used in this paper:

- $\boldsymbol{\mu}$ is a generic random variable with distribution function F.
- For $0 \le \epsilon \le 1$, let

$$D_0(\epsilon) = \inf_{D \ge 0} \{P(\boldsymbol{\mu} \ge \mu^* - D) \ge \epsilon\},$$

  Note that $P(\boldsymbol{\mu} \ge \mu^* - D_0(\epsilon)) \ge \epsilon$, with equality if $\mu^* - D_0(\epsilon)$ is a continuity point of F. We refer to $D_0(\epsilon)$ as the *tail function* of F.
- Let $\epsilon_0^*(n)$ be defined as[1]

$$\epsilon_0^*(n) = \sup\left\{\epsilon \in [0,1] : nD_0(\epsilon) \le \frac{1}{\epsilon}\right\}. \tag{2}$$

  Note that $nD_0(\epsilon_1) \le \frac{1}{\epsilon_0^*(n)}$ for $\epsilon_1 \le \epsilon_0^*(n)$, and $nD_0(\epsilon_2) \ge \frac{1}{\epsilon_0^*(n)}$ for $\epsilon_2 > \epsilon_0^*(n)$.

For example, when $\boldsymbol{\mu}$ is uniform on $[a, b]$, then $D_0(\epsilon) = \frac{\epsilon}{b-a}$, and $\epsilon_0^*(n) = \sqrt{\frac{b-a}{n}}$.

[1] If the support of $\boldsymbol{\mu}$ is a single interval, then $D_0(\epsilon)$ is continuous. In that case, definition (2) reduced to the equation $nD_0(\epsilon) = \frac{1}{\epsilon}$ which, by monotonicity, has a unique solution for n large enough.

– Furthermore, let $D(\epsilon)$ denote a given upper bound on the tail function $D_0(\epsilon)$, and let $\epsilon^*(n)$ be defined similarly to $\epsilon_0^*(n)$ with $D_0(\epsilon)$ replaced by $D(\epsilon)$, namely,

$$\epsilon^*(n) = \sup\left\{\epsilon \in [0,1] : nD(\epsilon) \leq \frac{1}{\epsilon}\right\}. \tag{3}$$

Note that $\epsilon^*(n) \leq \epsilon_0^*(n)$. Since $D_0(\epsilon)$ is a non-decreasing function, we assume, without loss of generality, that $D(\epsilon)$ is also a non-decreasing function.

In the following sections, we shall assume that the upper bound $D(\epsilon)$ on the tail function $D_0(\epsilon)$ is known to the learning agent, and that it satisfies the following growth property.

Assumption 1

$$D(\epsilon) \leq MD(\epsilon_0)\alpha^{\epsilon/\epsilon_0}$$

for every $0 < \epsilon_0 \leq \epsilon \leq 1$ *and constants* $M > 1$ *and* $1 \leq \alpha < e$.

A general class of distributions that satisfies Assumption 1 is given in the following example, which will further serves us throughout the paper.

Example 1. Suppose that $P(\boldsymbol{\mu} \geq \mu^* - \epsilon) = \Theta\left(\epsilon^\beta\right)$ for $\epsilon > 0$ small enough, where $\beta > 0$. This is the case considered in [11]. Then $D_0(\epsilon) = \Theta\left(\epsilon^{1/\beta}\right)$, and for $D(\epsilon) = A\epsilon^{1/\beta}$, where $\beta > 0$ and $A > 0$, it can be obtained that $D(\epsilon) \leq MD(\epsilon_0)\alpha^{\epsilon/\epsilon_0}$, where $1 < \alpha < e$, $M = \frac{\lambda^{1/\beta}}{\alpha^\lambda}$ and $\lambda = \frac{1}{\beta \ln(\alpha)}$. Hence, in this case Assumption 1 holds. Note that $\beta = 1$ corresponds to a uniform probability distribution which is the case considered in [3] and [4] for $\mu^* = 1$.

Remark 1. Assumption 1 can be extended to any upper bound $\overline{\alpha}$ on the value of α (instead of e). In that case, a proper modification to the algorithms below leads to upper bounds that are larger by a constant multiplicative factor of $\ln(\overline{\alpha})$. However, as the assumption above covers most cases of interest, for simplicity of presentation, we will not go further into this extension. We note that the algorithms presented here do not use the values of α and M.

For the case in which the tail function $D_0(\epsilon)$ itself is known to the learning agent, the following lower bound on the expected regret was established in [7].

Theorem 1. *The n-step regret is lower bounded by*

$$regret(n) \geq (1 - \delta_n)\frac{\mu^* - E[\mu]}{16}\frac{1}{\epsilon_0^*(n)}, \tag{4}$$

where $\epsilon_0^*(n)$ *satisfies* (2)*, and* $\delta_n = 1 - 2\exp\left(-\frac{(\mu^* - E[\mu])^2}{8\epsilon_0^*(n)}\right)$.

Note that when $\epsilon_0^*(n) \to 0$ as $n \to \infty$, $\delta_n \to 0$ as $n \to \infty$, so that its effect becomes negligible. Furthermore, this lower bound coincides with the lower bounds presented in [3] and in [11] in the more specific models studied in those papers.

In the following corollary we present a lower bound on the regret for the case in which only a bound on the tail function $D_0(\epsilon)$ is known.

Corollary 1. *Let $D(\epsilon)$ be an upper bound on the tail function $D_0(\epsilon)$ such that*

$$\frac{D(\epsilon)}{D_0(\epsilon)} \leq L < \infty, \quad \forall\ 0 \leq \epsilon \leq 1.$$

Then, the n-step regret is lower bounded by

$$regret(n) \geq (1-\delta_n)\,\frac{\mu^* - E[\mu]}{16L}\,\frac{1}{\epsilon^*(n)}\,, \tag{5}$$

where $\epsilon^*(n)$ satisfies (3) and δ_n is as defined in Theorem 1.

Proof: Let

$$\epsilon_L^*(n) = \sup\left\{\epsilon \in [0,1] : n\frac{D(\epsilon)}{L} \leq \frac{1}{\epsilon}\right\}. \tag{6}$$

Then, for every $0 \leq \epsilon_1 \leq 1$ such that $\epsilon_L^*(n) < \epsilon_1$, by (6) and the assumed condition of the Corollary, it follows that $\frac{1}{\epsilon_1} < n\frac{D(\epsilon_1)}{L} \leq nD_0(\epsilon_1)$. Therefore, by Equation (2), $\epsilon_0^*(n) < \epsilon_1$. Thus,

$$\epsilon_0^*(n) \leq \epsilon_L^*(n). \tag{7}$$

Now, we need to compare $\epsilon_L^*(n)$ to $\epsilon^*(n)$. Let $L\epsilon^*(n) < \epsilon_2$. Since the tail function is non-decreasing, it follows that $\frac{L}{\epsilon_2} < nD(\frac{\epsilon_2}{L}) \leq nD(\epsilon_2)$, so that $\frac{1}{\epsilon_2} < n\frac{D(\epsilon_2)}{L}$. Hence, $\epsilon_L^*(n) < \epsilon_2$, and

$$\epsilon_L^*(n) \leq L\epsilon^*(n). \tag{8}$$

Equations (7) and (8) imply that $\epsilon_0^*(n) \leq L\epsilon^*(n)$, or $\frac{1}{L\epsilon^*(n)} \leq \frac{1}{\epsilon_0^*(n)}$. By substituting in Equation (4), the Corollary is obtained.

3 Optimal Sample Size

Here we discuss our most basic model, namely, the retainable arms model for a known time horizon. We present an algorithm that under Assumption 1 achieves a regret of the same order as the lower bound presented in Equation (5). We also present an example for which Assumption 1 does not hold, and show that for this example the lower bound on the regret is larger by a logarithmic factor than the lower bound presented in Equation (4).

The presented algorithm is simple and is based on initially sampling a certain number of new arms, followed by constantly choosing the single best arm found in the initial phase.

The following theorem provides an upper bound on the regret incurred by Algorithm 1.

Theorem 2. *Under Assumption 1, for every $n > 1$, the regret of Algorithm 1 is upper bounded by*

$$regret(n) \leq \left(1 + Me\frac{\alpha}{e-\alpha}\right)\frac{1}{\epsilon^*(n)} + 1, \tag{9}$$

where $\epsilon^*(n)$ is defined in Equation (3), and M and α are as defined in Assumption 1.

Algorithm 1. The Optimal Sampling Algorithm for Retainable Arms – OSR Algorithm

1: **Input:** $D(\epsilon)$, an upper bound on the tail function and time horizon $n > 1$.
2: Compute $\epsilon^*(n)$ as defined in (3).
3: Sample $N = \lfloor \frac{1}{\epsilon^*(n)} \rfloor + 1$ arms and keep the best one.
4: Continue by pulling the saved best arm up to the last stage n.

The upper bound obtained in the above Theorem is of the same order as the lower bound in Equation (5). Note that the values of M and α in Assumption 1 are not used in the algorithm, but only appear in the regret bound.

Example 1 (continued). For $\beta = 1$ ($\boldsymbol{\mu}$ is uniform on $[a, b]$), Assumption 1 holds for any $\alpha \in [1, e]$, with $M = \frac{\lambda^{1/\beta}}{\alpha^{\lambda}}$, where $\lambda = \frac{1}{\beta \ln(\alpha)}$ and $\frac{1}{\epsilon^*(n)} = \frac{\sqrt{n}}{\sqrt{b-a}}$. Therefore, for $\beta = 1$, with the optimize choice of $\alpha = 1.47$, we obtain $regret(n) < \frac{4.1\sqrt{n}}{\sqrt{b-a}} + 1$.

Proof of Theorem 2: For $N \geq 1$, we denote by $V_N(1)$ the value of the best arm found by sampling N different arms. Clearly,

$$regret(n) \leq N + (n - N)\Delta(N)\,, \tag{10}$$

where $\Delta(N) = E[\mu^* - V_N(1)]$. Then, for $N = \lfloor \frac{1}{\epsilon^*(n)} \rfloor + 1$, since $D(0) = 0$ we obtain

$$\Delta(N) \leq \Delta^{N,\epsilon}, \tag{11}$$

where

$$\Delta^{N,\epsilon} = \sum_{i=1}^{N} D(i\epsilon) P\left(D(i\epsilon) \geq \mu^* - V_N(1) > D((i-1)\epsilon)\right).$$

Note that if $N\epsilon > 1$ we take $D(N\epsilon) = D(1)$. So, Assumption 1 still holds. Also, for any $0 \leq \epsilon \leq 1$,

$$P\left(\mu^* - V_N(1) > D(\epsilon)\right) \leq (1 - \epsilon)^N.$$

Therefore,

$$\begin{aligned} \Delta^{N,\epsilon} &\leq \sum_{i=1}^{N} D(i\epsilon) P\left(\mu^* - V_N(1) > D((i-1)\epsilon)\right) \\ &\leq \sum_{i=1}^{N} D(i\epsilon)(1 - (i-1)\epsilon)^N \triangleq \overline{\Delta}^{N,\epsilon}. \end{aligned} \tag{12}$$

Observe that $(1 - \epsilon)^{\frac{1}{\epsilon}} \leq e^{-1}$ for $\epsilon \in (0, 1]$. Then, for $\epsilon = \epsilon^*(n)$, since $N \geq \frac{1}{\epsilon^*(n)}$ it follows that $(1 - (i-1)\epsilon^*(n))^N \leq e^{1-i}$. Hence,

$$\begin{aligned} \overline{\Delta}^{N,\epsilon^*(n)} &= \sum_{i=1}^{N} D(i\epsilon^*(n))(1 - (i-1)\epsilon^*(n))^N \\ &\leq \sum_{i=1}^{N} D(i\epsilon^*(n)) e^{1-i} \triangleq \overline{\Delta}_0^{N,\epsilon^*(n)}\,. \end{aligned} \tag{13}$$

Now, by Assumption 1,

$$\overline{\Delta}_0^{N,\epsilon^*(n)} \leq \sum_{i=1}^{N} MD(\epsilon^*(n))\alpha^i e^{1-i} < Me\frac{\alpha}{e-\alpha}D(\epsilon^*(n)). \tag{14}$$

Therefore, by (10),

$$regret(n) \leq \lfloor \frac{1}{\epsilon^*(n)} \rfloor + 1 + nMe\frac{\alpha}{e-\alpha}D(\epsilon^*(n)) \leq (1 + Me\frac{\alpha}{e-\alpha})\frac{1}{\epsilon^*(n)} + 1\,.$$

Hence, the upper bound on (9) is obtained.

□

For the case that Assumption 1 does not hold, we provide an example for which the regret is larger than the lower bound presented in Equation (4) by a logarithmic term.

Example 2. Suppose that $P(\boldsymbol{\mu} \geq \mu^* - \epsilon) = -\frac{1}{\ln(\epsilon)}$. Then $D_0(\epsilon) = e^{-\frac{1}{\epsilon}}$, and it follows that $\frac{1}{\ln(n)} \leq \epsilon_0^*(n) \leq \frac{2}{\ln(n)}$.
Take $\epsilon_0 = \frac{1}{2}\epsilon$. Then, for any $\alpha > 1$ and $M > 0$, for ϵ small enough we obtain $\frac{D(\epsilon)}{D(\epsilon_0)} = e^{1/\epsilon_0 - 1/\epsilon} = e^{1/\epsilon} > M\alpha^2 = M\alpha^{\epsilon/\epsilon_0}$. Hence, Assumption 1 does not hold.

Lemma 1. *For the case considered in Example 2, the best regret which can be achieved is larger by multiplicative a logarithmic factor* ($\ln(n)$) *than the lower bound presented in Equation* (4).

Proof: Let N stand for the number of sampled arms, then, one can find that

$$regret(n) = NE[\boldsymbol{\mu}] + (n - N)\Delta(N), \tag{15}$$

where $\Delta(N) = E[\mu^* - V_N(1)]$. To bound the second term of Equation (15), note that, for any $\overline{N} \leq \lfloor \frac{1}{\epsilon} \rfloor$,

$$\begin{aligned}\Delta(N) &\geq \sum_{i=1}^{\overline{N}} D_0(i\epsilon)P\left(D_0((i+1)\epsilon) \geq \mu^* - V_N(1) > D_0((i)\epsilon)\right) \\ &= \sum_{i=1}^{\overline{N}} D_0(i\epsilon)(\Delta^{N,\epsilon}(i) - \Delta^{N,\epsilon}(i+1)) \triangleq \tilde{\Delta}(N),\end{aligned}$$

where

$$\Delta^{N,\epsilon}(i) = P\left(\mu^* - V_N(1) > D_0(i\epsilon)\right).$$

By the fact that $D_0(\epsilon)$ is continuous, it follows that

$$\Delta^{N,\epsilon}(i) = P\left(\mu^* - V_N(1) > D_0(i\epsilon)\right) = (1 - i\epsilon)^N,$$

and

$$\Delta^{N,\epsilon}(i+1) = P\left(\mu^* - V_N(1) > D_0((i+1)\epsilon)\right) = (1 - (i+1)\epsilon)^N.$$

Noting that $e^{-1} \geq (1-\epsilon)^{\frac{1}{\epsilon}} \geq \exp\left(-1-\frac{\epsilon}{1-\epsilon}\right)$ for $\epsilon \in (0,1]$, we obtain for the choice of $\epsilon = \frac{1}{N}$ that

$$\Delta^{N,\frac{1}{N}}(i) - \Delta^{N,\frac{1}{N}}(i+1) \geq e^{-i}\beta_N^i \ ,$$

where $\beta_N^i = e^{\frac{-i}{N-1}} - e^{-1}$.

Now, since $D_0(i\epsilon) = D_0(\epsilon)e^{\frac{i-1}{i\epsilon}}$, again for the choice of $\epsilon = \frac{1}{N}$, it follows that

$$\tilde{\Delta}(N) \geq \sum_{i=1}^{\overline{N}} D_0(\frac{1}{N})e^{N-\frac{N}{i}}e^{-i}\beta_N^i \geq \lfloor\sqrt{N}\rfloor D_0(\frac{1}{N})e^{N-2\sqrt{N}}\beta_N^{\sqrt{N}}.$$

Therefore, since $D_0(\frac{1}{N}) = e^{-\frac{1}{N}}$, for $N \geq 3$ we obtain that

$$regret(n) = NE[\boldsymbol{\mu}] + (n-N)\lfloor\sqrt{N}\rfloor e^{-2\sqrt{N}}\beta_N^{\sqrt{N}},$$

For $N < 3$, noting that $\Delta(N)$ is a non-increasing function of N, we have $\Delta(1), \Delta(2) \geq \Delta(3)$, hence

$$regret(n) = NE[\boldsymbol{\mu}] + (n-2)\lfloor\sqrt{3}\rfloor e^{-2\sqrt{3}}\beta_3^{\sqrt{3}}.$$

By optimizing over N, it can be found that

$$regret(n) \geq A\ln^2(n)$$

where $A = \frac{E[\boldsymbol{\mu}]}{5}$. But, since $\epsilon_0^*(n) \leq \frac{2}{\ln(n)}$, the order of the regret is larger by a logarithmic factor than the lower bound on the regret of Equation (4).

□

4 Extensions

In this section we discuss two extensions of the basic model, the first is the case in which the time horizon is not specified, leading to an anytime algorithm, and the second is the non-reatainable arms model.

4.1 Anytime Algorithm

Consider again the retainable arms model, but assuming now that the time horizon is unspecified. Under Assumption 1 and a mild condition on the tail of the value probability distribution, the proposed algorithm achieves a regret of the same order as the lower bound of Equation (5).

The presented algorithm is a natural extension of Algorithm 1. Here, instead of sampling a certain number of arms at the first phase (as a function of the time horizon) and then sampling the best one among them at the second phase, the algorithm insures that at every time step, the number of sampled arms is larger than a threshold which is a function of time. Since the number of sampled

Algorithm 2. The Anytime Optimal Sampling Algorithm for Retainable Arms – AT-OSR Algorithm

Input: $D(\epsilon)$, an upper bound on the tail function.
Initialization: $m = 0$ the number of sampled arms.
Compute $\epsilon^*(t)$ as defined in (3).
if $m < \lfloor \frac{1}{\epsilon^*(t)} \rfloor + 1$ **then**
    Sample a new arm, update $t = t + 1$ and return to step 3.
else
    Pull the best arm so far, update $t = t + 1$ and return to step 3.
end if

arms is increasing gradually, the upper bound on the regret obtained here is worse than that obtained in the case of known time horizon. However, we show in Corollary 2 that it is of the same order, under an additional condition.

We note that applying the standard doubling trick to Algorithm 1 does not serve our purpose here, as it would add a logarithmic factor to the regret bound.

In the following Theorem we provide an upper bound on the regret achieved by the proposed Algorithm.

Theorem 3. *Under Assumption 1, for every $n > 1$, the regret of Algorithm 2 is upper bounded by*

$$regret(n) \leq Me\frac{\alpha}{e-\alpha}\sum_{t=2}^{n}\frac{1}{t\epsilon^*(t)} + \frac{1}{\epsilon^*(n)} + 2\ , \tag{16}$$

where $\epsilon^*(n)$ is defined in (3), and M and α are as defined in Assumption 1.

As $\frac{1}{\epsilon^*(n)} \geq \frac{1}{\epsilon^*(t)}$ for $t \leq n$, it is obtained that in the worst case, the bound in Equation (16) is larger than the lower bound in Equation (5) by a logarithmic term. However, as shown in the following corollary, under reasonable conditions on the tail function $D(\epsilon)$, the bound in Equation (16) is of the same order as the lower bound in Equation (5).

Corollary 2. *If $B_1 t^\gamma \leq D(\epsilon) \leq B_2 t^\gamma$ for some constants $0 < B_1 \leq B_2$ and $0 < \gamma$, then*

$$regret(n) \leq \left(2Me\frac{\alpha}{e-\alpha}\left(\frac{B_2}{B_1}\right)^{\frac{1}{1+\gamma}}(1+\gamma)\,f(n) + 1\right)\frac{1}{\epsilon^*(n)} + 2 \tag{17}$$

where $f(n) = \left(\frac{n+1}{n}\right)^{\frac{1}{1+\gamma}}$; note that $f(n) \to 1$ asymptotically.

Example 1 (continued). When $D(\epsilon) = \Theta\left(\epsilon^{1/\beta}\right)$, it follows that $\frac{1}{\epsilon^*(t)} = \Theta\left(n^{\frac{\beta}{1+\beta}}\right)$. Therefore, the condition of Corollary 2 holds.

Proof of Corollary 2: Under the assumed condition $B_1' t^{\gamma'} \leq \frac{1}{\epsilon^*(t)} \leq B_2' t^{\gamma'}$, where $B_1' = B_1^{\frac{1}{1+\gamma}}$, $B_2' = B_2^{\frac{1}{1+\gamma}}$ and $\gamma' = \frac{1}{1+\gamma}$. Therefore,

$$\sum_{t=2}^{n} \frac{1}{t\epsilon^*(t)} \leq \int_{t=2}^{n+1} \frac{1}{(t-1)\epsilon^*(t)} \leq \int_{t=2}^{n+1} \frac{2}{t\epsilon^*(t)} \leq \frac{2B_2'}{\gamma'}(n+1)^{\gamma'} \leq \frac{2B_2' f(n)}{B_1'\gamma'} \frac{1}{\epsilon^*(n)} .$$

Therefore, by Equation (16), Equation (17) is obtained.

□

Proof of Theorem 3: Recall the notation $V_N(1)$ for the value of the best arm found by sampling N different arms. We bound the regret by

$$regret(n) \leq E\left[1 + \sum_{t=2}^{n} I\left(E_t\right) + \left(\mu^* - V_t(1)\right) I\left(\overline{E_t}\right)\right] ,$$

where $E_t = \left\{m_t < \lfloor \frac{1}{\epsilon^*(t)} \rfloor + 1\right\}$, and $I(\cdot)$ is the indicator function.

Since $\lfloor \frac{1}{\epsilon^*(t)} \rfloor + 1$ is a monotone increasing function it follows that

$$1 + \sum_{t=2}^{n} I\left(E_t\right) \leq \lfloor \frac{1}{\epsilon^*(n)} \rfloor + 2 .$$

Recall that $\Delta(t) = E[\mu^* - V_t(1)]$, then, since $V_t(1)$ is non-decreasing, by Equations (11)-(14), we obtain that

$$E\left[\sum_{t=2}^{n} \left(\mu^* - V_t(1)\right) I\left(\overline{E_t}\right)\right] \leq \sum_{t=2}^{n} \Delta\left(\lfloor \frac{1}{\epsilon^*(t)} \rfloor + 1\right) \leq \sum_{t=2}^{n} Me \frac{\alpha}{e-\alpha} D(\epsilon^*(t)) .$$

Also, by Equation (3),

$$\sum_{t=2}^{n} D(\epsilon^*(t)) \leq \sum_{t=2}^{n} \frac{1}{t\epsilon^*(t)} .$$

Combining the above yields the bound in Equation (16). □

4.2 Non-retainable Arms

Here the learning agent is not allowed to reuse a previously sampled arm, unless this arm was sampled in the last time step. So, Algorithm 1 cannot be applied to this model. However, the values of previously chosen arms can provide a useful information for the agent. In this section we present an algorithm that achieves a regret that is larger by a sublogarithmic term than the lower bound in Equation (5). The following additional assumption will be required here.

Assumption 2

$$D(\epsilon) \leq C\left(\frac{\epsilon}{\epsilon_0}\right)^{\tau} D(\epsilon_0)$$

for every $0 < \epsilon_0 \leq \epsilon \leq 1$ *and constants* $C > 0$ *and* $\tau > 0$.

Algorithm 3. The Optimal Sampling Algorithm for Non-Retainable Arms – OSN Algorithm

1: **Input:** An upper bound $D(\epsilon)$ on the tail function, time horizon $n > 1$, and τ under which Assumption 2 holds.
2: Compute $\epsilon^*(n)$ as defined in (3).
3: Sample $N = \lfloor \ln^{-\frac{1}{1+\tau}}(n) \frac{1}{\epsilon^*(n)} \rfloor + 1$ arms and keep the value of the best one.
4: Continue by pulling until observing a value equal or greater than the saved best value. Then, continue by pulling this arm up to the last stage n.

Assumption 2 holds, in particular, in Example 1 for $\tau = \frac{1}{\beta}$.

The proposed algorithm is based on sampling a certain number of arms, such that the value of the best one among them is on one hand large enough, and on the other hand the probability of finding another arm with a larger value is also high enough. Thus, after sampling that number of arms, the algorithm continues by pulling new arms until it finds one with a larger (or equal) value than all previously sampled arms.

In the following theorem, we provide an upper bound on the regret achieved by the presented algorithm.

Theorem 4. *Under Assumptions 1 and 2, for every $n > 1$, the regret of Algorithm 3 is upper bounded by*

$$regret(n) \leq \left(1 + \ln^{\frac{\tau}{1+\tau}}(n) \left(CMe\frac{\alpha}{e-\alpha} + \frac{2}{\ln(2)}\right)\right) \frac{1}{\epsilon^*(n)} + \frac{2\ln^{\frac{\tau}{1+\tau}}(n)}{\ln(2)} + 1 \quad (18)$$

where $\epsilon^*(n)$ is defined in Equation (3), M and α are as defined in Assumption 1 and C and τ are as defined in Assumption 2.

Proof: For $N \geq 1$, recall that $V_N(1)$ stands for the value of the best arm found by sampling N different arms and that $\Delta(N) = E[\mu^* - V_N(1)]$. Clearly,

$$regret(n) \leq N + (n-N)\Delta(N) + E[Y(V_N(1))], \quad (19)$$

where the random variable $Y(V)$ is the number of arms sampled until an arm with a value larger or equal to V is sampled. The second term in Equation (19) can be bounded similarly to the second term in Equation (10). Namely, since $N \geq \ln^{-\frac{1}{1+\tau}}(n) \frac{1}{\epsilon^*(n)}$,

$$\overline{\Delta}^{N,\ln^{\frac{1}{1+\tau}}(n)\epsilon^*(n)} \leq \overline{\Delta}_0^{N,\ln^{\frac{1}{1+\tau}}(n)\epsilon^*(n)},$$

and then, by Assumption 2,

$$\overline{\Delta}_0^{N,\ln^{\frac{1}{1+\tau}}(n)\epsilon^*(n)} \leq \sum_{i=1}^{N} MD(\ln^{\frac{1}{1+\tau}}(n)\epsilon^*(n))\alpha^i e^{1-i} < C\ln^{\frac{\tau}{1+\tau}}(n) Me\frac{\alpha}{e-\alpha} D(\epsilon^*(n)).$$

Thus, as shown in the proof of Theorem 2,

$$(n-N)\Delta(N) < nC\ln^{\frac{\tau}{1+\tau}}(n)Me\frac{\alpha}{e-\alpha}D(\epsilon^*(n)) \leq C\ln^{\frac{\tau}{1+\tau}}(n)Me\frac{\alpha}{e-\alpha}\frac{1}{\epsilon^*(n)}. \tag{20}$$

For bounding the third term, let

$$\hat{\epsilon}_\gamma = \sup\{\epsilon \in [0,1] | D(\epsilon) \leq \gamma\} \ ,$$

and note that

$$P\left(\boldsymbol{\mu} \geq \mu^* - \gamma\right) = \hat{\epsilon}_\gamma \ . \tag{21}$$

Now, let us define:

$$\epsilon_1 = \hat{\epsilon}_{\gamma_1}, \ \gamma_1 = D(\frac{1}{n}) \ ,$$

as well as the following sequence:

$$\epsilon_{i+1} = \hat{\epsilon}_{\gamma_{i+1}}, \text{ for } \gamma_{i+1} = D(2\epsilon_i), \ \forall i \geq 1 \ .$$

Let M be such that ϵ_M is the first element in the sequence which is larger or equal to one, and set $\epsilon_M = 1$. Then, since $D(\epsilon_{i+1}) = D(2\epsilon_i) = \gamma_{i+1}$ $\forall i \geq 1$, and $E[Y(V)]$ is non-decreasing in V, we obtain that

$$\begin{aligned} E\left[Y(V_N(1))\right] &= E\left[E\left[Y(V_N(1))\right]|V_N(1)\right] \\ &\leq \sum_{i=1}^{M} E\left[Y(\mu^* - \gamma_i)\right] P\left(\mu^* - \gamma_i \geq V_N(1) > \mu^* - \gamma_{i+1}\right) \\ &\leq \sum_{i=1}^{M} E\left[Y(\mu^* - \gamma_i)\right] P\left(V_N(1) > \mu^* - \gamma_{i+1}\right) \triangleq \Phi_N \ . \end{aligned} \tag{22}$$

Then, by the expected value of a Geometric distribution, Equation (21), and the fact that $\gamma_i = D(\epsilon_i)$, we obtain that

$$E\left[Y(\mu^* - \gamma_i)\right] = \frac{1}{\epsilon_i} \ .$$

Also, since $\gamma_{i+1} = D(2\epsilon_i)$, it follows that

$$P\left(V_N(1) > \mu^* - \gamma_{i+1}\right) \leq 2N\epsilon_i \ .$$

So, since $M \leq \frac{\ln(n)}{\ln(2)}$, we have

$$\Phi_N \leq \sum_{i=1}^{M} 2N \leq 2N\frac{\ln(n)}{\ln(2)} \leq \frac{2\ln^{\frac{\tau}{1+\tau}}(n)}{\ln(2)}\left(\frac{1}{\epsilon^*(n)} + 1\right). \tag{23}$$

By combining Equations (19), (20), (22) and (23) the claimed bound in Equation (18) is obtained.

□

We note that a combined model which considers the anytime problem for the non-retainable arms case can be analyzed by similar methods. However, we do not consider this variant here.

5 Experiments

We next investigate numerically the algorithms presented in this paper, and compare them to the relevant algorithms from [7,11]. We remind that the present deterministic model was only studied in [7], while the model considered in [11] is similar in its assumptions to the presented one in that only the form of the tail function (rather the exact value distribution) is assumed known. Since the algorithms in [11] are analyzed only for the case of Example 1 (i.e. $D(\epsilon) = \Theta\left(\epsilon^{\beta}\right)$), we adhere to this model with several values of β for our experiments. The maximal value is taken $\mu^* = 0.99$, but is not known to the learning agent. In addition to that, since the algorithms presented in [11] were planned for the stochastic model, they apply the UCB-V policy on the sampled set of arms. Here, we eliminate this stage which is evidently redundant for the deterministic model considered here.

5.1 Retainable Arms

For the case of retainable arms and a known time horizon, we compare Algorithm 1 with the *KT&RA* Algorithm presented in [7] and the *UCB-F* Algorithm presented in [11]. Since in [11], just an order of the number of arms needed to be sampled is specified (and not exact number), we consider two variations of the *UCB-F* Algorithm, one with a multiplicative factor of 10, and the other with a multiplicative factor of 0.2.

Table 1. Average regret for the retainable arms model for the known time horizon case.

	Time Horizon								
	$\beta = 0.9$			$\beta = 1$			$\beta = 1.1$		
Algorithm	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4
UCB-F-10	574	740	870	1022	1350	1612	1376	1847	2227
UCB-F-0.2	1043	1410	1778	1043	1410	1778	1129	1445	1764
KT&RA	423	578	705	568	787	970	738	1035	1276
Algorithm 1	242	307	360	287	388	460	381	515	626

In Figure 1, we present the average regret of 200 runs vs. the time horizon for $\beta = 0.9$, $\beta = 1$ and $\beta = 1.1$. The empirical standard deviation is smaller than 5% in all of our results. As shown in Figure 1 and detailed in Table 1, the performance of Algorithm 1 is significantly better than the other algorithms.

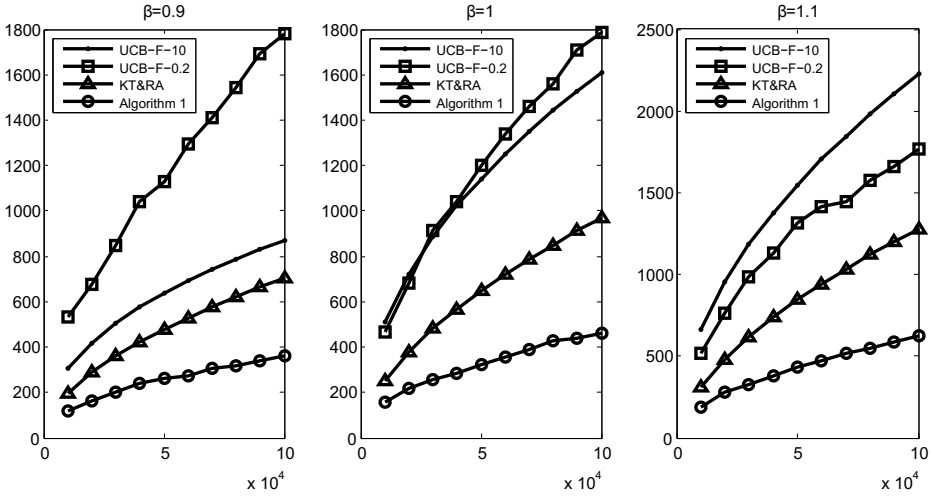

Fig. 1. Average regret (y-axis) vs. the time horizon (x-axis) for $\beta = 0.9$, $\beta = 1$ and $\beta = 1.1$.

5.2 Anytime Algorithm

For the retainable arms model and unspecified time horizon, we compare Algorithm 2 with the *UCB-AIR* Algorithm presented in [11]. Since, these algorithms are identical for $\beta \geq 1$, we run this experiment for $\beta = 0.7$, $\beta = 0.8$ and $\beta = 0.9$. In Figure 2 we present the average regret of 200 runs vs. the time. It is shown in Figure 2 and detailed in Table 2 that the average regret of Algorithm 2 is

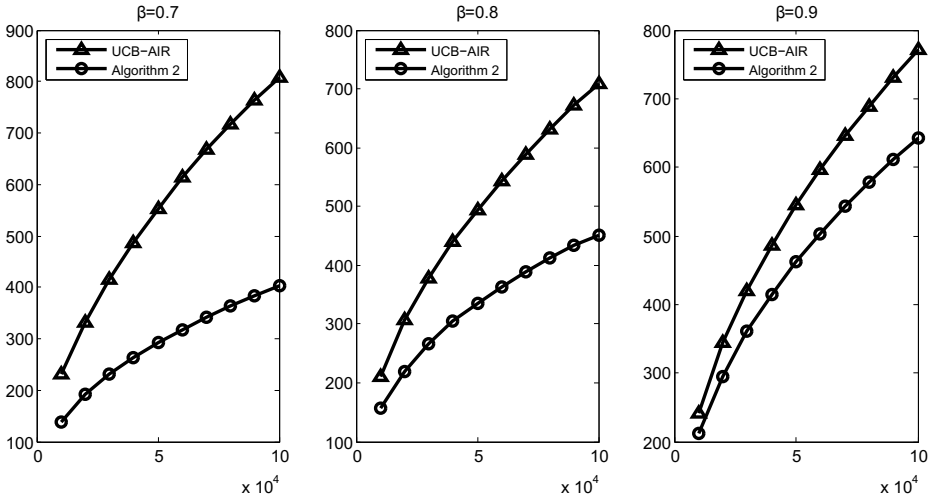

Fig. 2. Average regret (y-axis) vs. time (x-axis) for $\beta = 0.7$, $\beta = 0.8$ and $\beta = 0.9$.

Table 2. Average regret for the retainable arms model for the unknown time horizon case.

	Time Horizon								
	$\beta = 0.7$			$\beta = 0.8$			$\beta = 0.9$		
Algorithm	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4
UCB-AIR	414	667	808	440	589	710	486	646	771
Algorithm 2	264	341	402	305	389	414	542	642	1764

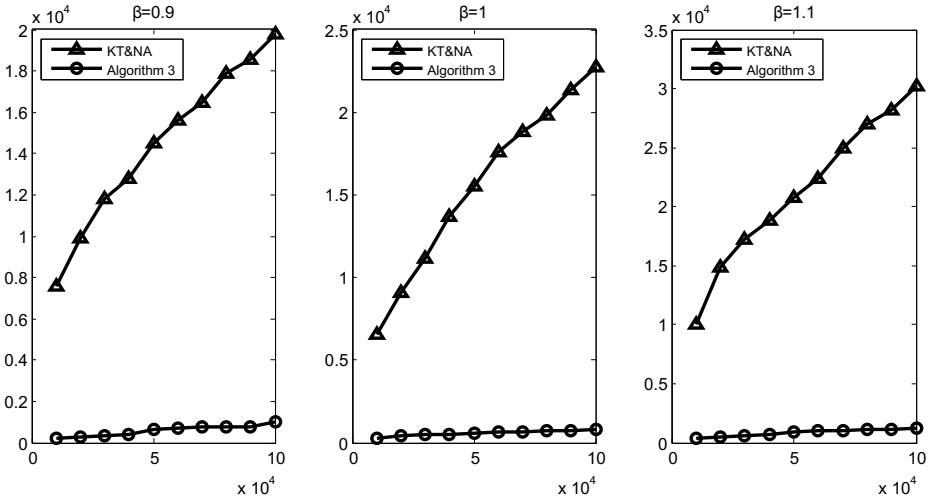

Fig. 3. Average regret (y-axis) vs. the time horizon (x-axis) for $\beta = 0.9$, $\beta = 1$ and $\beta = 1.1$.

smaller and increasing slower than that of the *UCB-AIR* Algorithm. Here the empirical standard deviation is smaller than 7% in all of our results.

Table 3. Average regret for the non-retainable arms model.

	Time Horizon								
	$\beta = 0.7$			$\beta = 0.8$			$\beta = 0.9$		
Algorithm	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4	4×10^4	7×10^4	10×10^4
KT&NA	12800	16460	19760	13670	18800	22760	18850	24950	30170
Algorithm 3	418	741	983	509	646	791	674	1048	1277

5.3 Non-Retainable Arms

In the case of non-retainable arms and a fixed time horizon, we compare Algorithm 3 with the *KT&NA* Algorithm presented in [7]. As in the previous case, we present in Figure 3 the average regret of 200 runs vs. the time horizon for $\beta = 0.9$, $\beta = 1$ and $\beta = 1.1$. Here, the empirical standard deviation is smaller than 10% in all of our results. As shown in Figure 3 and detailed in Table 3, Algorithm 3 outperforms the *KT&NA* Algorithm.

6 Conclusion and Discussion

In this work we provided algorithms with tight bounds on the cumulative regret for the infinitely-many armed problem with deterministic rewards. Our central assumption is that the tail function $D_0(\epsilon)$ is known, to within multiplicative constant. The basic algorithm was extended to the any-time case and to the model with non retainable arms.

A major challenge for future work is further relaxation of the requirement of a known upper bound on the tail function $D_0(\epsilon)$. Initial steps in this direction were presented in [7].

References

1. Auer, P., Ortner, R., Szepesvári, C.: Improved rates for the stochastic continuum-armed bandit problem. In: Bshouty, N.H., Gentile, C. (eds.) COLT. LNCS (LNAI), vol. 4539, pp. 454–468. Springer, Heidelberg (2007)
2. Babaioff, M., Immorlica, N., Kempe, D., Kleinberg, R.: Online auctions and generalized secretary problems. ACM SIGecom Exchanges **7**(2), 1–11 (2008)
3. Berry, D.A., Chen, R.W., Zame, A., Heath, D.C., Shepp, L.A.: Bandit problems with infinitely many arms. The Annals of Statistics, 2103–2116 (1997)
4. Bonald, T., Proutiere, A.: Two-target algorithms for infinite-armed bandits with bernoulli rewards. In: Advances in Neural Information Processing Systems, pp. 2184–2192 (2013)
5. Bubeck, S., Munos, R., Stoltz, G., Szepesvári, C.: X-armed bandits. Journal of Machine Learning Research **12**, 1655–1695 (2011)
6. Chakrabarti, D., Kumar, R., Radlinski, F., Upfal, E.: Mortal multi-armed bandits. In: Advances in Neural Information Processing Systems, pp. 273–280 (2009)
7. David, Y., Shimkin, N.: Infinitely many-armed bandits with unknown value distribution. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014, Part I. LNCS, vol. 8724, pp. 307–322. Springer, Heidelberg (2014)
8. Freeman, P.: The secretary problem and its extensions: A review. International Statistical Review, 189–206 (1983)
9. Kleinberg, R., Slivkins, A., Upfal, E.: Multi-armed bandits in metric spaces. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, pp. 681–690. ACM (2008)
10. Teytaud, O., Gelly, S., Sebag, M.: Anytime many-armed bandits. In: CAP (2007)
11. Wang, Y., Audibert, J.-Y., Munos, R.: Algorithms for infinitely many-armed bandits. In: Advances in Neural Information Processing Systems, pp. 1729–1736 (2009)