# Infinitely Many-Armed Bandits with Unknown Value Distribution

Yahel David and Nahum Shimkin

Department of Electrical Engineering,
Technion—Israel Institute of Technology,
Haifa 32000, Israel

**Abstract.** We consider a version of the classical stochastic Multi-Armed bandit problem in which the number of arms is large compared to the time horizon, with the goal of minimizing the cumulative regret. Here, the mean-reward (or *value*) of newly chosen arms is assumed to be i.i.d. We further make the simplifying assumption that the value of an arm is revealed once this arm is chosen. We present a general lower bound on the regret, and learning algorithms that achieve this bound up to a logarithmic factor. Contrary to previous work, we do not assume that the functional form of the tail of the value distribution is known. Furthermore, we also consider a variant of our model where sampled arms are non-retainable, namely are lost if not used continuously, with similar near-optimality results.

## 1 Introduction

We consider a statistical learning problem where a learning agent is facing a large pool of possible choices, or *arms*, each associated with a distinct numeric *value* which equals the one-stage reward that is obtained by choosing that arm. The goal is to minimize the cumulative $n$-step regret (relative to the best available arm). The agent has no prior knowledge on the value of unobserved arms, and assumes that the value of each newly observed arm is sampled independently from a common probability distribution. Once an arm is chosen its value is revealed, and the agent may continue to pick a new arm, or return to a previously chosen one. Clearly, this choice represents the essence of the *exploration vs. exploitation* dilemma for this model.

It is assumed that the pool of arms is large enough compared to the time horizon $n$, so that the agent cannot (or does not find it efficient) to sample them all, hence this pool can be effectively viewed as infinite. A similar model has been considered in [4,5,7,14,15]. In these papers, the observed reward of a given arm is assumed to be stochastic. In contrast, we consider here the simpler case where the reward of each arm is deterministic, so that a single observation is enough to evaluate it precisely[1]. This focuses the problem strictly on the issue of

[1] More generally, we may assume that the obtained reward is stochastic, but its mean is revealed once an arm is chosen. This does not affect our results as we consider the expected regret.

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 307–322, 2014.

obtaining new samples, rather than learning the expected value of ones already sampled. On the other hand, the present paper generalizes the models studied in these papers in the following two respects.

- **Prior knowledge:** No prior knowledge is assumed regarding the functional form of the value distribution. Thus, the required sample size need to be estimated from the observed samples.
- **Non-retainable arms:** In addition to the basic model that allows retaining previously observed arms for further use, we consider the case where previously sampled arms are lost if not used again immediately. Discarded arms cannot be used again, but their observed values are useful for learning purposes.

Relaxing the prior knowledge assumption is natural when facing an unknown population for the first time. The non-retainable arms model is motivated by applications where arms are associated with volatile resources such as job offers and positions, apartment rental, business contracts, established routs in an ad-hoc network, and so on. To elaborate on a particular example, consider the problem of video streaming of a movie file to a media client over a wireless channel. After the transmission of each segment of the movie, the provider obtains feedback on the quality of the used channel, and decides whether to use this channel again or try a new one. If a channel is dropped it may be used by another user and hence lost. This scenario may be captured in our model by associating channels with arms, and the perceived channel quality with the obtained rewards.

As mentioned, the infinitely-many arms model has been considered before in [4], [5], [7] [14] and [15]. In [4], the rewards of each arm are assumed to be Bernoulli distributed, while the mean rewards (or *values*) of the different arms are taken to be uniformly distributed. This paper presents algorithms for a fixed horizon $n$ which achieve a cumulative regret of an order of $\sqrt{n}$ for a fixed horizon $n$, and establishes a lower bound of the same order. Later, in [14] and [5], anytime algorithms were presented for similar reward and value distributions, where [5] also provides a fixed horizon time algorithm which achieves the optimal regret. A more general model was considered in [15], where arm value distribution (or at least its upper tail) is assumed known and to belong to a certain one parameter family. This paper provides a lower bound on the regret, that depends on this parameter, and proposes fixed horizon and anytime algorithms that approaches this bound up to logarithmic factors in $n$. Motivated by e-commerce applications, a deterministic reward model, similar to ours, is considered in [7]. That paper presents an algorithm which attains the optimal regret bound, under the assumption of known value distribution.

In a broader context, our model may be compared to the continuous multi-armed bandit problem discussed in [11], [2] and [6]. In this model the arm is chosen from a continuous set, and continuity conditions are assumed on the arm values. In contrast, in the model of the present paper no regularity or dependence assumptions are made on the arms; for further discussion and comparison of the two models see [15]. Another similar model is the contextual Multi-armed Bandit with an infinite number of arms or context sets, which is discussed in [12], [13], [10] and [1]. Again, in this model a continuity or another similarity condition is assumed on the arm values.

The non-retainable arm assumption (along with the deterministic reward property) are reminiscent of the celebrated *Secretary problem* of optimal stopping theory. In its basic form, a known number of candidates arrive sequentially for an interview, which reveals their relative merit. The interviewer should decide after each interview whether to stop and hire the last interviewed candidate, with the goal of maximizing the probability of hiring the best one. This problem has been extensively studied and extended, for example see [9] and [3]. An essential difference in our problem is the use of the regret as the performance criterion.

In this paper we present several classes of adaptive sampling algorithms for the infinitely many armed bandit problem. The algorithms are developed gradually, starting with the simpler case of a known tail distribution and generalizing to the unknown distribution case. The presentation proceeds as follows. After presenting the model in Sect. 2, we formulating in Sect. 3 a lower bound that applies to all the cases considered. All our proposed algorithms will be shown to achieve this lower bound up to a logarithmic factor. In Sect. 4 we consider the model with known tail distribution, and in Sect. 5 we address the problem with unknown distribution. Both the retainable arms and non-retainable arms cases are treated in these sections. Section 6 concludes the paper with some directions for further study.

## 2 Model Formulation

We consider an unlimited pool of possible objects or *arms*. The reward obtained by choosing a particular arm is deterministic, and considered as the *value* of that arm. The value of a newly chosen arm is determined as an independent sample from a fixed probability distribution, with a cumulative distribution function $F(\mu)$, $\mu \in \mathbb{R}$, that represents the empirical value distribution in the population. The obtained value is observed by the learning agent, and remains the same in future choices of that arm.

Let $I_F$ denote the support of the probability measure that corresponds to $F$. We denote $\mu^*$ as the supremal reward, i.e., the maximal value in the support $I_F$. Our performance measure will be the cumulative regret, which is defined as follows.

**Definition 1.** *The $n$-step regret 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 $t$.*

We assume that all arms values are in the interval $[0,1]$. We further use the following notations.

- $\boldsymbol{\mu}$ stands for a generic random variable with distribution $F$.
- $\boldsymbol{\mu_i}$ is the $i$-th sampled value from $F$, i.e., the revealed value of the $i$-th newly sampled arm.
- For $0 \leq \epsilon \leq 1$, let $\mu^*_\epsilon = \sup\{x \in \mathbb{R} : P(\boldsymbol{\mu} \geq x) \geq \epsilon\}$. Note that $\mu^* = \mu^*_0$. Furthermore, let
$$D(\epsilon) = \mu^* - \mu^*_\epsilon\,,$$
Note that $P\left(\boldsymbol{\mu} \geq \mu^* - D(\epsilon)\right) \geq \epsilon$, with equality if $\mu^*_{D(\epsilon)}$ is a continuity point of $F$. We refer to $D(\epsilon)$ as the *tail function* of $F$.
- Let $\epsilon^*(n)$ be defined as[2]
$$\epsilon^*(n) = \sup\left\{\epsilon \in [0,1] : nD(\epsilon) \leq \frac{1}{\epsilon}\right\}\,. \tag{2}$$
Note that $nD(\epsilon_1) \leq \frac{1}{\epsilon^*(n)}$ for $\epsilon_1 < \epsilon^*(n)$, and $nD(\epsilon_2) \geq \frac{1}{\epsilon^*(n)}$ for $\epsilon_2 > \epsilon^*(n)$.

The following property of the distribution $F$ will be needed in Sect. 5.

**Assumption 1**
$$D\left(2\epsilon\right) \geq \left(C+1\right) D\left(\epsilon\right) \tag{3}$$
*for some constant* $C > 0$ *and every* $0 \leq \epsilon \leq \frac{1}{2}$.

*Remark 1.* We observed that property (3) is satisfied in the following cases, among others:
(a) Suppose that the probability density function (p.d.f.) of $\boldsymbol{\mu}$ is strictly positive and bounded, i.e., $0 < c_1 \leq f_{\boldsymbol{\mu}}(x) \leq c_2$ for some positive constants $c_1$ and $c_2$ and for every $x \in I_F$. Then (3) is satisfied for $C = \frac{c_1}{c_2}$.
(b) If $P\left(\boldsymbol{\mu} \geq \mu^* - \epsilon\right) = c\epsilon^\beta$ for $\beta > 0$ and for every $0 \leq \epsilon \leq 1$, then $D\left(\epsilon\right) = c^{-\frac{1}{\beta}}\epsilon^{\frac{1}{\beta}}$, so that (3) is satisfied for $C = 2^{\frac{1}{\beta}}$ and every $0 \leq \epsilon \leq \frac{1}{2}$.
(c) Suppose that the p.d.f. of $\boldsymbol{\mu}$ is non decreasing. Then (3) is satisfied for $C = 1$.

## 3 Lower Bound and Some Examples

We next present a lower bound on the regret that holds for all our model variations (and, in particular, for the "easiest" case of *known distribution*, retainable arms, and given time horizon).

**Theorem 1.** *The* $n$*-step regret is lower bounded by*
$$regret(n) \geq \left(1 - \delta_n\right) \frac{\mu^* - E[\mu]}{16} \frac{1}{\epsilon^*(n)}\,, \tag{4}$$

[2] If the support of $\boldsymbol{\mu}$ is a single interval, then $D(\epsilon)$ is continuous. In that case, definition (2) reduced to the equation $nD(\epsilon) = \frac{1}{\epsilon}$ which, by monotonicity, has a unique solution for $n$ large enough. See Sect. 3 for examples.

*where $\epsilon^*(n)$ satisfies* (2)*, and*

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

*Note that when $\epsilon^*(n) \to 0$ as $n \to \infty$, $\delta_n \to 0$ as $n \to \infty$, so that its effect becomes negligible.*

*Proof.* Let $\{\mu_1, ...\mu_n\}$ denote the values of the first $n$ arms to be drawn from the pool, and assume that these values are revealed beforehand to the learning agent (even if it does not actually draw $n$ new arms in $n$ steps).

For any such sequence $\{\mu_1, ...\mu_n\}$, the smallest possible regret that can be obtained (by any algorithm) is

$$R_n^* = \min_{k\in\{1,\dots,n\}} \{\Gamma(n,k)\} ,$$

where

$$\Gamma(n,k) = n\mu^* - \left[\sum_{i=1}^{k} \mu_i + (n-k)\mu_k\right] .$$

This is due to the easily varified fact that the optimal policy for given $(\mu_i)$ is to continue sampling new arms up to some index $k^*$ and continue pulling the $k^*$-th arm thereafter.

Define the events

$$A(m,\delta_1) = \left\{\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\mu_i} < \mu^* - \delta_1\right\}$$

and

$$B(m,\delta_2) = \left\{\max_{i\in\{1,\dots,m\}} \boldsymbol{\mu_i} < \mu^* - \delta_2\right\}$$

for $m \in \{1, ..., n\}$, $0 \le \delta_1 \le \mu^*$ and $0 \le \delta_2 \le \mu^*$. If these two events are satisfied for some $m$, $\delta_1$, and $\delta_2$, we obtain that $R_n^* > m\delta_1$, for $m \le k^*$, and $R_n^* > n\delta_2$, for $m \ge k^*$, where

$$\arg\min_{k\in\{1,\dots,n\}} \{\Gamma(n,k)\} \triangleq k^* .$$

Therefore,

$$R_n^* > \min(m\delta_1, n\delta_2) .$$

Also,

$$P(A(m,\delta_1) \cap B(m,\delta_2)) \ge 1 - P(A(m,\delta_1)^c) - P(B(m,\delta_2)^c) ,$$

where $A^c$ denotes the complement of $A$. So, for $\delta_1 = \frac{1}{2}(\mu^* - E[\mu])$, by Hoeffding's inequality,

$$P(A(m,\delta_1)^c) \le \exp\left(-\frac{m}{2}(\mu^* - E[\mu])^2\right)$$

and for $\delta_2 = \frac{1}{2}D(2\epsilon^*(n))$,

$$P(B(m,\delta_2)^c) = 1 - \prod_i^m P\left(\boldsymbol{\mu_i} < \mu^* - \delta_2\right) \le 1 - (1-2\epsilon^*(n))^m \le 2\epsilon^*(n)m\,.$$

Therefore, for $m = \frac{1}{4\epsilon^*(n)}$, and $\delta_1$, $\delta_2$ as above

$$\begin{aligned} regret(n) &\ge (1 - P(A(m,\delta_1)^c) - P(B(m,\delta_2)^c)) \min\left(m\delta_1, n\delta_2\right) \\ &\ge \left(1 - \exp\left(-\frac{(\mu^* - E[\mu])^2}{8\epsilon^*(n)}\right) - \frac{1}{2}\right) \min\left(\frac{\mu^* - E[\mu]}{8\epsilon^*(n)}, \frac{n}{2}D(2\epsilon^*(n))\right) \\ &= \left(\frac{1}{2} - \exp\left(-\frac{(\mu^* - E[\mu])^2}{8\epsilon^*(n)}\right)\right) \frac{\mu^* - E[\mu]}{8\epsilon^*(n)}\,, \end{aligned}$$

where the last equality follows by (2), since $\frac{n}{2}D(2\epsilon^*(n)) \ge \frac{1}{2\epsilon^*(n)} \ge \frac{\mu^* - E[\mu]}{8\epsilon^*(n)}$. □

The main consequence of this bound is that the order of the regret is at least $\frac{1}{\epsilon^*(n)}$. As illustrate in the following examples, the order of $\frac{1}{\epsilon^*(n)}$ is typically polynomial in $n$. We will show below that all the algorithms presented in this paper attain the lower bound up to a logarithmic factors.

The papers [4] and [15] provide similar lower bounds for specific cases. In [4], a lower bound of $\sqrt{2n}$ is provided for the case where the arms values are uniformly distributed in $[0,1]$ and with Bernoulli rewards. In [15], a lower bound of order $\Omega\left(n^{\frac{\beta}{\beta+1}}\right)$ is provided for the case where $D(\epsilon) = O(\epsilon^\beta)$ with $\beta \ge 0$. Noting Example 1, our bound below is of the same order. Our proof approach is different than that of [15] and applies to more general distribution. Also, we provide a specific coefficient rather than just an order of magnitude.

The following examples serve to illustrate the dependence of $\epsilon^*(n)$ on $n$. Example 1 is the standard form studied in [15], while the others examples illustrate general cases that are covered by our model.

1. Suppose that for $\epsilon > 0$ (small enough), we have $P\left(\boldsymbol{\mu} \ge \mu^* - \epsilon\right) = \Theta\left(\epsilon^\beta\right)$, where $\beta > 0$. Then $D(\epsilon) = \Theta\left(\epsilon^{\frac{1}{\beta}}\right)$, so that $\epsilon^*(n) = \Theta\left(n^{-\frac{\beta}{\beta+1}}\right)$.
   This is the case considered in [15]. Note that $\beta = 1$ corresponds to a uniform probability distribution.
2. Suppose $\boldsymbol{\mu}$ has the CDF

$$F(\mu) = \begin{cases} (1-a)\,\frac{\mu}{\mu^*} & 0 \le \mu < \mu^* \\ 1 & \mu = \mu^* \end{cases}\,,$$

   where $0 \le a < 1$. This describes a uniform distribution with an added atom of probability $a$ at $\mu^*$. Then $D(\epsilon) = 0$ for $\epsilon \le a$, and $D(\epsilon) = \frac{\mu^*(\epsilon - a)}{1-a}$ for $\epsilon > a$. Therefore, it follows that $2\epsilon^*(n) = a + \left(a^2 + \frac{4c(1-a)}{n}\right)^{\frac{1}{2}}$.
   Note that in this case $\epsilon^*(n) > a$ for all $n$. Hence, contrary to Example 1, $\epsilon^*(n)$ does not converge to 0 as $n \to \infty$. So, the regret is finite.

3. Suppose we have $P(\boldsymbol{\mu} \geq \mu^* - \epsilon) = -\frac{c}{\ln(\epsilon)}$. We obtain that $D(\epsilon) = e^{-\frac{c}{\epsilon}}$. Therefore, it follows that $\frac{c}{\ln(n)} \leq \epsilon^*(n) \leq \frac{c+1}{\ln(n)}$.
   Note that in this case, $\epsilon^*(n)$ decays slower than any polynomial function of $n$, and the regret grows as $O(\ln(n))$.

# 4 Known Tail Function

This section discusses the model in which the tail function $D(\epsilon)$ is known (although, of course, the upper value $\mu^*$ is unknown). This model specializes the stochastic-arm model presented by Wang et al. [15] to deterministic arms. On the other hand, our model is more general in the sense that it is not restricted to tail functions of the form $D(\epsilon) = \epsilon^{\beta}$. Furthermore, we consider here both the retainable arms and the non-retainable arms problems, as described in the Introduction.

## 4.1 Retainable Arms

We propose the following algorithm.

**Algorithm 1 (KT&RA – Known Tail and Retainable Arms).**

1. *Parameters: Time horizon $n > 1$ and a constant $A > 0$.*
2. *Compute $\epsilon^*(n)$ as defined in (2).*
3. *Pull $N = \lfloor A\ln(n)\frac{1}{\epsilon^*(n)} \rfloor + 1$ arms and keep the best one so far.*
4. *Continue by pulling the saved best arm up to the last stage $n$.*

The right tradeoff between exploring new arms and pulling the best one so far is obtained by (2). The parameter $A$ allows a further tuning of the algorithm performance. Our regret bound for this algorithm is presented in the following Theorem.

**Theorem 2.** *For each $n > 1$, the regret of the KT&RA Algorithm with a constant $A$ is upper bounded by*

$$regret(n) \leq (1 + A\ln(n))\frac{1}{\epsilon^*(n)} + n^{1-A} + 1\,, \tag{5}$$

*where $\epsilon^*(n)$ is defined in (2).*

By properly choosing $A$, for example $A = 1$, we obtain an $O\left(\frac{\ln(n)}{\epsilon^*(n)}\right)$ bound on the regret. This bound is of the same order as the lower bound in (4), up to a logarithmic factor. We note that a slightly better choice of $A$ may be obtained by balancing the two terms in the bound (5).

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

$$regret(n) \leq N + (n - N)\Delta(N)\,,$$

where $\Delta(N) = E[\mu^* - V_N(1)]$. But for any $0 \leq \epsilon \leq 1$

$$P\left(\mu^* - V_N(1) > D(\epsilon)\right) \leq (1-\epsilon)^N \tag{6}$$

(note that equality holds if the distribution function of $\boldsymbol{\mu}$ is continuous) so that, since $\mu^* - V_N(1) \leq 1$,

$$\Delta(N) \leq (1-\epsilon)^N + D(\epsilon)\,. \tag{7}$$

Since in step 3 of the algorithm we chose $N = A\ln(n)\frac{1}{\epsilon(n)}$, where $\epsilon(n) < \epsilon^*(n)$, and noting that $(1-\epsilon)^{\frac{1}{\epsilon}} \leq e^{-1}$ for $\epsilon \in (0,1]$, we obtain that

$$(1-\epsilon(n))^N \leq n^{-A}\,. \tag{8}$$

Since, $\epsilon(n) < \epsilon^*(n)$, it follows that $nD\left(\epsilon(n)\right) \leq \frac{1}{\epsilon^*(n)}$. Therefore,

$$regret(n) \leq \lfloor A\ln(n)\frac{1}{\epsilon^*(n)}\rfloor + 1 + n^{1-A} + nD(\epsilon(n)) \leq A\ln(n)\frac{1}{\epsilon^*(n)} + 1 + n^{1-A} + \frac{1}{\epsilon^*(n)}\,.$$

Hence (5) is obtained. □

### 4.2 Non-retainable Arms

Here we are not allowed to keep any previously chosen arm except the last one. Therefore, the previous algorithm that keeps the best arm so far while trying out new arms cannot be applied in this case. However, the observed values of discarded arms provide usefull information for the learning agent. We introduce the notation $V_N(m)$ for the $m$-th largest value obtained after observing $N$ arms.

**Algorithm 2 (KT&NA – Known Tail and Non-retainable Arms).**

```
1. Parameters: Time horizon $n > 1$ and a constant $A \geq 2$.
2. Compute $\epsilon^*(n)$ as defined in (2).
3. Pull $N = \lfloor 5A\ln(n)\frac{1}{\epsilon^*(n)}\rfloor + 1$ arms and store their values.
4. a. Continue pulling new arms until observing a value not smaller than $V_N(m)$,
      where $m = \lceil 2A\ln(n)\rceil$.
   b. Once such a value is observed, continue pulling this arm up to the last
      stage $n$.
```

After observing $N$ arms, a threshold which is large on one hand, and on the other hand it is likely enough to find a new arm with a larger value than it is obtained. Then, the algorithm searches for an arm with a larger value than this threshold and keeps pulling this arm. Our regret bound for this algorithm is presented in the following Theorem.

**Theorem 3.** *For each $n > 1$, the regret of the KT&NA Algorithm with a constant $A$ is upper bounded by*

$$regret(n) \leq \left(5A\ln(n) + 8\right)\frac{1}{\epsilon^*(n)} + c_A(n)\,, \tag{9}$$

*where* $\epsilon^*(n)$ *is defined in* (2) *and for* $n \geq 10$ *it is obtained that*

$$c_A(n) \leq 4 \tag{10}$$

*The exact expression of* $c_A(n)$ *for* $n \geq 10$ *is found in* (15).

The algorithm starts with a learning period of length $N$, which allows to assess the values distribution near $\mu^*$. A threshold $V_N(m)$ is then set, and sampling new arms continues until an arm with that value is observed. The threshold $V_N(m)$ is chosen as the $m$-th largest value in the obtained samples, where $m$ is chosen so that the chances of quickly drawing a new arm with that value or over are high.

By a proper choice of $A$, for example $A = 2$ we obtain an $O\left(\frac{\ln(n)}{\epsilon^*(n)}\right)$ bound on the regret. This bound is of the same order as the lower bound in (4), up to a logarithmic factor. We note that by considering the exact expression of $c_A(n)$, a slightly better choice of $A$ may be obtained.

The proof of Theorem 3 relies on the following Lemma.

**Lemma 1.** *Let* $m$ *and* $N$ *be positive integers such that* $m < N$.

*(a) If* $\frac{m}{N} > \epsilon$, *then*

$$P\left(V_N(m) > \mu^*_\epsilon\right) \leq f_0(m, N, \epsilon)\,.$$

*(b) If* $\frac{m}{N} < \epsilon$, *then*

$$P\left(V_N(m) < \mu^*_\epsilon\right) \leq f_0(m, N, \epsilon)\,,$$

*where* $f_0(m, N, \epsilon) = \exp\left(-\frac{|m-N\epsilon|^2}{2(N\epsilon+|m-N\epsilon|/3)}\right)$.

For space considerations, the proof of that Lemma is presented in the technical report [8].

*Proof of Theorem 3.* The regret is bounded by

$$regret(n) \leq N + E[Y(V_N(m))] + nE[\mu^* - V_N(m)]\,, \tag{11}$$

where $N$ is the number of arms which are sampled in step 3 of the algorithm. The random variable $Y(V)$ is the number of arms which are sampled until an arm with a greater value than $V$ is sampled (or until the end of the time horizon, if such a value is never sampled again). We can find that for any $\epsilon_1 > 0$, the second term of (11) is bounded by

$$\begin{aligned} E[Y(V_N(m))] &\leq P\left(V_N(m) \leq \mu^*_{\epsilon_1}\right) E\left[Y(V_N(m))|V_N(m) \leq \mu^*_{\epsilon_1}\right] \\ &\quad + P\left(V_N(m) > \mu^*_{\epsilon_1}\right) E\left[Y(V_N(m))|V_N(m) > \mu^*_{\epsilon_1}\right] \\ &\leq \frac{1}{\epsilon_1} + nP\left(V_N(m) > \mu^*_{\epsilon_1}\right) \triangleq E_1(\epsilon_1)\,. \end{aligned} \tag{12}$$

By using the fact that $Y(V) \leq n$, the non decreasing of $Y(V)$ in $V$, and the expected value of a geometric variable. Also, for any $\epsilon_2 > 0$, the third term of (11) is bounded by

$$\begin{aligned} nE[\mu^* - V_N(m)] &\leq nP\left(V_N(m) \geq \mu^*_{\epsilon_2}\right) E\left[\mu^* - V_N(m) | V_N(m) \geq \mu^*_{\epsilon_2}\right] \\ &\quad + nP\left(V_N(m) < \mu^*_{\epsilon_2}\right) E\left[\mu^* - V_N(m) | V_N(m) < \mu^*_{\epsilon_2}\right] \\ &\leq nD(\epsilon_2) + nP\left(V_N(m) < \mu^*_{\epsilon_2}\right) \triangleq E_2(\epsilon_2) \,. \end{aligned} \tag{13}$$

Since it is known that $\mu^* - V_N(m) \leq 1$.

Therefore, by (11), (12), (13) and Lemma 1, for $\epsilon_1 = \frac{\epsilon(n)}{7}$ and $\epsilon_2 = \epsilon(n)$, where $N = 5A\ln(n)\frac{1}{\epsilon(n)}$ for some $\frac{5A\ln(n)\epsilon^*(n)}{5A\ln(n)+\epsilon^*(n)} \leq \epsilon(n) < \epsilon^*(n)$, it is obtained that

$$regret(n) \leq \lfloor \frac{5A\ln(n)}{\epsilon^*(n)} \rfloor + 1 + E_1(\epsilon_1) + E_2(\epsilon_2) \leq \frac{5A\ln(n)}{\epsilon^*(n)} + \frac{8}{\epsilon^*(n)} + c_A(n) \,, \tag{14}$$

where $\epsilon^*(n)$ is defined in (2), and

$$c_A(n) \leq 2n^{1-0.6A} + 2 \tag{15}$$

for $n \geq 10$. Note that (14) holds since $nD(\epsilon(n)) \leq \frac{1}{\epsilon^*(n)}$ and $\frac{1}{\epsilon(n)} \leq \frac{1}{\epsilon^*(n)} + \frac{1}{5A\ln(n)}$. Hence, (9) is obtained. □

## 5 Unknown Tail Function

We now proceed to the harder problem where the tail function $D(\epsilon)$ is unknown. Here, it is impossible to calculate beforehand the optimal number of arms to sample, as done in the algorithms of Sect. 4. To overcome this issue, the algorithms proposed in this section gradually increase the number of sampled arms until a certain condition is satisfied.

The analysis in this section will be carried out under Assumption 1. Note that, the values of the constant $C$ in the assumption is not used in the algorithm, but only in its analysis. Again, we consider here both the retainable arms and the non-retainable arms problems.

### 5.1 Retainable Arms

Recall that $V_N(m)$ stands for the $m$-th largest value obtained after observing $N$ arms.

**Algorithm 3 (UT&RA – Unknown Tail and Retainable Arms).**

```
1. Parameters: Time horizon $n > 1$, constants $N \geq 2$, $A \geq 2$.
   Set $N_0 = \lceil NA_n \rceil$, where $A_n = A\ln(n)$.
2. Pull $K = N_0$ arms.
3. If $\Psi(K,n) < \frac{K}{nA_n}$, where $\Psi(K,n) = V_K(1) - V_K(\lceil 5A_n \rceil)$:
   a. Pull another $K$ arms.
   b. Continue pulling the best arm so far up to time $n$.
   Else, if $\Psi(K,n) \geq \frac{K}{nA_n}$:
   a. Pull one more arm, and set $K = K + 1$.
   b. Return to 3.
```

In this algorithm, the number of sampled arms $K$ is increased until the condition in stage 3 is satisfied. Thereafter, the number of sampled arms is doubled, and then the best arm found is pulled up to time $n$.

The rational of this algorithm is as follows. Our goal is to ensure that, essentially, the number of samples $K$ is large enough, namely comparable to $\epsilon^*(n)^{-1}$ from (2). This translates to $K > nD(\frac{1}{K})$. The condition in stage 3 indicates that the gap $V_K(1) - V_K(5A_n)$, which is the difference between the largest value and the $5A_n$-th largest value in the first $K$ samples, is small in comparison to $K$. This gap is related, with high probability, to the difference $D(\frac{2}{K}) - D(\frac{1}{K})$, which, under Assumption 1, upper bounds the size of $D(\frac{1}{K})$. A second sample of $K$ arms is required due to the dependencies between the above stopping condition and values of the the first $K$ samples.

Our regret bound for this algorithm is presented in the following Theorem.

**Theorem 4.** *Let Assumption 1 hold for some $C > 0$, For each $n > 1$, the regret of the UT&RA Algorithm with a constant $A$ is upper bounded by*

$$\mathit{regret}(n) \leq \left(20A\ln(n) + \frac{1}{\min(1,C)}\right)\frac{1}{\epsilon^*(n)} + c_A(n)\,, \tag{16}$$

*where $\epsilon^*(n)$ solves* (2) *and*

$$c_A(n) \leq 2N_0 + 9 \tag{17}$$

*for $n \geq 10$. The exact expression of $c_A(n)$ for $n \geq 10$ is given in* (30).

Again, by a proper choice of $A$, for example $A = 2$ we obtain an $O\left(\frac{\ln(n)}{\epsilon^*(n)}\right)$ bound on the regret.

*Proof.* The regret is bounded by

$$\mathit{regret}(n) \leq E[2\hat{N}] + nE[\mu^* - V_{2\hat{N}}(1)]\,, \tag{18}$$

where $\hat{N}$ is the number of arms sampled by the algorithm until the condition in stage 3 is satisfied.

The first term of (18) is bounded by

$$E[2\hat{N}] \leq 2\left(N_1 + nP(\hat{N} > N_1)\right) \tag{19}$$

for every $N_1 \geq \lceil NA_n \rceil$. To bound the probability $P(\hat{N} > N_1)$, we note that

$$\begin{aligned}\left\{\hat{N} > N_1\right\} &\subseteq \left\{\Psi(N_1,n) \geq \frac{N_1}{nA_n}\right\} \subseteq \left\{\Psi(N_1,n) > \frac{N_1}{\gamma nA_n}\right\}\\ &\subseteq \left\{\Psi(N_1,n) > D\left(\frac{\gamma A_n}{N_1}\right)\right\}\bigcup E_4(\gamma,N_1)\\ &\subseteq E_3(\gamma,N_1)\bigcup E_4(\gamma,N_1)\,,\end{aligned}$$

where $\gamma > 1$,

$$E_3(\gamma,N_1) \triangleq \left\{V_{N_1}(\lceil 5A_n \rceil) < \mu^*_{\frac{\gamma A_n}{N_1}}\right\}$$

and

$$E_4(\gamma, N_1) \triangleq \left\{ D\left(\frac{\gamma A_n}{N_1}\right) > \frac{N_1}{\gamma n A_n} \right\} .$$

Note that $D(\epsilon) < \frac{1}{n\epsilon}$ for $\epsilon < \epsilon^*(n)$. So, it follows that $E_4(\gamma, N_1)$ is false, when $\frac{\gamma A_n}{N_1} < \epsilon^*(n)$, or $N_1 > \frac{\gamma A_n}{\epsilon^*(n)}$. So, for $N_1 = \max\left(\lceil \frac{\gamma A_n}{\epsilon^*(n)} + 1 \rceil, N_0\right)$, it is obtained that

$$\{\hat{N} > N_1\} \subseteq E_3(\gamma, N_1)$$

and by Lemma 1, it follows that

$$P\left(\hat{N} > N_1\right) \leq n^{-0.9A} \tag{20}$$

for $\gamma = 10$ and $n \geq 10$. Therefore, by (19),

$$E[2\hat{N}] \leq 2\left(\lceil \frac{10 A_n}{\epsilon^*(n)} + 1 \rceil + N_0 + n^{1-0.9A}\right) . \tag{21}$$

For bounding the second term of (18) we note that, for any $N_2 \geq 1$,

$$\begin{aligned} nE[\mu^* - V_{2\hat{N}}(1)] &\leq nE[\mu^* - V_{\hat{N}}(1)] \\ &\leq n\Big(E\left[\mu^* - V_{\hat{N}}(1) | \hat{N} \leq N_2\right] P(\hat{N} \leq N_2) \\ &\quad + E\left[\mu^* - V_{\hat{N}}(1) | \hat{N} > N_2\right] P(\hat{N} > N_2)\Big) \\ &\leq n\left(P(\hat{N} \leq N_2) + E\left[\mu^* - V_{N_2+1}(1)\right]\right) , \end{aligned} \tag{22}$$

where, starting from the first inequality, we consider only the $\hat{N}$ arms that were sampled after the condition in stage 3 of the algorithm has been satisfied, so that, $\hat{N}$ and the obtained values are independent. In the third inequality we use the fact that $E\left[V_m(1)\right]$ is non decreasing in $m$.

For bounding $P(\hat{N} \leq N_2)$, we note that for every $i \geq N_0$,

$$\left\{\hat{N} \leq N_2\right\} = \cup_{N_0 \leq i \leq N_2} \left\{A(i)\right\} , \tag{23}$$

where

$$\begin{aligned} A(i) &\triangleq \left\{\Psi\left(i, n\right) < \frac{i}{nA_n}\right\} \\ &\subseteq \left\{\Psi\left(i, n\right) < D\left(\frac{2A_n}{i}\right) - D\left(\frac{A_n}{i}\right)\right\} \bigcup \left\{D\left(\frac{2A_n}{i}\right) - D\left(\frac{A_n}{i}\right) < \frac{i}{nA_n}\right\} . \end{aligned}$$

Since

$$\Psi\left(i, n\right) = V_i(1) - V_i(\lceil 5A_n \rceil)$$

and

$$D\left(\frac{2A_n}{i}\right) - D\left(\frac{A_n}{i}\right) = \mu^*_{\frac{A_n}{i}} - \mu^*_{\frac{2A_n}{i}} ,$$

it follows that

$$A(i) \subseteq B(i) \bigcup C(i)\,, \tag{24}$$

where

$$B(i) \triangleq \left\{ V_i(\lceil 5A_n \rceil) > \mu^*_{\frac{2A_n}{i}} \right\} \cup \left\{ V_i(1) < \mu^*_{\frac{A_n}{i}} \right\}$$
$$C(i) \triangleq \left\{ \min(1,C) D\left(\frac{A_n}{i}\right) < \frac{i}{nA_n} \right\}$$

and the constant $C$ satisfies that $CD(\epsilon) \leq D(2\epsilon) - D(\epsilon)$ for every $0 \leq \epsilon \leq \frac{1}{2}$. So, by (23) and (24), and since

$$\cup_{N_0 \leq i \leq N_2} \{C(i)\} \subseteq C(N_2)\,,$$

it is obtained that for any $N_2 \geq N_0$ such that $C(N_2)$ is false,

$$\left\{ \hat{N} \leq N_2 \right\} = \cup_{N_0 \leq i \leq N_2} \{A(i)\} \subseteq \cup_{N_0 \leq i \leq N_2} \{B(i)\}\ .$$

Therefore, by Lemma 1, and similarly to (6) and (8) with $\epsilon = \frac{A_n}{i}$ and $N = i$, it follows that

$$P\left(\hat{N} \leq N_2\right) \leq n(n^{-1.4A} + n^{-A}) \tag{25}$$

for $n \geq 10$. Note that for $N_2 < N_0$ it is obtained that $P\left(\hat{N} \leq N_2\right) = 0$.

The remaining issue is to bound the term $E\left[\mu^* - V_{N_2+1}(1)\right]$ from (22) under the same condition that $C(N_2)$ is false. Since $\varDelta \triangleq \mu^* - V_{N_2+1}(1) \leq 1$

$$\begin{aligned} E\left[\varDelta\right] &\leq D(\frac{A_n}{N_2+1}) + P\left(\varDelta > D(\frac{A_n}{N_2+1})\right) \\ &\leq D(\frac{A_n}{N_2+1}) + n^{-A}\,. \end{aligned} \tag{26}$$

The last inequality follows similarly to (6) and (8) with $\epsilon = \frac{A_n}{N_2+1}$ and $N = N_2+1$. Let $\epsilon(n)$ be defined as

$$\epsilon(n) = \sup\left\{ \epsilon \in [0,1] : n \min(1,C) D(\epsilon) \leq \frac{1}{\epsilon} \right\}\ .$$

If it is satisfied that

$$E(C) \triangleq \left\{ \min(1,C) D(\epsilon(n)) \geq \frac{1}{n\epsilon(n)} \right\}$$

then, let us choose $N_2$ as the largest integer for which $\frac{N_2}{A_n} \leq \frac{1}{\epsilon(n)}$. Then, $C(N_2)$ is false, and furthermore $\frac{A_n}{N_2+1} < \epsilon(n)$. So,

$$D(\frac{A_n}{N_2+1}) \leq \frac{1}{n \min(1,C)\epsilon(n)}\,.$$

On the other hand, if $E(C)$ is not satisfied, then, let us choose $N_2$ as the largest integer for which $\frac{N_2}{A_n} < \frac{1}{\epsilon(n)}$. Then, again, $C(N_2)$ is false, and furthermore $\frac{A_n}{N_2+1} \leq \epsilon(n)$. So,

$$D(\frac{A_n}{N_2+1}) \leq D(\epsilon(n)) < \frac{1}{n\min(1,C)\epsilon(n)}\,.$$

Therefore, since $\min(1,C) \leq 1$, it is obtained that $\frac{1}{\epsilon(n)} \leq \frac{1}{\epsilon^*(n)}$. So,

$$D(\frac{A_n}{N_2+1}) \leq \frac{1}{n\min(1,C)\epsilon^*(n)}\,. \tag{27}$$

Therefore, by (22), (25), (26) and (27), it follows that

$$nE[\mu^* - V_{2\hat{N}}(1)] \leq n\left(n\left(n^{-1.4A} + n^{-A}\right) + \frac{1}{n\min(1,C)\epsilon^*(n)} + n^{-A}\right)\,. \tag{28}$$

Finally, by (18), (21) and (28), it is obtained that

$$regret(n) \leq \left(20A_n + \frac{1}{\min(1,C)}\right)\frac{1}{\epsilon^*(n)} + c_A(n)\,, \tag{29}$$

where

$$c_A(n) = 2n^{1-0.9A} + n^{2-1.4A} + n^{2-A} + n^{1-A} + 2NA_n + 4 \tag{30}$$

for $n \geq 10$. Hence, since $A \geq 2$, it follows that $c_A(n) \leq 2N_0 + 9$ for $n \geq 10$, so Theorem 4 is obtained. □

### 5.2 Non-retainable Arms

Here, as in Sect. 4.2, it is impossible to pull a group of arms and keep the best one of them. So, we combine the UT&RA algorithm from the previous section with the KT&NA algorithm from Sect. 4.2. Recall that (3) is satisfied for a positive constant $C$ and $\epsilon \leq \epsilon_0$, where $\epsilon_0$ is known for the learning agent.

**Algorithm 4 (UT&NA – Unknown Tail and Non-retainable Arms).**

```
1. Parameters: Time horizon $n > 1$, constants $N \geq 10$, $A \geq 4$.
   Set $N_0 = \lceil NA_n \rceil$, where $A_n = A\ln(n)$.
2. Pull $K = N_0$ arms.
3. If $\Psi(K,n) < \frac{K}{nA_n}$, where $\Psi(K,n) = V_K(1) - V_K(\lceil 5A_n \rceil)$:
   a. Pull another $K$ arms.
   b. Continue pulling new arms until observing a value equal or larger than
      $V_K(m)$, where $m = \lceil \frac{3A_n}{10} \rceil$.
   c. Continue pulling this arm up to time $n$.
   Else, if $\Psi(K,n) \geq \frac{K}{nA_n}$:
   a. Pull one more arm, and set $K = K + 1$.
   b. Return to 3.
```

This algorithm begins, similarly to the UT&RA Algorithm 3, to find a large enough sample size $K$. Then, since observed arms cannot be retained, it proceeds similarly to KT&NA Algorithm 2, to compute a desired value threshold and sample new arms until such an arm is obtained. Our regret bound for this algorithm is as follows.

**Theorem 5.** *Let Assumption 1 hold for some $C > 0$. For each $n > 1$, the regret of the UT&NA Algorithm with a constant $A$ is upper bounded by*

$$regret(n) \leq \left(20A\ln(n) + 140 + \frac{1}{\min(1,C)}\right)\frac{1}{\epsilon^*(n)} + c_A(n)\,, \tag{31}$$

*where $\epsilon^*(n)$ solves* (2) *and*

$$c_A(n) \leq 2N_0 + 14N + 13 \tag{32}$$

*for $A \geq 7$ and $n \geq 100$. The full expression of $c_A(n)$ can be found in [8].*

Similarly to the UT-LB and the KT-LB Algorithms, by a proper choice of $A$, for example $A = 7$, we obtain an $O\left(\frac{\ln(n)}{\epsilon^*(n)}\right)$ bound on the regret.

For space considerations, the proof of Theorem 5 is presented in the technical report [8].

## 6 Conclusion

For the problem of infinitely many armed-bandits with unknown value distribution, we have proposed algorithms that obtain the optimal regret up to a logarithmic factors. Our treatment was focused on the case of deterministic rewards. Further work should naturally consider the extension of our results to the stochastic rewards model, which requires repeated trials of sampled arms (possibly using a UCB-like bandit algorithm similarly to [15]). Another extension of our results, which were presented here for a given time horizon, is to the case of anytime algorithms. This can of course be accomplished using a simple doubling trick, however the development of specific and more effective algorithms for this case should be of interest. Note that in the stochastic rewards problem, it should be of interest to consider the intermediate case, where only a limited number of arms (rather than all or none) can be retained. As mentioned, in the present deterministic rewards model, it is enough to retain only the one arm with the best value so far.

**Acknowledgments.** This research was partly supported by the Technion-Microsoft Electronic Commerce Research Center.

## References

1. Amin, K., Kearns, M., Draief, M., Abernethy, J.D.: Large-scale bandit problems and KWIK learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 588–596 (2013)

2. 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)
3. Babaioff, M., Immorlica, N., Kempe, D., Kleinberg, R.: Online auctions and generalized secretary problems. ACM SIGecom Exchanges 7(2), 1–11 (2008)
4. 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)
5. Bonald, T., Proutiere, A.: Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In: Advances in Neural Information Processing Systems 26, pp. 2184–2192. Curran Associates, Inc. (2013)
6. Bubeck, S., Munos, R., Stoltz, G., Szepesvári, C.: X-armed bandits. Journal of Machine Learning Research 12, 1655–1695 (2011)
7. Chakrabarti, D., Kumar, R., Radlinski, F., Upfal, E.: Mortal multi-armed bandits. In: Advances in Neural Information Processing Systems 21, pp. 273–280. Curran Associates, Inc. (2009)
8. David, Y., Shimkin, N.: Infinitely many-armed bandits with unknown value distribution. Technical report, Technion—Israel Institute of Technology (2014), `http://webee.technion.ac.il/people/shimkin/PAPERS/ECML14-full.pdf`
9. Freeman, P.: The secretary problem and its extensions: A review. International Statistical Review, 189–206 (1983)
10. Kakade, S.M., von Luxburg, U. (eds.): COLT 2011 - The 24th Annual Conference on Learning Theory, Budapest, Hungary, June 9-11. JMLR Proceedings, vol. 19. JMLR.org (2011)
11. 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)
12. Langford, J., Zhang, T.: The epoch-greedy algorithm for multi-armed bandits with side information. In: Advances in Neural Information Processing Systems, pp. 817–824 (2007)
13. Lu, T., Pál, D., Pál, M.: Contextual multi-armed bandits. In: International Conference on Artificial Intelligence and Statistics, pp. 485–492 (2010)
14. Teytaud, O., Gelly, S., Sebag, M.: Anytime many-armed bandits. In: CAP, Grenoble, France (2007)
15. Wang, Y., Audibert, J.-Y., Munos, R.: et al. Infinitely many-armed bandits. In: Advances in Neural Information Processing Systems, vol. 8, pp. 1–8 (2008)