Infinitely many-armed bandits

Yizao Wang, Jean-Yves Audibert, Rémi Munos

Algorithms for Infinitely Many-Armed Bandits

Yizao Wang
University of Michigan, 439 West Hall, 1085 South University Street
Ann Arbor, MI 48109-1107 USA
yizwang@umich.edu

Jean-Yves Audibert
Certis - Ecole des Ponts, 6 av. Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée France
Willow - ENS / INRIA, 45 rue d'Ulm, 75005 Paris, France
audibert@certis.enpc.fr

Rémi Munos
INRIA Lille - Nord Europe, SequeL project,
40 avenue Halley, 59650 Villeneuve d'Ascq, France
remi.munos@inria.fr

Abstract

We consider multi-armed bandit problems where the number of arms is larger than the possible number of experiments. We make a stochastic assumption on the mean-reward of a new selected arm which characterizes its probability of being a near-optimal arm. Our assumption is weaker than in previous works. We describe algorithms based on upper-confidence-bounds applied to a restricted set of randomly selected arms and provide upper-bounds on the resulting expected regret. We also derive a lower-bound which matches (up to a logarithmic factor) the upper-bound in some cases.

1 Introduction

Multi-armed bandit problems describe typical situations where learning and optimization should be balanced in order to achieve good cumulative performances. Usual multi-armed bandit problems (see e.g. [9]) consider a finite number of possible actions (or arms) from which the learner may choose at each iteration. The number of arms is typically much smaller than the number of experiments allowed, so exploration of all possible options is usually performed and combined with exploitation of the apparently best ones.

In this paper, we investigate the case when the number of arms is infinite (or larger than the available number of experiments), which makes the exploration of all the arms an impossible task to achieve: if no additional assumption is made, it may be arbitrarily hard to find a near-optimal arm. Here we consider a stochastic assumption on the mean-reward of any new selected arm. When a new arm k is pulled, its mean-reward μ_k is assumed to be an independent sample of a fixed distribution. Our assumption essentially characterizes the probability of pulling near-optimal arms. It depends on two parameters $\beta \geq 0$ and $\mu^* \in [0,1]$, and writes: the probability that a new arm is ϵ-optimal is of order ϵ^β for small ϵ, i.e. $\mathbb{P}(\mu_k \geq \mu^* - \epsilon) = \Theta(\epsilon^\beta)$ for $\epsilon \to 0$[1]. Note that as a consequence, μ^* is the best possible mean-reward.

[1]We write $f(\epsilon) = \Theta(g(\epsilon))$ for $\epsilon \to 0$ when $\exists c_1, c_2, \epsilon_0 > 0$ such that $\forall \epsilon \leq \epsilon_0$, $c_1 g(\epsilon) \leq f(\epsilon) \leq c_2 g(\epsilon)$.

Like in multi-armed bandits, this setting exhibits a tradeoff between exploitation (selection of the arms that are believed to perform well) and exploration. The exploration takes two forms here: discovery (pulling a new arm that has never been tried before) and sampling (pulling an arm already discovered in order to gain information about its actual mean-reward).

Numerous applications can be found e.g. in [5]. It includes labor markets (a worker has many opportunities for jobs), mining for valuable resources (such as gold or oil) when there are many areas available for exploration (the miner can move to another location or continue in the same location, depending on results), and path planning under uncertainty in which the path planner has to decide among a route that has proven to be efficient in the past (exploitation), or a known route that has not been explored many times (sampling), or a brand new route that has never been tried before (discovery).

Let us write k_t the arm selected by our algorithm at time t. We define the regret up to time n as $R_n = n\mu^* - \sum_{t=1}^{n} \mu_{k_t}$. From the tower rule, $\mathbb{E}R_n$ is the expectation of the difference between the rewards we would have obtained by drawing an optimal arm (an arm having a mean-reward equal to μ^*) and the rewards we did obtain during the time steps $1, \ldots, n$. Our goal is to design an arm-pulling strategy such as to minimize this regret.

Overview of our results: We write $v_n = \tilde{\mathrm{O}}(u_n)$ when for some $n_0, C > 0$, $v_n \leq C u_n (\log(u_n))^2$, for all $n \geq n_0$. We assume that the rewards of the arms lie in $[0, 1]$. Our regret bounds depend on whether $\mu^* = 1$ or $\mu^* < 1$. For $\mu^* = 1$, our algorithms are such that $\mathbb{E}R_n = \tilde{O}(n^{\beta/(1+\beta)})$. For $\mu^* < 1$, we have $\mathbb{E}R_n = \tilde{O}(n^{\beta/(1+\beta)})$ if $\beta > 1$, and (only) $\mathbb{E}R_n = \tilde{O}(n^{1/2})$ if $\beta \leq 1$. Moreover we derive the lower-bound: for any $\beta > 0$, $\mu^* \leq 1$, any algorithm satisfies $\mathbb{E}R_n \geq C n^{\beta/(1+\beta)}$ for some $C > 0$. Finally we propose an algorithm having the anytime property, which is based on an arm-increasing rule.

Our algorithms essentially consist in pulling K different arms randomly chosen, where K is of order $n^{\beta/2}$ if $\mu^* < 1$ and $\beta < 1$, and $n^{\beta/(1+\beta)}$ otherwise, and using a variant of the UCB algorithm on this set of K arms, which takes into account the empirical variance of the rewards. This last point is crucial to get the proposed rate for $\mu^* = 1$ and $\beta < 1$, i.e. in cases where there are many arms with small variance.

Previous works on many-armed bandits: In [5], a specific setting of an infinitely many-armed bandit is considered, namely that the rewards are Bernoulli random variables with parameter p, where p follows a uniform law over a given interval $[0, \mu^*]$. All mean-rewards are therefore in $[0, \mu^*]$. They proposed three algorithms. (1) The *1-failure strategy* where an arm is played as long as 1s are received. When a 0 is received, a new arm is played and this strategy is repeated forever. (2) The *m-run strategy* uses the 1-failure strategy until either m continuous 1s are received (from the same arm) or m different arms have been played. In the first case, we continue to play forever the current arm. In the second case, the arm that gave the most wins is chosen to play for the remaining rounds. Finally, (3) the *m-learning strategy* uses the 1-failure strategy during the first m rounds, and for the remaining rounds it chooses the arm that gave the most 1s during the first m rounds.

For $\mu^* = 1$, the authors of [5] have shown that 1-failure strategy, $\sqrt{n}$-run strategy, and $\log(n)\sqrt{n}$-learning strategy have a regret $\mathbb{E}R_n \leq 2\sqrt{n}$. They also provided a lower bound on the regret of any strategy: $\mathbb{E}R_n \geq \sqrt{2n}$. For $\mu^* < 1$, the corresponding optimal strategies are $\sqrt{n\mu^*}$-run strategy and $\sqrt{n\mu^*}\log(n\mu^*)$-learning strategy. All these algorithms require the knowledge of the horizon n of the game. In many applications, it is important to design algorithms having the anytime property, that is, the upper bounds on the expected regret $\mathbb{E}R_n$ have the similar order for all n. Under the same Bernoulli assumption on the reward distributions, such algorithms has been obtained in [10].

In comparison to their setting (uniform distribution corresponds to $\beta = 1$), our upper- and lower-bounds are also of order $\sqrt{n}$ up to a logarithmic factor, and we do not assume that we know exactly the distribution of the mean-reward. However it is worth noting that the proposed algorithms in [5, 10] heavily depend on the Bernoulli assumption of the rewards and are not easily transposable to general distributions. Note also that the Bernoulli assumption does not work for the real problems mentioned above, where the outcomes may take several possible values.

Thus an important aspect of our work, compared to previous many-armed bandits, is that our setting allows general reward distributions for the arms, under a simple assumption on the mean-reward.

2 Main results

In our framework, each arm of a bandit is characterized by the distribution of the rewards (obtained by drawing that arm). The essential parameter of the distribution of rewards is its expectation. Another parameter of interest is the standard deviation. With low variance, poor arms will be easier to spot while good arms will have higher probability of not being disregarded at the beginning due to unlucky trials. To draw an arm is equivalent to draw a distribution ν of rewards. Let $\mu = \int w\nu(dw)$ and $\sigma^2 = \int (w-\mu)^2\nu(dw)$ denote the expectation and variance of ν. The quantities μ and σ are random variables. Our assumptions are the following:

(A) all rewards are uniformly bounded: without loss of generality, we may assume that the rewards are in $[0,1]$ (more general distributions can be considered[2])

(B) the expected reward of a randomly drawn arm satisfies: there exist $\mu^* \in (0,1]$ and $\beta > 0$ s.t.

$$\mathbb{P}\{\mu > \mu^* - \epsilon\} = \Theta(\epsilon^\beta), \text{ for } \epsilon \to 0 \tag{1}$$

(C) there is a function $V : [0,1] \to \mathbb{R}$ such that $\mathbb{P}\{\sigma^2 \leq V(\mu^* - \mu)\} = 1$.

The key assumption here is (B). It gives us (the order of) the number of arms that needs to be drawn before finding an arm that is ϵ-close to the optimum[3] (i.e., an arm for which $\mu \geq \mu^* - \epsilon$). Assumption (B) implies that there exists positive constants c_1 and c_2 such that for any $\epsilon \in [0, \mu^*]$, we have[4]

$$c_1\epsilon^\beta \leq \mathbb{P}\{\mu > \mu^* - \epsilon\} \leq \mathbb{P}\{\mu \geq \mu^* - \epsilon\} \leq c_2\epsilon^\beta. \tag{2}$$

Assumption (C) is weak to the extent that it always holds for $V(u) = \mu^*(1-\mu^*+u)$ (here we use $\mathbb{V}\text{ar}\, W \leq \mathbb{E}W(1-\mathbb{E}W)$), but it is convenient when one may want to deal with bandit problems where near-optimal arms have low variance (for instance, this happens when $\mu^* = 1$).

Let $X_{k,1}, X_{k,2}, \ldots$ denote the rewards obtained when pulling arm k. These are i.i.d. random variables with common expected value denoted μ_k. Let $\overline{X}_{k,s} \triangleq \frac{1}{s}\sum_{j=1}^{s} X_{k,j}$ and $V_{k,s} \triangleq \frac{1}{s}\sum_{j=1}^{s}(X_{k,j} - \overline{X}_{k,s})^2$ be the empirical mean and variance associated with the first s draws of arm k. Let $T_k(t)$ denote the number of times arm k is chosen by the policy during the first t plays. We will use as a subroutine of our algorithms the following version of UCB as introduced in [2]. Let $(\mathcal{E}_t)_{t\geq 0}$ be a nondecreasing sequence of nonnegative real numbers. It will be referred to as the exploration sequence since the larger it is, the more UCB explores. For any arm k and nonnegative integers s, t, introduce

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\mathcal{E}_t}{s}} + \frac{3\mathcal{E}_t}{s} \tag{3}$$

with the convention $1/0 = +\infty$. Define the UCB-V (for Variance estimate) policy:

UCB-V policy for a set $\mathcal{K}$ of arms:
At time t, play an arm in $\mathcal{K}$ maximizing $B_{k,T_k(t-1),t}$.

From [2, Theorem 1], the main property of $B_{k,s,t}$ is that with probability at least $1 - 5(\log t)e^{-\mathcal{E}_t/2}$, for any $s \in [0,t]$ we have $\mu_k \leq B_{k,s,t}$. So provided that $\mathcal{E}_t$ is large, $B_{k,T_k(t-1),t}$ is an observable quantity at time t which upper bounds μ_k with high probability. We consider nondecreasing sequence $(\mathcal{E}_t)$ in order that these bounds hold with probability increasing with time. This ensures that the low probability event, that the algorithm might concentrate the draws on suboptimal arms, has a decreasing probability with time.

[2] This is done in [7, see p.49 and Chap.4] where it is assumed that the distributions have uniformly bounded exponential moments.

[3] Precise computations lead to a number which is of order $\epsilon^{-\beta}$ up to possibly a logarithmic factor.

[4] Indeed, (1) implies that for some $0 < c_1' < c_2'$, there exists $0 < \epsilon_0 < \mu^*$ such that for any $\epsilon \leq \epsilon_0$, $c_1'\epsilon^\beta \leq \mathbb{P}\{\mu > \mu^* - \epsilon\} \leq \mathbb{P}\{\mu \geq \mu^* - \epsilon\} \leq c_2'\epsilon^\beta$. Straightforward computations show that one may take $c_1 = c_1'\epsilon_0^\beta$ and $c_2 = \max(\epsilon_0^{-\beta}, c_2')$.

2.1 UCB revisited for the infinitely many-armed bandit

When the number of arms of the bandit is greater than the total number of plays, it makes no sense to apply UCB-V algorithm (or other variants of UCB [3]) since its first step is to draw each arm once (to have $B_{k,T_k(t-1),t}$ finite). A more meaningful and natural approach is to decide at the beginning that only K arms will be investigated in the entire experiment. The K should be sufficiently small with respect to n (the total number of plays), as in this way we have fewer plays on bad arms and most of the plays will be on the best of K arms. The number K should not be too small either, since we want that the best of the K arms has an expected reward close to the best possible arm.

It is shown in [2, Theorem 4] that in the multi-armed bandit, taking a too small exploration sequence (e.g. such as $\mathcal{E}_t \leq \frac{1}{2}\log t$) might lead to polynomial regret (instead of logarithmic for e.g. $\mathcal{E}_t = 2\log t$) in a simple 2-armed bandit problem. However, we will show that this is not the case in the infinitely many-armed bandit, where one may (and should) take much smaller exploration sequences (typically of order $\log\log t$). The reason for this phenomenon is that in this setting, there are typically many near-optimal arms so that the subroutine UCB-V may miss some good arms (by unlucky trials) without being hurt: there are many other near-optimal arms to discover! This illustrates a tradeoff between the two aspects of exploration: sample the current, not well-known, arms or discover new arms.

We will start our analysis by considering the following UCB-V(∞) algorithm:

UCB-V(∞) algorithm: Given parameters K and the exploration sequence $(\mathcal{E}_t)$

- Randomly choose K arms,
- Run the UCB-V policy on the set of the K selected arms.

Theorem 1 *If the exploration sequence satisfies* $2\log(10\log t) \leq \mathcal{E}_t \leq \log t$, *then for* $n \geq 2$ *and* $K \geq 2$ *the expected regret of the UCB-V(∞) algorithm satisfies:*

$$\mathbb{E}R_n \leq C\Big\{(\log K)nK^{-1/\beta} + K(\log n)\mathbb{E}\Big[\big(\tfrac{V(\Delta)}{\Delta}+1\big)\wedge(n\Delta)\Big]\Big\}, \tag{4}$$

where $\Delta = \mu^* - \mu$ *with* μ *the random variable corresponding to the expected reward of a sampled arm from the pool, and where* C *is a positive constant depending only on* c_1 *and* β *(see* (2)*).*

Proof: The UCB-V(∞) algorithm has two steps: randomly choose K arms and run a UCB subroutine on the selected arms. The first part of the proof studies what happens during the UCB subroutine, that is, conditionally to the arms that have been randomly chosen during the first step of the algorithm. In particular we consider in the following that $\mu_1, \ldots, \mu_K$ are fixed. From the equality (obtained using Wald's theorem):

$$\mathbb{E}R_n = \textstyle\sum_{k=1}^K \mathbb{E}\{T_k(n)\}\Delta_k \tag{5}$$

with $\Delta_k = \mu^* - \mu_k$, it suffices to bound $\mathbb{E}T_k(n)$. The proof is inspired from the ones of Theorems 2 and 3 in [2]. The novelty of the following lemma is to make appear the product of probabilities in the last term of the right-hand-side. This enables us to incorporate the idea that if there are a lot of near-optimal arms, it is very unlikely that suboptimal arms are often drawn.

Lemma 1 *For any real number* τ *and any positive integer* u, *we have*

$$\mathbb{E}T_k(n) \leq u + \textstyle\sum_{t=u+1}^n \sum_{s=u}^t \mathbb{P}\big(B_{k,s,t} > \tau\big) + \sum_{t=u+1}^n \prod_{k'\neq k} \mathbb{P}\big(\exists s' \in [0,t],\ B_{k',s',t} \leq \tau\big) \tag{6}$$

where the expectations and probabilities are conditionally to the set of selected arms.

Proof: We have $T_k(n) - u \leq \sum_{t=u+1}^n Z_k(u,t)$ where $Z_k(u,t) = \mathbf{1}_{I_t=k;T_k(t)>u}$. We have

$$\begin{aligned} Z_k(u,t) &\leq \mathbf{1}_{\forall k'\neq k\ B_{k,T_k(t-1),t} \geq B_{k',T_{k'}(t-1),t};T_k(t-1)\geq u} \\ &\leq \mathbf{1}_{\exists s\in[u,t]\ B_{k,s,t}>\tau} + \mathbf{1}_{\forall k'\neq k\ \exists s'\in[0,t]\ B_{k',s',t}\leq\tau} \end{aligned}$$

where the last inequality holds since if the two terms in the last sum are equal to zero, then it implies that there exists $k' \neq k$ such that for any $s' \in [0,t]$ and any $s \in [u,t]$, $B_{k',s',t} > \tau \geq B_{k,s,t}$. Taking the expectation of both sides, using a union bound and the independence between rewards obtained from different arms, we obtain Lemma 1. □

Now we use Inequality (6) with $\tau = \frac{\mu^* + \mu_k}{2} = \mu_k + \frac{\Delta_k}{2} = \mu^* - \frac{\Delta_k}{2}$, and u the smallest integer larger than $32\Big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{1}{\Delta_k}\Big)\log n$. These choices are made to ensure that the probabilities in the r.h.s. of (6) are small. Precisely, for any $s \geq u$ and $t \leq n$, we have

$$\begin{aligned}\sqrt{\frac{2[\sigma_k^2 + \Delta_k/4]\mathcal{E}_t}{s}} + 3\frac{\mathcal{E}_t}{s} &\leq \sqrt{\frac{[2\sigma_k^2 + \Delta_k/2]\log n}{u}} + 3\frac{\log n}{u} \\ &\leq \sqrt{\tfrac{[2\sigma_k^2+\Delta_k/2]\Delta_k^2}{32[\sigma_k^2+\Delta_k]}} + \tfrac{3\Delta_k^2}{32[\sigma_k^2+\Delta_k]} = \tfrac{\Delta_k}{4}\left[\sqrt{\tfrac{\sigma_k^2+\Delta_k/4}{\sigma_k^2+\Delta_k}} + \tfrac{3}{8}\tfrac{\Delta_k}{\sigma_k^2+\Delta_k}\right] \leq \tfrac{\Delta_k}{4},\end{aligned}$$

where the last inequality holds since it is equivalent to $(x-1)^2 \geq 0$ for $x = \sqrt{\frac{\sigma_k^2+\Delta_k/4}{\sigma_k^2+\Delta_k}}$. Thus:

$$\begin{aligned}\mathbb{P}(B_{k,s,t} > \tau) &\leq \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2V_{k,s}\mathcal{E}_t}{s}} + 3\frac{\mathcal{E}_t}{s} > \mu_k + \Delta_k/2\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2[\sigma_k^2+\Delta_k/4]\mathcal{E}_t}{s}} + 3\tfrac{\mathcal{E}_t}{s} > \mu_k + \Delta_k/2\big) + \mathbb{P}\big(V_{k,s} \geq \sigma_k^2 + \Delta_k/4\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} - \mu_k > \Delta_k/4\big) + \mathbb{P}\Big(\tfrac{\sum_{j=1}^s (X_{k,j}-\mu_k)^2}{s} - \sigma_k^2 \geq \Delta_k/4\Big) \\ &\leq 2e^{-s\Delta_k^2/(32\sigma_k^2+8\Delta_k/3)},\end{aligned} \tag{7}$$

where in the last step we used Bernstein's inequality twice.

Summing up these probabilities we obtain

$$\begin{aligned}\sum_{s=u}^{t}\mathbb{P}(B_{k,s,t} > \tau) &\leq 2\sum_{s=u}^{\infty} e^{-s\Delta_k^2/(32\sigma_k^2+8\Delta_k/3)} = 2\frac{e^{-u\Delta_k^2/(32\sigma_k^2+8\Delta_k/3)}}{1-e^{-\Delta_k^2/(32\sigma_k^2+8\Delta_k/3)}} \\ &\leq \left(\tfrac{80\sigma_k^2}{\Delta_k^2} + \tfrac{7}{\Delta_k}\right)e^{-u\Delta_k^2/(32\sigma_k^2+8\Delta_k/3)} \leq \left(\tfrac{80\sigma_k^2}{\Delta_k^2} + \tfrac{7}{\Delta_k}\right)n^{-1},\end{aligned} \tag{8}$$

where we have used that $1 - e^{-x} \geq 4x/5$ for $0 \leq x \leq 3/8$. Now let us bound the product of probabilities in (6). Since $\tau = \mu^* - \Delta_k/2$, we have

$$\prod_{k' \neq k}\mathbb{P}(\exists s \in [0,t],\ B_{k',s,t} \leq \tau\big) \leq \prod_{k':\mu_{k'} > \mu^* - \Delta_k/2}\mathbb{P}(\exists s \in [0,t],\ B_{k',s,t} < \mu_k'\big)\,.$$

Now from [2, Theorem 1], with probability at least $1 - 5(\log t)e^{-\mathcal{E}_t/2}$, for any $s \in [0,t]$ we have $\mu_k \leq B_{k,s,t}$. For $\mathcal{E}_t \geq 2\log(10\log t)$, this gives $\mathbb{P}(\exists s \in [0,t],\ B_{k',s,t} < \mu_k'\big) \leq 1/2$. Putting all the bounds of the different terms of (6) leads to

$$\mathbb{E}T_k(n) \leq 1 + 32\Big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{1}{\Delta_k}\Big)\log n + \left(\frac{80\sigma_k^2}{\Delta_k^2} + \frac{7}{\Delta_k}\right) + n2^{-N_{\Delta_k}},$$

with N_{Δ_k} the cardinal of $\big\{k' \in \{1,\dots,K\} : \mu_{k'} > a - \Delta_k/2\big\}$. Since $\Delta_k \leq \mu^* \leq 1$ and $T_k(n) \leq n$, the previous inequality can be simplified into

$$\mathbb{E}T_k(n) \leq \Big\{\Big[50\Big(\tfrac{\sigma_k^2}{\Delta_k^2} + \tfrac{1}{\Delta_k}\Big)\log n\Big] \wedge n\Big\} + n2^{-N_{\Delta_k}}, \tag{9}$$

Here, for sake of simplicity, we are not interested in having tight constants. From now and on, we will take the expectations with respect to all sources of randomness, that is including the one coming from the first step of UCB-V(∞). The quantities $\Delta_1, \dots, \Delta_K$ are i.i.d. random variables satisfying $0 \leq \Delta_k \leq \mu^*$ and $\mathbb{P}(\Delta_k \leq \epsilon) = \Theta(\epsilon^\beta)$. The quantities $\sigma_1, \dots, \sigma_k$ are i.i.d. random variables satisfying almost surely $\sigma_k^2 \leq V(\Delta_k)$. From (5) and (9), we have

$$\mathbb{E}R_n = K\mathbb{E}\big\{T_1(n)\Delta_1\big\} \leq K\mathbb{E}\Big\{\Big[50\Big(\tfrac{V(\Delta_1)}{\Delta_1} + 1\Big)\log n\Big] \wedge (n\Delta_1) + n\Delta_1 2^{-N_{\Delta_1}}\Big\} \tag{10}$$

Let p denote the probability that the expected reward μ of a randomly drawn arm satisfies $\mu > \mu^* - \delta/2$ for a given δ. Conditioning on $\Delta_1 = \delta$, the quantity N_{Δ_1} follows a binomial distribution with parameters $K-1$ and p, hence $\mathbb{E}(2^{-N_{\Delta_1}}|\Delta_1 = \delta) = (1 - p + p/2)^{K-1}$. By using (2), we get:

$$\mathbb{E}\big\{\Delta_1 2^{-N_{\Delta_1}}\big\} = \mathbb{E}\big\{\Delta_1(1 - \mathbb{P}(\mu > \mu^* - \Delta_1/2)/2)^{K-1}\big\} \leq \mathbb{E}\chi(\Delta_1),$$

with $\chi(u) = u(1-c_3u^\beta)^{K-1}$ and $c_3 = c_1/2^\beta$. We have $\chi'(u) = (1-c_3u^\beta)^{K-2}\big[1-c_3(1+(K-1)\beta)u^\beta\big]$ so that $\chi(u) \le \chi(u_0)$ with $u_0 = \frac{1}{[c_3(1+(K-1)\beta)]^{1/\beta}}$ and $\chi(u_0) = \frac{(1-\frac{1}{1+(K-1)\beta})^{K-1}}{[c_3(1+(K-1)\beta)]^{1/\beta}} \le C'K^{-1/\beta}$ for C' a positive constant depending only c_1 and β. For any $u_1 \in [u_0, \mu^*]$, we have

$$\mathbb{E}\chi(\Delta_1) \le \chi(u_0)\mathbb{P}(\Delta_1 \le u_1) + \chi(u_1)\mathbb{P}(\Delta_1 > u_1) \le \chi(u_0)\mathbb{P}(\Delta_1 \le u_1) + \chi(u_1)\,.$$

Let us take $u_1 = C''\big(\frac{\log K}{K}\big)^{1/\beta}$ for C'' a positive constant depending on c_1 and β sufficiently large to ensure $u_1 \ge u_0$ and $\chi(u_1) \le K^{-1-1/\beta}$. We obtain $\mathbb{E}\chi(\Delta_1) \le CK^{-1/\beta}\frac{\log K}{K}$ for an appropriate constant C depending on c_1 and β. Putting this into (10), we obtain the result of Theorem 1. □

The r.h.s. of Inequality (4) contains two terms. The first term is the bias: when we randomly draw K arms, the expected reward of the best drawn arm is $\tilde{O}(K^{-1/\beta})$-optimal. So the best algorithm, once the K arms are fixed, will lead to a regret $\tilde{O}(nK^{-1/\beta})$. The second term is the estimation term. It indicates to what extent the UCB subroutine performs almost as well as the best arm among the K selected ones.

2.2 Strategy for fixed play number

Consider that we know in advance the total number of plays n and the value of β. In this case, one can use the UCB-V(∞) algorithm with parameter K of order of the minimizer of the r.h.s. of Inequality (4). This leads to the following UCB-F (for Fixed horizon) algorithm.

UCB-F (fixed horizon): given total number of plays n, and parameters μ^* and β of (1)

- Choose K arms with K of order $\begin{cases} n^{\frac{\beta}{2}} & \text{if } \beta < 1, \mu^* < 1 \\ n^{\frac{\beta}{\beta+1}} & \text{otherwise, i.e. if } \mu^* = 1 \text{ or } \beta \ge 1 \end{cases}$
- Run the UCB-V algorithm with the K chosen arms and an exploration sequence satisfying

$$2\log(10\log t) \le \mathcal{E}_t \le \log t \tag{11}$$

Theorem 2 *For any $n \ge 2$, the expected regret of the UCB-F algorithm satisfies*

$$\mathbb{E}R_n \le \begin{cases} C(\log n)\sqrt{n} & \textit{if } \beta < 1 \textit{ and } \mu^* < 1 \\ C(\log n)^2\sqrt{n} & \textit{if } \beta = 1 \textit{ and } \mu^* < 1 \\ C(\log n)n^{\frac{\beta}{1+\beta}} & \textit{otherwise, i.e. if } \mu^* = 1 \textit{ or } \beta > 1 \end{cases} \tag{12}$$

with C a constant depending only on c_1, c_2 and β (see (2)*).*

Proof: The result comes from Theorem 1 by bounding the expectation $E = \mathbb{E}\big[\big(\frac{V(\Delta)}{\Delta}+1\big)\wedge(n\Delta)\big]$. First, as mentioned before, Assumption (C) is satisfied for $V(\Delta) = \mu^*(1-\mu^*+\Delta)$. So for $\mu^* = 1$ and this choice of function V, we have $E \le 2$. For $\mu^* < 1$, since $\Delta \le \mu^*$, we have $E \le \mathbb{E}\Psi(\Delta)$ with $\Psi(t) = \frac{2\mu^*}{t} \wedge (nt)$. The function Ψ is continuous and differentiable by parts. Using Fubini's theorem and Inequality (2), we have

$$\begin{aligned} \mathbb{E}\Psi(\Delta) &= \Psi(\mu^*) - \mathbb{E}\textstyle\int_\Delta^{\mu^*}\Psi'(t)dt = \Psi(\mu^*) - \int_0^{\mu^*}\Psi'(t)\mathbb{P}(\Delta \le t)dt \\ &\le 2 + \textstyle\int_{\sqrt{2/n}}^1 \frac{2}{t^2}c_2t^\beta dt \le \begin{cases} 2 + \frac{2^{(1+\beta)/2}c_2}{1-\beta}n^{\frac{1-\beta}{2}} & \text{if } \beta < 1 \\ 2 + c_2\log(n/2) & \text{if } \beta = 1 \\ 2 + \frac{2c_2}{\beta-1} & \text{if } \beta > 1 \end{cases}. \end{aligned}$$

Putting these bounds in Theorem 1, we get

$$\mathbb{E}R_n \le \begin{cases} C\Big\{(\log K)nK^{-1/\beta} + (\log n)Kn^{\frac{1-\beta}{2}}\Big\} & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ C\Big\{(\log K)nK^{-1/\beta} + (\log n)^2K\Big\} & \text{if } \beta = 1 \text{ and } \mu^* < 1 \\ C\Big\{(\log K)nK^{-1/\beta} + (\log n)K\Big\} & \text{otherwise: } \mu^* = 1 \text{ or } \beta > 1 \end{cases}$$

with C a constant only depending on c_1, c_2 and β. The number K of selected arms in UCB-F is taken of the order of the minimizer of these bounds up to a logarithmic factor. □

Theorem 2 makes no difference between a logarithmic exploration sequence and an iterated logarithmic exploration sequence. However in practice, it is clearly better to take an iterated logarithmic exploration sequence, for which the algorithm spends much less time on exploring all suboptimal arms. For sake of simplicity, we have fixed the constants in (11). It is easy to check that for $\mathcal{E}_t = \zeta \log_t$ and $\zeta \geq 1$, Inequality (12) still holds but with a constant C depending linearly in ζ.

Theorem 2 shows that when $\mu^* = 1$ or $\beta \geq 1$, the bandit subroutine takes no time in spotting near-optimal arms (the use of UCB-V algorithm using variance estimate is crucial for this), whereas for $\beta < 1$ and $\mu^* < 1$, which means a lot of near-optimal arms with possibly high variances, the bandit subroutine has difficulties in achieving low regret.

The next theorem shows that our regret upper bounds are optimal up to logarithmic terms except for the case $\beta < 1$ and $\mu^* < 1$ for which it remains an open problem.

Theorem 3 *Any algorithm suffers a regret larger than $cn^{\frac{\beta}{1+\beta}}$ for some small enough constant c depending on c_2 and β.*

Sketch of proof. (The full proof is provided in Appendix A) If we want to have a regret smaller than $cn^{\beta/(1+\beta)}$ we need that most draws are done on an arm having an individual regret smaller than $\epsilon = cn^{-1/(1+\beta)}$. To find such an arm, we need to try a number of arms larger than $C'\epsilon^{-\beta} = C'c^{-\beta}n^{\beta/(1+\beta)}$ arms for some $C' > 0$ depending on c_2 and β. Since these arms are drawn at least once and since most of these arms give a constant regret, it leads to a regret larger than $C''c^{-\beta}n^{\beta/(1+\beta)}$ with C'' depending on c_2 and β. For c small enough, this contradicts that the regret is smaller than $cn^{\beta/(1+\beta)}$. So it is not possible to improve on the $n^{\beta/(1+\beta)}$ rate. We do not know whether the rate $n^{\beta/2}$ for $\beta < 1$ and $\mu^* < 1$ is improvable. □

2.3 Strategy for unknown play number

To apply the UCB-F algorithm we need to know the total number of plays n and we choose the corresponding K arms before starting. When n is unknown ahead of time, we propose here an anytime algorithm with a simple and reasonable way of choosing K by adding a new arm from time to time into the set of sampled arms. Let K_n denote the number of arms played up to time n. We set $K_0 = 0$. We define the UCB-AIR (for Arm-Increasing Rule):

UCB-AIR (Arm-Increasing Rule): given parameters μ^* and β of (1),

- at time n, try a new arm if

$$K_{n-1} < \begin{cases} n^{\frac{\beta}{2}} & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ n^{\frac{\beta}{\beta+1}} & \text{otherwise: } \mu^* = 1 \text{ or } \beta \geq 1 \end{cases}$$

- otherwise apply the UCB-V policy on the already drawn arms with an exploration sequence satisfying

$$2\log(10\log t) \leq \mathcal{E}_t \leq \log t$$

This arm-increasing rule makes our algorithm applicable for the anytime problem. This is a more reasonable approach in practice than restarting-based algorithms like the ones using the doubling trick (see e.g. [4, Section 5.3]). Our second main result is to show that the UCB-AIR algorithm has the same properties as the UCB-F algorithm (proof given in Appendix B).

Theorem 4 *For any horizon time $n \geq 2$, the expected regret of the UCB-AIR algorithm satisfies*

$$\mathbb{E}R_n \leq \begin{cases} C(\log n)^2\sqrt{n} & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ C(\log n)^2 n^{\frac{\beta}{1+\beta}} & \text{otherwise, i.e. if } \mu^* = 1 \text{ or } \beta \geq 1 \end{cases} \tag{13}$$

with C a constant depending only on c_1, c_2 and β (see (2)*).*

3 Comparison with continuum-armed bandits and conclusion

In continuum-armed bandits (see e.g. [1, 6, 4]), an infinity of arms is also considered. The arms lie in some Euclidean (or metric) space and their mean-reward is a deterministic and smooth (e.g. Lipschitz) function of the arms. This setting is different from ours since our assumption is stochastic and does not consider regularities of the mean-reward w.r.t. the arms. However, if we choose an arm-pulling strategy which consists in selecting randomly the arms, then our setting encompasses continuum-armed bandits. For example, consider the domain $[0,1]^d$ and a mean-reward function μ assumed to be locally equivalent to a Hölder function (of order $\alpha \in [0, +\infty)$) around any maximum x^* (the number of maxima is assumed to be finite), i.e.

$$\mu(x^*) - \mu(x) = \Theta(||x^* - x||^\alpha) \text{ when } x \to x^*. \tag{14}$$

Pulling randomly an arm X according to the Lebesgue measure on $[0,1]^d$, we have: $\mathbb{P}(\mu(X) > \mu^* - \epsilon) = \Theta(\mathbb{P}(||X - x^*||^\alpha < \epsilon)) = \Theta(\epsilon^{d/\alpha})$, for $\epsilon \to 0$. Thus our assumption (1) holds with $\beta = d/\alpha$, and our results say that if $\mu^* = 1$, we have $\mathbb{E}R_n = \tilde{O}(n^{\beta/(1+\beta)}) = \tilde{O}(n^{d/(\alpha+d)})$.

For $d = 1$, under the assumption that μ is α-Hölder (i.e. $|\mu(x) - \mu(y)| \leq c||x - y||^\alpha$ for $0 < \alpha \leq 1$), [6] provides upper- and lower-bounds on the regret $R_n = \Theta(n^{(\alpha+1)/(2\alpha+1)})$. Our results gives $\mathbb{E}R_n = \tilde{O}(n^{1/(\alpha+1)})$ which is better for all values of α. The reason for this apparent contradiction is that the lower bound in [6] is obtained by the construction of a function very irregular, which actually does not satisfy our local assumption (14).

Now, under assumptions (14) for any $\alpha > 0$ (around a finite set of maxima), [4] provides the rate $\mathbb{E}R_n = \tilde{O}(\sqrt{n})$. Our result gives the same rate when $\mu^* < 1$ but in the case $\mu^* = 1$ we obtain the improved rate $\mathbb{E}R_n = \tilde{O}(n^{1/(\alpha+1)})$ which is better whenever $\alpha > 1$ (because we are able to exploit the low variance of the good arms). Note that like our algorithm, the algorithms in [4], as well as in [6], do not make an explicit use (in the procedure) of the smoothness of the function. They just use a 'uniform' discretization of the domain.

On the opposite, the zooming algorithm of [8] adapts to the smoothness of μ (more arms are sampled at areas where μ is high). For any dimension d, they obtain $\mathbb{E}R_n = \tilde{O}(n^{(d'+1)/(d'+2)})$, where $d' \leq d$ is their 'zooming dimension'. Under assumptions (14) we deduce $d' = \frac{\alpha-1}{\alpha}d$ using the Euclidean distance as metric, thus their regret is $\mathbb{E}R_n = \tilde{O}(n^{\frac{d(\alpha-1)+\alpha}{d(\alpha-1)+2\alpha}})$. For locally quadratic functions (i.e. $\alpha = 2$), their rate is $\tilde{O}(n^{(d+2)/(d+4)})$, whereas ours is $\tilde{O}(n^{d/(2+d)})$. Again, we have a smaller regret although we do not use the smoothness of μ in our algorithm. Here the reason is that the zooming algorithm does not make full use of the fact that the function is locally quadratic (it considers a Lipschitz property only). However, in the case $\alpha < 1$, our rates are worse than algorithms specifically designed for continuum armed bandits.

Hence, the comparison between the many-armed and continuum-armed bandits settings is not easy because of the difference in nature of the basis assumptions. Our setting is an alternative to the continuum-armed bandit setting which does not require that there exists an underlying metric space in which the mean-reward function would be smooth. Our assumption (1) naturally deals with possible very complicated functions where maxima may be located in any part of the space. Our algorithm will be competitive when applied to (and compared to algorithms specifically designed for) the continuum-armed bandit problem when there are relatively many near-optimal arms, which matches the intuition that in such cases, a random selection strategy will perform well.

To conclude, our contributions are: (i) Compared to previous many-armed bandits, our setting allows general reward distributions for the arms, under a simple assumption on the probability of pulling near-optimal arms. (ii) We show that, since there is an infinity of arms, we need much less exploration of each arm than UCB variants applied to finite-armed bandits (the usual $\log$ term may be replaced by $\log\log$). (iii) Our UCB variant, making use of the variance estimate, enables to obtain higher rates in cases when the variance of the good arms is small. (iv) The UCB-AIR algorithm is anytime, making use of an arm increasing rule (instead of a doubling trick). (v) We provide a lower-bound matching the upper-bound (up to a logarithmic factor) in the case $\beta \geq 1$ or $\mu^* = 1$.

A Theorem 3

Proof of Theorem 3. An elementary event of the probability space is characterized by the infinite sequence $I_1, I_2, \dots$ of arms and by the infinite sequences of rewards corresponding to each of the arm: $X_{I_1,1}, X_{I_1,2}, \dots, X_{I_2,1}, X_{I_2,1}, \dots$, and so on. Arm I_1 is the first arm drawn, $I_2 \neq I_1$ is the second one, and so on. Let $0 < \delta < \delta' < \mu^*$. Let K^* denote the smallest ℓ such that $\mu_{I_\ell} > \mu^* - \delta$. Let $\bar{K}$ be the number of arms in $\{I_1, \dots, I_{K^*-1}\}$ with expected reward smaller than or equal to $\mu^* - \delta'$. An algorithm will request a number of arms K, which is a random variable (possibly depending on the obtained rewards). Let $\hat{\mu}$ be the expected reward of the best arm in $\{I_1, \dots, I_K\}$. Let $\kappa > 0$ a parameter to be chosen. We have

$$\begin{aligned} R_n &= R_n \mathbf{1}_{\hat{\mu} \leq \mu^* - \delta} + R_n \mathbf{1}_{\hat{\mu} > \mu^* - \delta} \\ &\geq n\delta \mathbf{1}_{\hat{\mu} \leq \mu^* - \delta} + \bar{K}\delta' \mathbf{1}_{\hat{\mu} > \mu^* - \delta} \\ &\geq n\delta \mathbf{1}_{\hat{\mu} \leq \mu^* - \delta} + \kappa\delta' \mathbf{1}_{\hat{\mu} > \mu^* - \delta; \bar{K} \geq \kappa}, \end{aligned}$$

where the first inequality uses that $\hat{\mu} > \mu^* - \delta$ implies that the arms $I_1, \dots, I_{K^*}$ have been at least tried once. By taking expectations on both sides and taking $\kappa = n\delta/\delta'$, we get

$$\mathbb{E}R_n \geq n\delta \mathbb{P}(\hat{\mu} \leq \mu^* - \delta) + \kappa\delta' \big(\mathbb{P}(\hat{\mu} > \mu^* - \delta) - \mathbb{P}(\bar{K} < \kappa)\big) = \delta'\kappa \mathbb{P}(\bar{K} \geq \kappa).$$

Now the random variable $\bar{K}$ follows a geometric distribution with parameter $p = \frac{\mathbb{P}(\mu > \mu^* - \delta)}{\mathbb{P}(\mu \notin (\mu^* - \delta', \mu^* - \delta])}$. So we have $\mathbb{E}R_n \geq \delta'\kappa(1-p)^\kappa$. Taking $\delta = \delta' n^{-1/(\beta+1)}$ and δ' a constant value in $(0, \mu^*)$ (for instance $(2c_2)^{-1/\beta}$ to ensure $p \leq 2c_2\delta^\beta$), we have $\kappa = n^{\frac{\beta}{1+\beta}}$ and p is of order $1/\kappa$ and obtain the desired result.

B Theorem 4

Proof of Theorem 4. We essentially need to adapt the proof of Theorem 1. We recall that K_n denote the number of arms played up to time n. Let $I_1, \dots, I_{K_n}$ denote the selected arms: I_1 is the first arm drawn, I_2 the second, and so on. Let S_k denote the time arm k being played for the first time. $1 = S_{I_1} < S_{I_2} < \cdots < S_{I_{K_n}}$. Since arms $I_1, \dots, I_{K_n}$ progressively enter in competition, Lemma 1 no longer holds but an easy adaptation of its proof shows that for $k \in \{I_1, \dots, I_{K_n}\}$,

$$\mathbb{E}(T_k(n) | I_1, \dots, I_{K_n}) \leq u + \textstyle\sum_{t=u+1}^{n} \sum_{s=u}^{t} \mathbb{P}\big(B_{k,s,t} > \tau\big) + \Omega_k \tag{15}$$

with

$$\Omega_k = \sum_{t=u+1}^{n} \prod_{k' \neq k\,, S_{k'} \leq t} \mathbb{P}(\exists s' \in [0,t],\ B_{k',s',t} \leq \tau)\,.$$

As in the proof of Theorem 1, since the exploration sequence satisfies $\mathcal{E}_t \geq 2\log(10\log t)$, we have $\mathbb{P}(\exists s' \in [0,t],\ B_{k',s',t} \leq \tau) \leq 1/2$ for arms k' such that $\mu_{k'} \geq \tau$. Consequently, letting $N_{\tau,k,t}$ denote the cardinal of the set $\{k' : k' \neq k, \mu_{k'} \geq \tau, S_{k'} \leq t\}$, we have

$$\Omega_k \leq \textstyle\sum_{t=1}^{n} 2^{-N_{\tau,k,t}}\,.$$

Let us first consider the case $\mu^* = 1$ or $\beta \geq 1$. In the case of UCB-AIR, S_{I_j} is the smallest integer strictly larger than $(j-1)^{(\beta+1)/\beta}$. To shorten notation, let us write S_j for S_{I_j}. According to the arm-increasing rule (try a new arm if $K_{t-1} < t^{\beta/(\beta+1)}$), $[S_j, S_{j+1})$ is the time interval in which the competing arms are $I_1, I_2, \dots, I_j$.

As in the proof of Theorem 1, we consider $\tau = \mu^* - \Delta_k/2$. We have

$$\begin{aligned} \mathbb{E}(\Omega_{I_\ell} | I_\ell = k) &\leq \textstyle\sum_{j=1}^{K_n} \sum_{t=S_j}^{S_{j+1}-1} \mathbb{E}\Big(2^{-N_{\tau,k,S_j}} | I_\ell = k\Big) \\ &= \textstyle\sum_{j=1}^{K_n} (S_{j+1} - S_j) \mathbb{E}\Big(2^{-N_{\tau,k,S_j}} | I_\ell = k\Big) \\ &\leq \textstyle\sum_{j=1}^{K_n} (S_{j+1} - S_j) \mathbb{E}\big(2^{-N_{\tau,\infty,S_{j-1}}}\big)\,. \end{aligned} \tag{16}$$

Since $N_{\tau,\infty,S_{j-1}}$ follows a binomial distribution with parameter $j-1$ and $\mathbb{P}(\mu \geq \tau)$, we have

$$\mathbb{E}\Big(2^{-N_{\tau,\infty,S_{j-1}}}\Big) = (1 - \mathbb{P}(\mu \geq \tau)/2)^{j-1},$$

and

$$\begin{aligned}\sum_{j=1}^{K_n}(S_{j+1}-S_j)\mathbb{E}\Big(2^{-N_{\tau,\infty,S_{j-1}}}\Big) &= \sum_{j=1}^{K_n}\big(S_{j+1}-S_j\big)(1-\mathbb{P}(\mu\geq\tau)/2)^{j-1} \\ &\leq \sum_{j=1}^{K_n}(1+\tfrac{\beta+1}{\beta}j^{\frac{1}{\beta}})(1-\tilde{c}[2(\mu^*-\tau)]^{\beta})^{j-1},\end{aligned} \tag{17}$$

where $\tilde{c}=c_1 2^{-1-\beta}$. Plugging (17) into (16), we obtain

$$\mathbb{E}(\Delta_{I_\ell}\Omega_{I_\ell}) \leq \tfrac{2\beta+1}{\beta}\sum_{j=1}^{K_n} j^{\frac{1}{\beta}}\mathbb{E}\big(\Delta_{I_\ell}\big[1-\tilde{c}\Delta_{I_\ell}^{\beta}\big]^{j-1}\big).$$

Now this last expectation can be bounded by the same computations as for $\mathbb{E}\chi(\Delta_1)$ in the proof of Theorem 1. We have, for appropriate positive constants C_1 and C_2 depending on c_1 and β,

$$\mathbb{E}(\Delta_{I_\ell}\Omega_{I_\ell}) \leq C_1\sum_{j=1}^{K_n} j^{\frac{1}{\beta}}j^{-\frac{1}{\beta}}\tfrac{\log j}{j} \leq C_2(\log K_n)^2\,. \tag{18}$$

Using (15) and $\mathbb{E}R_n=\sum_{\ell=1}^{K_n}\mathbb{E}(\Delta_{I_\ell}\Omega_{I_\ell})$, we obtain

$$\mathbb{E}R_n \leq K_n\mathbb{E}\bigg\{\Big[50\Big(\frac{V(\Delta_1)}{\Delta_1}+1\Big)\log n\Big]\wedge(n\Delta_1)+C_2(\log K_n)^2\bigg\},$$

from which Theorem 4 follows for the case $\mu^*=1$ or $\beta\geq 1$. For the case $\beta<1$ and $\mu^*<1$, replacing $\frac{\beta}{\beta+1}$ by $\frac{\beta}{2}$ leads to a similar version of (18) as

$$\mathbb{E}(\Delta_{I_\ell}\Omega_{I_\ell}) \leq C_1\sum_{j=1}^{K_n} j^{\frac{2}{\beta}-1}j^{-\frac{1}{\beta}}\tfrac{\log j}{j} \leq C_2(\log K_n)K_n^{\frac{1-\beta}{\beta}},$$

which gives the desired convergence rate since K_n is of order $n^{\beta/2}$.

References

[1] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.

[2] J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In M. Hutter, R. A. Servedio, and E. Takimoto, editors, *ALT*, volume 4754 of *Lecture Notes in Computer Science*, pages 150–165. Springer, 2007.

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

[4] P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. *20th COLT, San Diego, CA, USA*, 2007.

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

[6] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS-2004*, 2004.

[7] R. Kleinberg. *Online decision problems with large strategy sets*. PhD thesis, MIT, 2005.

[8] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandit problems in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.

[9] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[10] O. Teytaud, S. Gelly, and M. Sebag. Anytime many-armed bandit. *Conférence francophone sur l'Apprentissage automatique (CAp) Grenoble, France*, 2007.