Journal of Machine Learning Research 14 (2013) 187-207 Submitted 12/11; Revised 9/12; Published 1/13

Lower Bounds and Selectivity of Weak-Consistent Policies in Stochastic Multi-Armed Bandit Problem

Antoine Salomon SALOMONA@IMAGINE.ENPC.FR
Jean-Yves Audibert* AUDIBERT@IMAGINE.ENPC.FR
Issam El Alaoui ISSAM.EL-ALAOUI.2007@POLYTECHNIQUE.ORG
Imagine, Université Paris-Est
6 Avenue Blaise Pascal
77455 Champs-sur-Marne, France

Editor: Nicolo Cesa-Bianchi

Abstract

This paper is devoted to regret lower bounds in the classical model of stochastic multi-armed bandit. A well-known result of Lai and Robbins, which has then been extended by Burnetas and Katehakis, has established the presence of a logarithmic bound for all consistent policies. We relax the notion of consistency, and exhibit a generalisation of the bound. We also study the existence of logarithmic bounds in general and in the case of Hannan consistency. Moreover, we prove that it is impossible to design an adaptive policy that would select the best of two algorithms by taking advantage of the properties of the environment. To get these results, we study variants of popular Upper Confidence Bounds (UCB) policies.

Keywords: stochastic bandits, regret lower bounds, consistency, selectivity, UCB policies

1. Introduction and Notations

Multi-armed bandits are a classical way to illustrate the difficulty of decision making in the case of a dilemma between exploration and exploitation. The denomination of these models comes from an analogy with playing a slot machine with more than one arm. Each arm has a given (and unknown) reward distribution and, for a given number of rounds, the agent has to choose one of them. As the goal is to maximize the sum of rewards, each round decision consists in a trade-off between exploitation (i.e., playing the arm that has been the more lucrative so far) and exploration (i.e., testing another arm, hoping to discover an alternative that beats the current best choice). One possible application is clinical trial: a doctor wants to heal as many patients as possible, the patients arrive sequentially, and the effectiveness of each treatment is initially unknown (Thompson, 1933). Bandit problems have initially been studied by Robbins (1952), and since then they have been applied to many fields such as economics (Lamberton et al., 2004; Bergemann and Valimaki, 2008), games (Gelly and Wang, 2006), and optimisation (Kleinberg, 2005; Coquelin and Munos, 2007; Kleinberg et al., 2008; Bubeck et al., 2009).

∗. Also at Willow, CNRS/ENS/INRIA - UMR 8548.

1.1 Setting

In this paper, we consider the following model. A stochastic multi-armed bandit problem is defined by:

- a number of rounds n,
- a number of arms $K \geq 2$,
- an environment $\theta = (\nu_1, \cdots, \nu_K)$, where each ν_k ($k \in \{1, \cdots, K\}$) is a real-valued measure that represents the distribution reward of arm k.

The number of rounds n may or may not be known by the agent, but this will not affect the present study.

We assume that rewards are bounded. Thus, for simplicity, each ν_k is a probability on $[0,1]$. Environment θ is initially unknown by the agent but lies in some known set Θ. For the problem to be interesting, the agent should not have great knowledges of its environment, so that Θ should not be too small and/or only contain too trivial distributions such as Dirac measures. To make it simple, we may assume that Θ contains all environments where each reward distribution is a Dirac distribution or a Bernoulli distribution. This will be acknowledged as Θ having the *Dirac/Bernoulli property*. For technical reason, we may also assume that Θ is of the form $\Theta_1 \times \ldots \times \Theta_K$, meaning that Θ_k is the set of possible reward distributions of arm k. This will be acknowledged as Θ having the *product property*.

The game is as follows. At each round (or time step) $t = 1, \cdots, n$, the agent has to choose an arm I_t in the set $\{1, \cdots, K\}$. This decision is based on past actions and observations, and the agent may also randomize his choice. Once the decision is made, the agent gets and observes a reward that is drawn from ν_{I_t} independently from the past. Thus a policy (or strategy) can be described by a sequence $(\sigma_t)_{t \geq 1}$ (or $(\sigma_t)_{1 \leq t \leq n}$ if the number of rounds n is known) such that each σ_t is a mapping from the set $\{1, \ldots, K\}^{t-1} \times [0,1]^{t-1}$ of past decisions and rewards into the set of arm $\{1, \ldots, K\}$ (or into the set of probabilities on $\{1, \ldots, K\}$, in case the agent randomizes his choices).

For each arm k and all time step t, let $T_k(t) = \sum_{s=1}^{t} \mathbb{1}_{I_s = k}$ denote the sampling time, that is, the number of times arm k was pulled from round 1 to round t, and $X_{k,1}, X_{k,2}, \ldots, X_{k,T_k(t)}$ the corresponding sequence of rewards. We denote by $\mathbb{P}_\theta$ the distribution on the probability space such that for any $k \in \{1, \ldots, K\}$, the random variables $X_{k,1}, X_{k,2}, \ldots, X_{k,n}$ are i.i.d. realizations of ν_k, and such that these K sequences of random variables are independent. Let $\mathbb{E}_\theta$ denote the associated expectation.

Let $\mu_k = \int x d\nu_k(x)$ be the mean reward of arm k. Introduce $\mu^* = \max_{k \in \{1,\ldots,K\}} \mu_k$ and fix an arm $k^* \in \operatorname{argmax}_{k \in \{1,\ldots,K\}} \mu_k$, that is, k^* has the best expected reward. The agent aims at minimizing its *regret*, defined as the difference between the cumulative reward he would have obtained by always drawing the best arm and the cumulative reward he actually received. Its regret is thus

$$R_n = \sum_{t=1}^{n} X_{k^*,t} - \sum_{t=1}^{n} X_{I_t, T_{I_t}(t)}.$$

As most of the publications on this topic, we focus on the expected regret, for which one can check that:

$$\mathbb{E}_\theta R_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}_\theta[T_k(n)], \tag{1}$$

where Δ_k is the *optimality gap* of arm k, defined by $\Delta_k = \mu^* - \mu_k$. We also define Δ as the gap between the best arm and the second best arm, that is, $\Delta := \min_{k:\Delta_k>0} \Delta_k$.

Other notions of regret exist in the literature. One of them is the quantity

$$\max_k \sum_{t=1}^{n} X_{k,t} - X_{I_t,T_{I_t}(t)},$$

which is mostly used in adversarial settings. Results and ideas we want to convey here are more suited to expected regret, and considering other definitions of regret would only bring some more technical intricacies.

1.2 Consistency and Regret Lower Bounds

Former works have shown the existence of lower bounds on the expected regret of a large class of policies: intuitively, to perform well the agent has to explore all arms, and this requires a significant amount of suboptimal choices. In this way, Lai and Robbins (1985) proved a lower bound of order $\log n$ in a particular parametric framework, and they also exhibited optimal policies. This work has then been extended by Burnetas and Katehakis (1996). Both papers deal with *consistent* policies, meaning that they only consider policies such that:

$$\forall a > 0,\ \forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n^a). \tag{2}$$

Let us detail the bound of Burnetas and Katehakis, which is valid when Θ has the product property. Given an environment $\theta = (\nu_1, \cdots, \nu_K)$ and an arm $k \in \{1, \dots, K\}$, define:

$$D_k(\theta) := \inf_{\tilde{\nu}_k \in \Theta_k : \mathbb{E}[\tilde{\nu}_k] > \mu^*} KL(\nu_k, \tilde{\nu}_k),$$

where $KL(\nu, \mu)$ denotes the Kullback-Leibler divergence of measures ν and μ. Now fix a consistent policy and an environment $\theta \in \Theta$. If k is a suboptimal arm (i.e., $\mu_k \neq \mu^*$) such that $0 < D_k(\theta) < \infty$, then

$$\forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}\left[T_k(n) \geq \frac{(1-\varepsilon)\log n}{D_k(\theta)}\right] = 1.$$

This readily implies that:

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{\log n} \geq \frac{1}{D_k(\theta)}.$$

Thanks to Formula (1), it is then easy to deduce a lower bound of the expected regret.

One contribution of this paper is to generalize the study of regret lower bounds, by considering weaker notions of consistency: α-consistency and Hannan consistency. We will define α-consistency ($\alpha \in [0,1)$) as a variant of Equation (2), where equality $\mathbb{E}_\theta[R_n] = o(n^a)$ only holds for all $a > \alpha$. We show that the logarithmic bound of Burnetas and Katehakis still holds, but coefficient $\frac{1}{D_k(\theta)}$ is turned into $\frac{1-\alpha}{D_k(\theta)}$. We also prove that the dependence of this new bound with respect to the term $1-\alpha$ is asymptotically optimal when $n \to +\infty$ (up to a constant).

We will also consider the case of Hannan consistency. Indeed, any policy achieves at most an expected regret of order n: because of the equality $\sum_{k=1}^{K} T_k(n) = n$ and thanks to Equation (1), one can show that $\mathbb{E}_\theta R_n \leq n \max_k \Delta_k$. More intuitively, this comes from the fact that the average cost of pulling an arm k is a constant Δ_k. As a consequence, it is natural to wonder what happens when

dealing with policies whose expected regret is only required to be $o(n)$, which is equivalent to Hannan consistency. This condition is less restrictive than any of the previous notions of consistency. In this larger class of policy, we show that the lower bounds on the expected regret are no longer logarithmic, but can be much smaller.

Finally, even if no logarithmic lower bound holds on the whole set Θ, we show that there necessarily exist some environments θ for which the expected regret is at least logarithmic. The latter result is actually valid without any assumptions on the considered policies, and only requires a simple property on Θ.

1.3 Selectivity

As we exhibit new lower bounds, we want to know if it is possible to provide optimal policies that achieve these lower bounds, as it is the case in the classical class of consistent policies. We answer negatively to this question, and for this we solve the more general problem of selectivity. Given a set of policies, we define selectivity as the ability to perform at least as good as the policy that is best suited to the current environment θ. Still, such an ability can not be implemented. As a by-product it is not possible to design a procedure that would specifically adapt to some kinds of environments, for example by selecting a particular policy. This contribution is linked with selectivity in on-line learning problem with perfect information, commonly addressed by prediction with expert advice (see, e.g., Cesa-Bianchi et al., 1997). In this spirit, a closely related problem is the one of regret against the best strategy from a pool studied by Auer et al. (2003). The latter designed an algorithm in the context of adversarial/nonstochastic bandit whose decisions are based on a given number of recommendations (experts), which are themselves possibly the rewards received by a set of given policies. To a larger extent, model selection has been intensively studied in statistics, and is commonly solved by penalization methods (Mallows, 1973; Akaike, 1973; Schwarz, 1978).

1.4 UCB Policies

Some of our results are obtained using particular Upper Confidence Bound algorithms. These algorithms were introduced by Lai and Robbins (1985): they basically consist in computing an index for each arm, and selecting the arm with the greatest index. A simple and efficient way to design such policies is as follows: choose each index as low as possible such that, conditional to past observations, it is an upper bound of the mean reward of the considered arm with high probability (or, say, with high confidence level). This idea can be traced back to Agrawal (1995), and has been popularized by Auer et al. (2002), who notably described a policy called UCB1. In this policy, each index $B_{k,s,t}$ is defined by an arm k, a time step t, an integer s that indicates the number of times arm k has been pulled before round t, and is given by:

$$B_{k,s,t} = \hat{X}_{k,s} + \sqrt{\frac{2\log t}{s}},$$

where $\hat{X}_{k,s}$ is the empirical mean of arm k after s pulls, that is, $\hat{X}_{k,s} = \frac{1}{s}\sum_{u=1}^{s} X_{k,u}$.

To summarize, UCB1 policy first pulls each arm once and then, at each round $t > K$, selects an arm k that maximizes $B_{k,T_k(t-1),t}$. Note that, by means of Hoeffding's inequality, the index $B_{k,T_k(t-1),t}$ is indeed an upper bound of μ_k with high probability (i.e., the probability is greater than $1 - 1/t^4$). Another way to understand this index is to interpret the empiric mean $\hat{X}_{k,T_k(t-1)}$ as an "exploitation"

term, and the square root $\sqrt{2\log t/s}$ as an "exploration" term (because the latter gradually increases when arm k is not selected).

Policy UCB1 achieves the logarithmic bound (up to a multiplicative constant), as it was shown that:

$$\forall \theta \in \Theta,\ \forall n \geq 3,\ \ \mathbb{E}_\theta[T_k(n)] \leq 12\frac{\log n}{\Delta_k^2} \ \text{ and } \ \mathbb{E}_\theta R_n \leq 12\sum_{k=1}^{K}\frac{\log n}{\Delta_k} \leq 12K\frac{\log n}{\Delta}.$$

Audibert et al. (2009) studied some variants of UCB1 policy. Among them, one consists in changing the $2\log t$ in the exploration term into $\rho\log t$, where $\rho > 0$. This can be interpreted as a way to tune exploration: the smaller ρ is, the better the policy will perform in simple environments where information is disclosed easily (for example when all reward distributions are Dirac measures). On the contrary, ρ has to be greater to face more challenging environments (typically when reward distributions are Bernoulli laws with close parameters).

This policy, that we denote UCB(ρ), was proven by Audibert et al. to achieve the logarithmic bound when $\rho > 1$, and the optimality was also obtained when $\rho > \frac{1}{2}$ for a variant of UCB(ρ). Bubeck (2010) showed in his PhD dissertation that their ideas actually enable to prove optimality of UCB(ρ) for $\rho > \frac{1}{2}$. Moreover, the case $\rho = \frac{1}{2}$ corresponds to a confidence level of $\frac{1}{t}$ (because of Hoeffding's inequality, as above), and several studies (Lai and Robbins, 1985; Agrawal, 1995; Burnetas and Katehakis, 1996; Audibert et al., 2009; Honda and Takemura, 2010) have shown that this level is critical.

We complete these works by a precise study of UCB(ρ) when $\rho \leq \frac{1}{2}$. We prove that UCB(ρ) is $(1-2\rho)$-consistent and that it is not α-consistent for any $\alpha < 1-2\rho$ (in view of the definition above, this means that the expected regret is roughly of order $n^{1-2\rho}$). Thus it does not achieve the logarithmic bound, but it performs well in simple environments, for example, environments where all reward distributions are Dirac measures.

Moreover, we exhibit expected regret bounds of general UCB policies, with the $2\log t$ in the exploration term of UCB1 replaced by an arbitrary function. We give sufficient conditions for such policies to be Hannan consistent and, as mentioned before, find that lower bounds need not be logarithmic any more.

1.5 Outline

The paper is organized as follows: in Section 2, we give bounds on the expected regret of general UCB policies and of UCB(ρ) ($\rho \leq \frac{1}{2}$), as preliminary results. In Section 3, we focus on α-consistent policies. Then, in Section 4, we study the problem of selectivity, and we conclude in Section 5 by general results on the existence of logarithmic lower bounds.

Throughout the paper $\lceil x \rceil$ denotes the smallest integer not less than x whereas $\lfloor x \rfloor$ denotes the largest integer not greater than x, $\mathbb{1}_A$ stands for the indicator function of event A, $Ber(p)$ is the Bernoulli law with parameter p, and δ_x is the Dirac measure centred on x.

2. Preliminary Results

In this section, we estimate the expected regret of UCB policies. This will be useful for the rest of the paper.

2.1 Bounds on the Expected Regret of General UCB Policies

We first study general UCB policies, defined by:

- Draw each arm once,
- then, at each round t, draw an arm

$$I_t \in \operatorname*{argmax}_{k\in\{1,\dots,K\}} B_{k,T_k(t-1),t},$$

where $B_{k,s,t}$ is defined by $B_{k,s,t} = \hat{X}_{k,s} + \sqrt{\frac{f_k(t)}{s}}$ and where functions f_k ($1 \le k \le K$) are increasing.

This definition is inspired by popular UCB1 algorithm, for which $f_k(t) = 2\log t$ for all k.

The following lemma estimates the performances of UCB policies in simple environments, for which reward distributions are Dirac measures.

Lemma 1 *Let $0 \le b < a \le 1$ and $n \ge 1$. For $\theta = (\delta_a, \delta_b)$, the random variable $T_2(n)$ is uniformly upper bounded by $\frac{1}{\Delta^2} f_2(n) + 1$. Consequently, the expected regret of* UCB *is upper bounded by $\frac{1}{\Delta} f_2(n) + 1$.*

Proof In environment θ, best arm is arm 1 and $\Delta = \Delta_2 = a - b$. Let us first prove the upper bound of the sampling time. The assertion is true for $n = 1$ and $n = 2$: the first two rounds consists in drawing each arm once, so that $T_2(n) \le 1 \le \frac{1}{\Delta^2} f_2(n) + 1$ for $n \in \{1,2\}$. If, by contradiction, the assertion is false, then there exists $t \ge 3$ such that $T_2(t) > \frac{1}{\Delta^2} f_2(t) + 1$ and $T_2(t-1) \le \frac{1}{\Delta^2} f_2(t-1) + 1$. Since $f_2(t) \ge f_2(t-1)$, this leads to $T_2(t) > T_2(t-1)$, meaning that arm 2 is drawn at round t. Therefore, we have $a + \sqrt{\frac{f_1(t)}{T_1(t-1)}} \le b + \sqrt{\frac{f_2(t)}{T_2(t-1)}}$, hence $a - b = \Delta \le \sqrt{\frac{f_2(t)}{T_2(t-1)}}$, which implies $T_2(t-1) \le \frac{1}{\Delta^2} f_2(t)$ and thus $T_2(t) \le \frac{1}{\Delta^2} f_2(t) + 1$. This contradicts the definition of t, and this ends the proof of the first statement. The second statement is a direct consequence of Formula (1). ■

Remark: throughout the paper, we will often use environments with $K = 2$ arms to provide bounds on expected regrets. However, we do not lose generality by doing so, because all corresponding proofs can be written almost identically to suit to any $K \ge 2$, by simply assuming that the distribution of each arm $k \ge 3$ is δ_0.

We now give an upper bound of the expected sampling time of any arm such that $\Delta_k > 0$. This bound is valid in any environment, and not only those of the form (δ_a, δ_b).

Lemma 2 *For any $\theta \in \Theta$ and any $\beta \in (0,1)$, if $\Delta_k > 0$ the following upper bound holds:*

$$\mathbb{E}_\theta[T_k(n)] \le u + \sum_{t=u+1}^{n} \left(1 + \frac{\log t}{\log(\frac{1}{\beta})}\right)\left(e^{-2\beta f_k(t)} + e^{-2\beta f_{k^*}(t)}\right),$$

where $u = \left\lceil \frac{4 f_k(n)}{\Delta_k^2} \right\rceil$.

An upper bound of the expected regret can be deduced from this lemma thanks to Formula 1.

Proof The core of the proof is a peeling argument and the use of Hoeffding's maximal inequality (see, e.g., Cesa-Bianchi and Lugosi, 2006, section A.1.3 for details). The idea is originally taken from Audibert et al. (2009), and the following is an adaptation of the proof of an upper bound of UCB(ρ) in the case $\rho > \frac{1}{2}$ which can be found in S. Bubeck's PhD dissertation.

First, let us notice that the policy selects an arm k such that $\Delta_k > 0$ at time step $t \leq n$ only if at least one of the three following equations holds:

$$B_{k^*,T_{k^*}(t-1),t} \leq \mu^*, \tag{3}$$

$$\hat{X}_{k,t} \geq \mu_k + \sqrt{\frac{f_k(t)}{T_k(t-1)}}, \tag{4}$$

$$T_k(t-1) < \frac{4f_k(n)}{\Delta_k^2}. \tag{5}$$

Indeed, if none of the equations is true, then:

$$B_{k^*,T_{k^*}(t-1),t} > \mu^* = \mu_k + \Delta_k \geq \mu_k + 2\sqrt{\frac{f_k(n)}{T_k(t-1)}} > \hat{X}_{k,t} + \sqrt{\frac{f_k(t)}{T_k(t-1)}} = B_{k,T_k(t-1),t},$$

which implies that arm k can not be chosen at time step t.

We denote respectively by $\xi_{1,t}, \xi_{2,t}$ and $\xi_{3,t}$ the events corresponding to Equations (3), (4) and (5).

We have:

$$\mathbb{E}_\theta[T_k(n)] \quad = \quad \mathbb{E}_\theta\left[\sum_{t=1}^n \mathbb{1}_{I_t=k}\right] = \mathbb{E}_\theta\left[\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}}\right] + \mathbb{E}_\theta\left[\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\setminus\xi_{3,t}}\right].$$

Let us show that the sum $\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}}$ is almost surely lower than $u := \lceil 4f_k(n)/\Delta_k^2 \rceil$. We assume by contradiction that $\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} > u$. Then there exists $m < n$ such that $\sum_{t=1}^{m-1} \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} < 4f_k(n)/\Delta_k^2$ and $\sum_{t=1}^m \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} = \lceil 4f_k(n)/\Delta_k^2 \rceil$. Therefore, for any $s > m$, we have:

$$T_k(s-1) \geq T_k(m) = \sum_{t=1}^m \mathbb{1}_{\{I_t=k\}} \geq \sum_{t=1}^m \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} = \left\lceil \frac{4f_k(n)}{\Delta_k^2} \right\rceil \geq \frac{4f_k(n)}{\Delta_k^2},$$

so that $\mathbb{1}_{\{I_s=k\}\cap\xi_{3,s}} = 0$. But then

$$\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} = \sum_{t=1}^m \mathbb{1}_{\{I_t=k\}\cap\xi_{3,t}} = \left\lceil \frac{4f_k(n)}{\Delta_k^2} \right\rceil \leq u,$$

which is the contradiction expected.

We also have $\sum_{t=1}^n \mathbb{1}_{\{I_t=k\}\setminus\xi_{3,t}} = \sum_{t=u+1}^n \mathbb{1}_{\{I_t=k\}\setminus\xi_{3,t}}$: since $T_k(t-1) \leq t-1$, event $\xi_{3,t}$ always happens at time step $t \in \{1,\ldots,u\}$.

And then, since event $\{I_t = k\}$ is included in $\xi_{1,t} \cup \xi_{2,t} \cup \xi_{3,t}$:

$$\mathbb{E}_\theta\left[\sum_{t=u+1}^n \mathbb{1}_{\{I_t=k\}\setminus\xi_{3,t}}\right] \leq \mathbb{E}_\theta\left[\sum_{t=u+1}^n \mathbb{1}_{\xi_{1,t}\cup\xi_{2,t}}\right] \leq \sum_{t=u+1}^n \mathbb{P}_\theta(\xi_{1,t}) + \mathbb{P}_\theta(\xi_{2,t}).$$

It remains to find upper bounds of $\mathbb{P}_\theta(\xi_{1,t})$ and $\mathbb{P}_\theta(\xi_{2,t})$. To this aim, we apply the peeling argument with a geometric grid over the time interval $[1,t]$:

$$\begin{aligned}
\mathbb{P}_\theta(\xi_{1,t}) &= \mathbb{P}_\theta\left(B_{k^*,T_{k^*}(t-1),t} \le \mu^*\right) = \mathbb{P}_\theta\left(\hat{X}_{k^*,T_{k^*}(t-1)} + \sqrt{\frac{f_{k^*}(t)}{T_{k^*}(t-1)}} \le \mu^*\right) \\
&\le \mathbb{P}_\theta\left(\exists s \in \{1,\cdots,t\},\, \hat{X}_{k^*,s} + \sqrt{\frac{f_{k^*}(t)}{s}} \le \mu^*\right) \\
&\le \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \mathbb{P}_\theta\left(\exists s : \{\beta^{j+1}t < s \le \beta^j t\},\, \hat{X}_{k^*,s} + \sqrt{\frac{f_{k^*}(t)}{s}} \le \mu^*\right) \\
&\le \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \mathbb{P}_\theta\left(\exists s : \{\beta^{j+1}t < s \le \beta^j t\},\, \sum_{l=1}^{s}(X_{k^*,l} - \mu^*) \le -\sqrt{s f_{k^*}(t)}\right) \\
&\le \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \mathbb{P}_\theta\left(\exists s : \{\beta^{j+1}t < s \le \beta^j t\},\, \sum_{l=1}^{s}(X_{k^*,l} - \mu^*) \le -\sqrt{\beta^{j+1} t f_{k^*}(t)}\right) \\
&= \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \mathbb{P}_\theta\left(\max_{\beta^{j+1}t < s \le \beta^j t} \sum_{l=1}^{s}(\mu^* - X_{k^*,l}) \ge \sqrt{\beta^{j+1} t f_{k^*}(t)}\right) \\
&\le \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \mathbb{P}_\theta\left(\max_{s \le \beta^j t} \sum_{l=1}^{s}(\mu^* - X_{k^*,l}) \ge \sqrt{\beta^{j+1} t f_{k^*}(t)}\right).
\end{aligned}$$

As the range of the random variables $(X_{k^*,l})_{1\le l\le s}$ is $[0,1]$, Hoeffding's maximal inequality gives:

$$\mathbb{P}_\theta(\xi_{1,t}) \le \sum_{j=0}^{\left\lfloor \frac{\log t}{\log(1/\beta)} \right\rfloor} \exp\left(-\frac{2\left(\sqrt{\beta^{j+1} t f_{k^*}(t)}\right)^2}{\beta^j t}\right) \le \left(\frac{\log t}{\log(1/\beta)} + 1\right) e^{-2\beta f_{k^*}(t)}.$$

Similarly, we have:

$$\mathbb{P}_\theta(\xi_{2,t}) \le \left(\frac{\log t}{\log(1/\beta)} + 1\right) e^{-2\beta f_k(t)},$$

and the result follows from the combination of previous inequalities. ■

2.2 Bounds on the Expected Regret of UCB(ρ), $\rho \le \frac{1}{2}$

We study the performances of UCB(ρ) policy, with $\rho \in (0, \frac{1}{2}]$. We recall that UCB($\rho$) is the UCB policy defined by $f_k(t) = \rho \log(t)$ for all k, that is, $B_{k,s,t} = \hat{X}_{k,s} + \sqrt{\frac{\rho \log t}{s}}$. Small values of ρ can be interpreted as a low level of experimentation in the balance between exploration and exploitation. Precise regret bound orders of UCB(ρ) when $\rho \in (0, \frac{1}{2}]$ are not documented in the literature.

We first give an upper bound of expected regret in simple environments, where it is supposed to perform well. As stated in the following proposition (which is a direct consequence of Lemma 1), the order of the bound is $\frac{\rho \log n}{\Delta}$.

Lemma 3 *Let $0 \leq b < a \leq 1$ and $n \geq 1$. For $\theta = (\delta_a, \delta_b)$, the random variable $T_2(n)$ is uniformly upper bounded by $\frac{\rho}{\Delta^2}\log(n) + 1$. Consequently, the expected regret of* UCB(ρ) *is upper bounded by $\frac{\rho}{\Delta}\log(n) + 1$.*

One can show that the expected regret of UCB(ρ) is actually equivalent to $\frac{\rho \log n}{\Delta}$ as n goes to infinity. These good performances are compensated by poor results in more complex environments, as showed in the following theorem. We exhibit an expected regret upper bound which is valid for any $\theta \in \Theta$, and which is roughly of order $n^{1-2\rho}$. We also show that this upper bound is asymptotically optimal. Thus, with $\rho \in (0, \frac{1}{2})$, UCB($\rho$) does not perform enough exploration to achieve the logarithmic bound, as opposed to UCB(ρ) with $\rho \in (\frac{1}{2}, +\infty)$.

Theorem 4 *For any $\rho \in (0, \frac{1}{2}]$, any $\theta \in \Theta$ and any $\beta \in (0,1)$, one has*

$$\mathbb{E}_\theta[R_n] \leq \sum_{k:\Delta_k > 0} \frac{4\rho \log n}{\Delta_k} + \Delta_k + 2\Delta_k \left(\frac{\log n}{\log(1/\beta)} + 1 \right) \frac{n^{1-2\rho\beta}}{1-2\rho\beta}.$$

Moreover, if Θ has the Dirac/Bernoulli property, then for any $\varepsilon > 0$ there exists $\theta \in \Theta$ such that

$$\lim_{n \to +\infty} \frac{\mathbb{E}_\theta[R_n]}{n^{1-2\rho-\varepsilon}} = +\infty.$$

The value $\rho = \frac{1}{2}$ is critical, but we can deduce from the upper bound of this theorem that UCB($\frac{1}{2}$) is consistent in the classical sense of Lai and Robbins (1985) and of Burnetas and Katehakis (1996).

Proof We set $u = \left\lceil \frac{4\rho \log n}{\Delta_k^2} \right\rceil$. By Lemma 2 we get:

$$\begin{aligned}
\mathbb{E}_\theta[T_k(n)] &\leq u + 2\sum_{t=u+1}^{n} \left(\frac{\log t}{\log(1/\beta)} + 1 \right) e^{-2\beta\rho\log(t)} \\
&= u + 2\sum_{t=u+1}^{n} \left(\frac{\log t}{\log(1/\beta)} + 1 \right) \frac{1}{t^{2\rho\beta}} \\
&\leq u + 2\left(\frac{\log n}{\log(1/\beta)} + 1 \right) \sum_{t=1}^{n} \frac{1}{t^{2\rho\beta}} \\
&\leq u + 2\left(\frac{\log n}{\log(1/\beta)} + 1 \right) \left(1 + \sum_{t=2}^{n} \frac{1}{t^{2\rho\beta}} \right) \\
&\leq u + 2\left(\frac{\log n}{\log(1/\beta)} + 1 \right) \left(1 + \int_1^{n-1} \frac{1}{t^{2\rho\beta}} dt \right) \\
&\leq u + 2\left(\frac{\log n}{\log(1/\beta)} + 1 \right) \frac{n^{1-2\rho\beta}}{1-2\rho\beta}.
\end{aligned}$$

As usual, the upper bound of the expected regret follows from Formula (1).

Now, let us show the lower bound. The result is obtained by considering an environment θ of the form $\left(Ber(\frac{1}{2}), \delta_{\frac{1}{2}-\Delta} \right)$, where Δ lies in $(0, \frac{1}{2})$ and is such that $2\rho(1+\sqrt{\Delta})^2 < 2\rho + \varepsilon$. This notation is obviously consistent with the definition of Δ as an optimality gap. We set $T_n := \lceil \frac{\rho \log n}{\Delta} \rceil$, and define the event ξ_n by:

$$\xi_n = \left\{ \hat{X}_{1,T_n} < \frac{1}{2} - (1 + \frac{1}{\sqrt{\Delta}})\Delta \right\}.$$

When event ξ_n occurs, one has for any $t \in \{T_n, \dots, n\}$

$$\begin{aligned} \hat{X}_{1,T_n} + \sqrt{\frac{\rho \log t}{T_n}} &\leq \hat{X}_{1,T_n} + \sqrt{\frac{\rho \log n}{T_n}} < \frac{1}{2} - (1 + \frac{1}{\sqrt{\Delta}})\Delta + \sqrt{\Delta} \\ &\leq \frac{1}{2} - \Delta, \end{aligned}$$

so that arm 1 is chosen no more than T_n times by UCB(ρ) policy. Consequently:

$$\mathbb{E}_\theta [T_2(n)] \geq \mathbb{P}_\theta(\xi_n)(n - T_n).$$

We will now find a lower bound of the probability of ξ_n thanks to Berry-Esseen inequality. We denote by C the corresponding constant, and by Φ the c.d.f. of the standard normal distribution. For convenience, we also define the following quantities:

$$\sigma := \sqrt{\mathbb{E}\left[\left(X_{1,1} - \frac{1}{2}\right)^2\right]} = \frac{1}{2},\ M_3 := \mathbb{E}\left[\left|X_{1,1} - \frac{1}{2}\right|^3\right] = \frac{1}{8}.$$

Using the fact that $\Phi(-x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x}\beta(x)$ with $\beta(x) \xrightarrow[x \to +\infty]{} 1$, we have:

$$\begin{aligned} \mathbb{P}_\theta(\xi_n) &= \mathbb{P}_\theta\left(\frac{\hat{X}_{1,T_n} - \frac{1}{2}}{\sigma}\sqrt{T_n} \leq -2\left(1 + \frac{1}{\sqrt{\Delta}}\right)\Delta\sqrt{T_n}\right) \\ &\geq \Phi\left(-2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}} \\ &\geq \frac{\exp\left(-2(\Delta + \sqrt{\Delta})^2 T_n\right)}{2\sqrt{2\pi}(\Delta + \sqrt{\Delta})\sqrt{T_n}}\beta\left(2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}} \\ &\geq \frac{\exp\left(-2(\Delta + \sqrt{\Delta})^2(\frac{\rho \log n}{\Delta} + 1)\right)}{2\sqrt{2\pi}(\Delta + \sqrt{\Delta})\sqrt{T_n}}\beta\left(2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}} \\ &\geq \frac{n^{-2\rho(1+\sqrt{\Delta})^2}}{\sqrt{T_n}} \frac{\exp\left(-2(\Delta + \sqrt{\Delta})^2\right)}{2\sqrt{2\pi}(\Delta + \sqrt{\Delta})}\beta\left(2(\Delta + \sqrt{\Delta})\sqrt{T_n}\right) - \frac{CM_3}{\sigma^3\sqrt{T_n}}. \end{aligned}$$

Previous calculations and Formula (1) give

$$\mathbb{E}_\theta[R_n] = \Delta\mathbb{E}_\theta[T_2(n)] \geq \Delta\mathbb{P}_\theta(\xi_n)(n - T_n),$$

so that we finally obtain a lower bound of $\mathbb{E}_\theta[R_n]$ of order $\frac{n^{1-2\rho(1+\sqrt{\Delta})^2}}{\sqrt{\log n}}$. Therefore, $\frac{\mathbb{E}_\theta[R_n]}{n^{1-2\rho-\varepsilon}}$ is at least of order $\frac{n^{2\rho+\varepsilon-2\rho(1+\sqrt{\Delta})^2}}{\sqrt{\log n}}$. Since $2\rho + \varepsilon - 2\rho(1 + \sqrt{\Delta})^2 > 0$, the numerator goes to infinity, faster than $\sqrt{\log n}$. This concludes the proof. ■

3. Bounds on the Class α-consistent Policies

In this section, our aim is to find how the classical results of Lai and Robbins (1985) and of Burnetas and Katehakis (1996) can be generalised if we do not restrict the study to consistent policies. As a by-product, we will adapt their results to the present setting, which is simpler than their parametric frameworks.

We recall that a policy is consistent if its expected regret is $o(n^a)$ for all $a > 0$ in all environments $\theta \in \Theta$. A natural way to relax this definition is the following.

Definition 5 *A policy is α-consistent if*

$$\forall a > \alpha,\ \forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n^a).$$

For example, we showed in the previous section that, for any $\rho \in (0, \frac{1}{2}]$, UCB($\rho$) is $(1-2\rho)$-consistent and not α-consistent if $\alpha < 1-2\rho$.

Note that the relevant range of α in this definition is $[0,1)$: the case $\alpha = 0$ corresponds to the standard definition of consistency (so that throughout the paper the term "consistent" also means "0-consistent"), and any value $\alpha \geq 1$ is pointless as any policy is then α-consistent. Indeed, the expected regret of any policy is at most of order n. This also lead us to wonder what happens if we only require the expected regret to be $o(n)$:

$$\forall \theta \in \Theta,\ \mathbb{E}_\theta[R_n] = o(n).$$

This requirement corresponds to the definition of Hannan consistency. The class of Hannan consistent policies includes consistent policies and α-consistent policies for any $\alpha \in [0,1)$. Some results about this class will be obtained in Section 5.

We focus on regret lower bounds on α-consistent policies. We first show that the main result of Burnetas and Katehakis can be extended in the following way.

Theorem 6 *Assume that Θ has the product property. Fix an α-consistent policy and $\theta \in \Theta$. If $\Delta_k > 0$ and if $0 < D_k(\theta) < \infty$, then*

$$\forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}_\theta \left[T_k(n) \geq (1-\varepsilon) \frac{(1-\alpha)\log n}{D_k(\theta)} \right] = 1.$$

Consequently

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{\log n} \geq \frac{1-\alpha}{D_k(\theta)}.$$

Remind that the lower bound of the expected regret is then deduced from Formula (1), and that coefficient $D_k(\theta)$ is defined by:

$$D_k(\theta) := \inf_{\tilde{\nu}_k \in \Theta_k : \mathbb{E}[\tilde{\nu}_k] > \mu^*} KL(\nu_k, \tilde{\nu}_k),$$

where $KL(\nu,\mu)$ denotes the Kullback-Leibler divergence of measures ν and μ.

Note that, as opposed to Burnetas and Katehakis (1996), there is no optimal policy in general (i.e., a policy that would achieve the lower bound in all environment θ). This can be explained intuitively as follows. If by contradiction there existed such a policy, its expected regret would be of order $\log n$ and consequently it would be (0-)consistent. Then the lower bounds in the case of

0-consistency would necessarily hold. This can not happen if $\alpha > 0$ because $\frac{1-\alpha}{D_k(\theta)} < \frac{1}{D_k(\theta)}$.
Nevertheless, this argument is not rigorous because the fact that the regret would be of order $\log n$ is only valid for environments θ such that $0 < D_k(\theta) < \infty$. The non-existence of optimal policies is implied by a stronger result of the next section (yet, only if $\alpha > 0.2$).

Proof We adapt Proposition 1 in Burnetas and Katehakis (1996) and its proof. Let us denote $\theta = (\nu_1, \dots, \nu_K)$. We fix $\varepsilon > 0$, and we want to show that:

$$\lim_{n\to+\infty} \mathbb{P}_\theta \left(\frac{T_k(n)}{\log n} < \frac{(1-\varepsilon)(1-\alpha)}{D_k(\theta)} \right) = 0.$$

Set $\delta > 0$ and $\delta' > \alpha$ such that $\frac{1-\delta'}{1+\delta} > (1-\varepsilon)(1-\alpha)$. By definition of $D_k(\theta)$, there exists $\tilde{\nu}_k$ such that $\mathbb{E}[\tilde{\nu}_k] > \mu^*$ and

$$D_k(\theta) < KL(\nu_k, \tilde{\nu}_k) < (1+\delta) D_k(\theta).$$

Let us set $\tilde{\theta} = (\nu_1, \dots, \nu_{k-1}, \tilde{\nu}_k, \nu_{k+1}, \dots, \nu_K)$. Environment $\tilde{\theta}$ lies in Θ by the product property and arm k is its best arm. Define $I^\delta = KL(\nu_k, \tilde{\nu}_k)$ and

$$A_n^{\delta'} := \left\{ \frac{T_k(n)}{\log n} < \frac{1-\delta'}{I^\delta} \right\}, \; C_n^{\delta''} := \left\{ \log L_{T_k(n)} \le \left(1-\delta''\right) \log n \right\},$$

where δ'' is such that $\alpha < \delta'' < \delta'$ and L_t is defined by $\log L_t = \sum_{s=1}^t \log\left(\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,s}) \right)$.

Now, we show that $\mathbb{P}_\theta(A_n^{\delta'}) = \mathbb{P}_\theta(A_n^{\delta'} \cap C_n^{\delta''}) + \mathbb{P}_\theta(A_n^{\delta'} \setminus C_n^{\delta''}) \xrightarrow[n\to+\infty]{} 0$.

On the one hand, one has:

$$\begin{aligned}
\mathbb{P}_\theta(A_n^{\delta'} \cap C_n^{\delta''}) &\le n^{1-\delta''} \mathbb{P}_{\tilde{\theta}}(A_n^{\delta'} \cap C_n^{\delta''}) && (6)\\
&\le n^{1-\delta''} \mathbb{P}_{\tilde{\theta}}(A_n^{\delta'}) = n^{1-\delta''} \mathbb{P}_{\tilde{\theta}} \left(n - T_k(n) > n - \frac{1-\delta'}{I^\delta} \log n \right) \\
&\le \frac{n^{1-\delta''} \mathbb{E}_{\tilde{\theta}}[n - T_k(n)]}{n - \frac{1-\delta'}{I^\delta} \log n} && (7)\\
&= \frac{n^{-\delta''} \mathbb{E}_{\tilde{\theta}} \left[\sum_{l=1}^K T_\ell(n) - T_k(n) \right]}{n - \frac{1-\delta'}{I^\delta} \frac{\log n}{n}} \\
&\le \frac{\sum_{\ell \neq k} n^{-\delta''} \mathbb{E}_{\tilde{\theta}}[T_\ell(n)]}{1 - \frac{1-\delta'}{I^\delta} \frac{\log n}{n}} \xrightarrow[n\to+\infty]{} 0. && (8)
\end{aligned}$$

Equation (6) results from a partition of $A_n^{\delta'}$ into events $\{T_k(n) = a\}$, $0 \le a < \left\lceil \frac{1-\delta'}{I^\delta} \log n \right\rceil$. Each event $\{T_k(n) = a\} \cap C_n^{\delta''}$ equals $\{T_k(n) = a\} \cap \left\{ \prod_{s=1}^a \frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,s}) \le n^{1-\delta''} \right\}$ and is measurable with respect to $X_{k,1}, \dots, X_{k,a}$ and to $X_{\ell,1}, \dots, X_{\ell,n}$ ($\ell \neq k$). Thus, $\mathbb{1}_{\{T_k(n)=a\} \cap C_n^{\delta''}}$ can be written as a function f of the latter r.v. and we have:

$$\begin{aligned}
\mathbb{P}_\theta \left(\{T_k(n) = a\} \cap C_n^{\delta''} \right) &= \int f\left((x_{k,s})_{1\le s\le a}, (x_{\ell,s})_{\ell\neq k, 1\le s\le n} \right) \prod_{\substack{\ell \neq k \\ 1 \le s \le n}} d\nu_\ell(x_{\ell,s}) \prod_{1\le s\le a} d\nu_k(x_{k,s}) \\
&\le \int f\left((x_{k,s})_{1\le s\le a}, (x_{\ell,s})_{\ell\neq k, 1\le s\le n} \right) \prod_{\substack{\ell \neq k \\ 1 \le s \le n}} d\nu_\ell(x_{\ell,s}) n^{1-\delta''} \prod_{1\le s\le a} d\tilde{\nu}_k(x_{k,s}) \\
&= n^{1-\delta''} \mathbb{P}_{\tilde{\theta}} \left(\{T_k(n) = a\} \cap C_n^{\delta''} \right).
\end{aligned}$$

Equation (7) is a consequence of Markov's inequality, and the limit in (8) is a consequence of α-consistency.

On the other hand, we set $b_n := \frac{1-\delta'}{I^\delta}\log n$, so that

$$\begin{aligned}\mathbb{P}_\theta(A_n^{\delta'} \setminus C_n^{\delta''}) &\leq \mathbb{P}\left(\max_{j\leq\lfloor b_n\rfloor} \log L_j > (1-\delta'')\log n\right)\\ &\leq \mathbb{P}\left(\frac{1}{b_n}\max_{j\leq\lfloor b_n\rfloor} \log L_j > I^\delta \frac{1-\delta''}{1-\delta'}\right).\end{aligned}$$

This term tends to zero, as a consequence of the law of large numbers.

Now that $\mathbb{P}_\theta(A_n^{\delta'})$ tends to zero, the conclusion results from

$$\frac{1-\delta'}{I^\delta} > \frac{1-\delta'}{(1+\delta)D_k(\theta)} \geq \frac{(1-\varepsilon)(1-\alpha)}{D_k(\theta)}.$$

■

The previous lower bound is asymptotically optimal with respect to its dependence in α, as claimed in the following proposition.

Proposition 7 *Assume that Θ has the Dirac/Bernoulli property. There exist $\theta \in \Theta$ and a constant $c > 0$ such that, for any $\alpha \in [0,1)$, there exists an α-consistent policy such that:*

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_k(n)]}{(1-\alpha)\log n} \leq c,$$

for any k satisfying $\Delta_k > 0$.

Proof In any environment of the form $\theta = (\delta_a, \delta_b)$ with $a \neq b$, Lemma 3 implies the following estimate for UCB(ρ):

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta T_k(n)}{\log n} \leq \frac{\rho}{\Delta^2},$$

where $k \neq k^*$.

Because $\frac{1-\alpha}{2} \in (0, \frac{1}{2})$ and since UCB(ρ) is $(1-2\rho)$-consistent for any $\rho \in (0, \frac{1}{2}]$ (Theorem 4), we obtain the result by choosing the α-consistent policy UCB($\frac{1-\alpha}{2}$) and by setting $c = \frac{1}{2\Delta^2}$.

■

4. Selectivity

In this section, we address the problem of selectivity. By selectivity, we mean the ability to adapt to the environment as and when rewards are observed. More precisely, a set of two (or more) policies is given. The one that performs the best depends on environment θ. We wonder if there exists an adaptive procedure that, given any environment θ, would be as good as any policy in the given set. Two major reasons motivate this study.

On the one hand this question was answered by Burnetas and Katehakis within the class of consistent policies. They exhibits an asymptotically optimal policy, that is, that achieves the regret

lower bounds they have proven. The fact that a policy performs as best as any other one obviously solves the problem of selectivity.

On the other hand, this problem has already been studied in the context of adversarial bandit by Auer et al. (2003). Their setting differs from our not only because their bandits are nonstochastic, but also because their adaptive procedure takes only into account a given number of recommendations, whereas in our setting the adaptation is supposed to come from observing rewards of the chosen arms (only one per time step). Nevertheless, one can wonder if an "exponentially weighted forecasters" procedure like EXP4 could be transposed to our context. The answer is negative, as stated in the following theorem.

To avoid confusion, we make the notations of the regret and of sampling time more precise by adding the considered policy: under policy $\mathcal{A}$, R_n and $T_k(n)$ will be respectively denoted $R_n(\mathcal{A})$ and $T_k(n,\mathcal{A})$.

Theorem 8 *Let $\tilde{\mathcal{A}}$ be a consistent policy and let ρ be a real number in $(0,0.4)$. If Θ has the Dirac/Bernoulli property and the product property, there is no policy which can both beat $\tilde{\mathcal{A}}$ and* UCB*(ρ), that is:*

$$\forall \mathcal{A},\ \exists \theta \in \Theta,\ \limsup_{n\to+\infty} \frac{\mathbb{E}_\theta[R_n(\mathcal{A})]}{\min(\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})], \mathbb{E}_\theta[R_n(\mathrm{UCB}(\rho))])} > 1.$$

Thus the existence of optimal policies does not hold when we extend the notion of consistency. Precisely, as UCB(ρ) is $(1-2\rho)$-consistent, we have shown that there is no optimal policy within the class of α-consistent policies, with $\alpha > 0.2$. Consequently, there do not exist optimal policies in the class of Hannan consistent policies either.

Moreover, Theorem 8 shows that methods that would be inspired by related literature in adversarial bandit can not apply to our framework. As we said, this impossibility may come from the fact that we can not observe at each time step the decisions and rewards of more than one algorithm. If we were able to observe a given set of policies from step to step, then it would be easy to beat them all: it would be sufficient to aggregate all the observations and simply pull the arm with the greater empiric mean. The case where we only observe decisions (and not rewards) of a set of policies may be interesting, but is left outside of the scope of this paper.

Proof Assume by contradiction that

$$\exists \mathcal{A},\ \forall \theta \in \Theta,\ \limsup_{n\to+\infty} u_{n,\theta} \leq 1,$$

where $u_{n,\theta} = \frac{\mathbb{E}_\theta[R_n(\mathcal{A})]}{\min(\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})], \mathbb{E}_\theta[R_n(UCB(\rho))])}$.

For any θ, we have

$$\mathbb{E}_\theta[R_n(\mathcal{A})] = \frac{\mathbb{E}_\theta[R_n(\mathcal{A})]}{\mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})]} \mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})] \leq u_{n,\theta} \mathbb{E}_\theta[R_n(\tilde{\mathcal{A}})], \tag{9}$$

so that the fact that $\tilde{\mathcal{A}}$ is a consistent policy implies that $\mathcal{A}$ is also consistent. Consequently the lower bound of Theorem 6 also holds for policy $\mathcal{A}$.

For the rest of the proof, we focus on environments of the form $\theta = (\delta_0, \delta_\Delta)$ with $\Delta > 0$. In this case, arm 2 is the best arm, so that we have to compute $D_1(\theta)$. On the one hand, we have:

$$D_1(\theta) = \inf_{\tilde{\nu}_1 \in \Theta_1 : \mathbb{E}[\tilde{\nu}_1] > \mu^*} KL(\nu_1, \tilde{\nu}_1) = \inf_{\tilde{\nu}_1 \in \Theta_1 : \mathbb{E}[\tilde{\nu}_1] > \Delta} KL(\delta_0, \tilde{\nu}_1) = \inf_{\tilde{\nu}_1 \in \Theta_1 : \mathbb{E}[\tilde{\nu}_1] > \Delta} \log\left(\frac{1}{\tilde{\nu}_1(0)}\right).$$

As $\mathbb{E}[\tilde{\nu}_1] \leq 1 - \tilde{\nu}_1(0)$, we get:

$$D_1(\theta) \geq \inf_{\tilde{\nu}_1 \in \Theta_1 : 1-\tilde{\nu}_1(0) \geq \Delta} \log\left(\frac{1}{\tilde{\nu}_1(0)}\right) \geq \log\left(\frac{1}{1-\Delta}\right).$$

One the other hand, we have for any $\varepsilon > 0$:

$$D_1(\theta) \leq KL(\delta_0, Ber(\Delta+\varepsilon)) = \log\left(\frac{1}{1-\Delta-\varepsilon}\right)$$

Consequently $D_1(\theta) = \log\left(\frac{1}{1-\Delta}\right)$, and the lower bound of Theorem 6 reads:

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_1(n,\mathcal{A})]}{\log n} \geq \frac{1}{\log\left(\frac{1}{1-\Delta}\right)}.$$

Just like Equation (9), we have:

$$\mathbb{E}_\theta[R_n(\mathcal{A})] \leq u_{n,\theta}\mathbb{E}_\theta[R_n(UCB(\rho))].$$

Moreover, Lemma 3 provides:

$$\mathbb{E}_\theta[R_n(UCB(\rho))] \leq 1 + \frac{\rho \log n}{\Delta}.$$

Now, by gathering the three previous inequalities and Formula (1), we get:

$$\begin{aligned}
\frac{1}{\log\left(\frac{1}{1-\Delta}\right)} &\leq \liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[T_1(n,\mathcal{A})]}{\log n} = \liminf_{n\to+\infty} \frac{\mathbb{E}_\theta[R_n(\mathcal{A})]}{\Delta \log n} \\
&\leq \liminf_{n\to+\infty} \frac{u_{n,\theta}\mathbb{E}_\theta[R_n(UCB(\rho))]}{\Delta \log n} \leq \liminf_{n\to+\infty} \frac{u_{n,\theta}}{\Delta \log n}\left(1 + \frac{\rho \log n}{\Delta}\right) \\
&\leq \liminf_{n\to+\infty} \frac{u_{n,\theta}}{\Delta \log n} + \liminf_{n\to+\infty} \frac{\rho u_{n,\theta}}{\Delta^2} = \frac{\rho}{\Delta^2} \liminf_{n\to+\infty} u_{n,\theta} \leq \frac{\rho}{\Delta^2} \limsup_{n\to+\infty} u_{n,\theta} \\
&\leq \frac{\rho}{\Delta^2}.
\end{aligned}$$

This means that ρ has to be lower bounded by $\frac{\Delta^2}{\log\left(\frac{1}{1-\Delta}\right)}$, but this is greater than 0.4 if $\Delta = 0.75$, hence the contradiction.

■

Note that this proof gives a simple alternative to Theorem 4 to show that UCB(ρ) is not consistent (if $\rho \leq 0.4$). Indeed if it were consistent, then in environment $\theta = (\delta_0, \delta_\Delta)$ the same contradiction between the lower bound of Theorem 6 and the upper bound of Lemma 3 would hold.

5. General Bounds

In this section, we study lower bounds on the expected regret with few requirements on Θ and on the class of policies. With a simple property on Θ but without any assumption on the policy, we show that there always exist logarithmic lower bounds for some environments θ. Then, still with a

simple property on Θ, we show that there exists a Hannan consistent policy for which the expected regret is sub-logarithmic for some environment θ.

Note that the policy that always pulls arm 1 has a 0 expected regret in environments where arm 1 has the best mean reward, and an expected regret of order n in other environments. So, for this policy, expected regret is sub-logarithmic in some environments. Nevertheless, this policy is not Hannan consistent because its expected regret is not always $o(n)$.

5.1 The Necessity of a Logarithmic Regret in Some Environments

The necessity of a logarithmic regret in some environments can be explained by a simple sketch proof. Assume that the agent knows the number of rounds n, and that he balances exploration and exploitation in the following way: he first pulls each arm $s(n)$ times, and then selects the arm that has obtained the best empiric mean for the rest of the game. Denote by $p_{s(n)}$ the probability that the best arm does not have the best empiric mean after the exploration phase (i.e., after the first $Ks(n)$ rounds). The expected regret is then of the form

$$c_1(1-p_{s(n)})s(n)+c_2 p_{s(n)}n. \tag{10}$$

Indeed, if the agent manages to match the best arm then he only suffers the pulls of suboptimal arms during the exploration phase. That represents an expected regret of order $s(n)$. If not, the number of pulls of suboptimal arms is of order n, and so is the expected regret.

Now, let us approximate $p_{s(n)}$. It has the same order as the probability that the best arm gets an empiric mean lower than the second best mean reward. Moreover, $\frac{X_{k^*,s(n)}-\mu^*}{\sigma}\sqrt{s(n)}$ (where σ is the variance of $X_{k^*,1}$) has approximately a standard normal distribution by the central limit theorem. Therefore, we have:

$$\begin{aligned}
p_{s(n)} &\approx \mathbb{P}_\theta(X_{k^*,s(n)} \le \mu^* - \Delta) = \mathbb{P}_\theta\left(\frac{X_{k^*,s(n)}-\mu^*}{\sigma}\sqrt{s(n)} \le -\frac{\Delta\sqrt{s(n)}}{\sigma}\right) \\
&\approx \frac{1}{\sqrt{2\pi}}\frac{\sigma}{\Delta\sqrt{s(n)}}\exp\left(-\frac{1}{2}\left(\frac{\Delta\sqrt{s(n)}}{\sigma}\right)^2\right) \\
&\approx \frac{1}{\sqrt{2\pi}}\frac{\sigma}{\Delta\sqrt{s(n)}}\exp\left(-\frac{\Delta^2 s(n)}{2\sigma^2}\right).
\end{aligned}$$

It follows that the expected regret has to be at least logarithmic. Indeed, to ensure that the second term $c_2 p_{s(n)}n$ of Equation (10) is sub-logarithmic, $s(n)$ has to be greater than $\log n$. But then first term $c_1(1-p_{s(n)})s(n)$ is greater than $\log n$.

Actually, the necessity of a logarithmic regret can be written as a consequence of Theorem 6. Indeed, if we assume by contradiction that $\limsup_{n\to+\infty}\frac{\mathbb{E}_\theta R_n}{\log n}=0$ for all θ (i.e., $\mathbb{E}_\theta R_n = o(\log n)$), the considered policy is consistent. Consequently, Theorem 6 implies that

$$\limsup_{n\to+\infty}\frac{\mathbb{E}_\theta R_n}{\log n} \ge \liminf_{n\to+\infty}\frac{\mathbb{E}_\theta R_n}{\log n} > 0.$$

Yet, this reasoning needs Θ having the product property, and conditions of the form $0 < D_k(\theta) < \infty$ also have to hold.

The following proposition is a rigorous version of our sketch, and it shows that the necessity of a logarithmic lower bound can be based on a simple property on Θ.

Proposition 9 *Assume that there exist two environments* $\theta = (\nu_1, \ldots, \nu_K) \in \Theta$, $\tilde{\theta} = (\tilde{\nu}_1, \ldots, \tilde{\nu}_K) \in \Theta$, *and an arm* $k \in \{1, \ldots, K\}$ *such that*

1. *k has the best mean reward in environment θ,*
2. *k is not the winning arm in environment $\tilde{\theta}$,*
3. $\nu_k = \tilde{\nu}_k$ *and there exists* $\eta \in (0,1)$ *such that*

$$\prod_{\ell \neq k} \frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) \geq \eta \;\; \mathbb{P}_{\tilde{\theta}} - a.s. \tag{11}$$

Then, for any policy, there exists $\hat{\theta} \in \Theta$ *such that*

$$\limsup_{n \to +\infty} \frac{\mathbb{E}_{\hat{\theta}} R_n}{\log n} > 0.$$

Let us explain the logic of the three conditions of the proposition. If $\nu_k = \tilde{\nu}_k$, and in case ν_k seems to be the reward distribution of arm k, then arm k has to be pulled often enough for the regret to be small if the environment is θ. Nevertheless, one has to explore other arms to know whether the environment is actually $\tilde{\theta}$. Moreover, Inequality (11) makes sure that the distinction between θ and $\tilde{\theta}$ is tough to make: it ensures that pulling any arm $\ell \neq k$ gives a reward which is likely in both environments. Without such an assumption the problem may be very simple, and providing a logarithmic lower bound is hopeless. Indeed, the distinction between any pair of tricky environments $(\theta, \tilde{\theta})$ may be solved in only one pull of a given arm $\ell \neq k$, that would almost surely give a reward that is possible in only one of the two environments.

The third condition can be seen as an alternate version of condition $0 < D_k(\theta) < \infty$ in Theorem 6, though there is no logical connection with it. Finally, let us remark that one can check that any set Θ that has the Dirac/Bernoulli property satisfies the conditions of Proposition 9.

Proof The proof consists in writing a proper version of Expression (10). To this aim we compute a lower bound of $\mathbb{E}_{\tilde{\theta}} R_n$, expressed as a function of $\mathbb{E}_\theta R_n$ and of an arbitrary function $g(n)$.

In the following, $\tilde{\Delta}_k$ denotes the optimality gap of arm k in environment $\tilde{\theta}$. As event $\left\{\sum_{\ell \neq k} T_\ell(n) \leq g(n)\right\}$ is measurable with respect to $X_{\ell,1}, \ldots, X_{\ell,\lfloor g(n) \rfloor}$ ($\ell \neq k$) and to $X_{k,1}, \ldots, X_{k,n}$, we also introduce the function q such that

$$\mathbb{1}_{\left\{\sum_{\ell \neq k} T_\ell(n) \leq g(n)\right\}} = q\big((X_{\ell,s})_{\ell \neq k,\, s=1..\lfloor g(n) \rfloor}, (X_{k,s})_{s=1..n}\big).$$

We have:

$$
\begin{aligned}
\mathbb{E}_{\tilde{\theta}} R_n &\geq \tilde{\Delta}_k \mathbb{E}_{\tilde{\theta}}[T_k(n)] \geq \tilde{\Delta}_k (n - g(n)) \mathbb{P}_{\tilde{\theta}} \left(T_k(n) \geq n - g(n) \right) && (12) \\
&= \tilde{\Delta}_k (n - g(n)) \mathbb{P}_{\tilde{\theta}} \left(n - \sum_{\ell \neq k} T_\ell(n) \geq n - g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \mathbb{P}_{\tilde{\theta}} \left(\sum_{\ell \neq k} T_\ell(n) \leq g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \int q\big((x_{\ell,s})_{\ell \neq k,\ s=1..\lfloor g(n) \rfloor}, (x_{k,s})_{s=1..n} \big) \prod_{\substack{\ell \neq k \\ s=1..\lfloor g(n) \rfloor}} d\tilde{\nu}_\ell(x_{\ell,s}) \prod_{s=1..n} d\tilde{\nu}_k(x_{k,s}) \\
&\geq \tilde{\Delta}_k (n - g(n)) \int q\big((x_{\ell,s})_{\ell \neq k,\ s=1..\lfloor g(n) \rfloor}, (x_{k,s})_{s=1..n} \big) \eta^{\lfloor g(n) \rfloor} \prod_{\substack{\ell \neq k \\ s=1..\lfloor g(n) \rfloor}} d\nu_\ell(x_{\ell,s}) \prod_{s=1..n} d\nu_k(x_{k,s}) && (13) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \int q\big((x_{\ell,s})_{\ell \neq k,\ s=1..\lfloor g(n) \rfloor}, (x_{k,s})_{s=1..n} \big) \prod_{\substack{\ell \neq k \\ s=1..\lfloor g(n) \rfloor}} d\nu_\ell(x_{\ell,s}) \prod_{s=1..n} d\nu_k(x_{k,s}) \\
&= \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \mathbb{P}_\theta \left(\sum_{\ell \neq k} T_\ell(n) \leq g(n) \right) \\
&= \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \mathbb{P}_\theta \left(\sum_{\ell \neq k} T_\ell(n) > g(n) \right) \right) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta \left(\sum_{\ell \neq k} T_\ell(n) \right)}{g(n)} \right) && (14) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta \left(\sum_{\ell \neq k} \Delta_\ell T_\ell(n) \right)}{\Delta g(n)} \right) && (15) \\
&\geq \tilde{\Delta}_k (n - g(n)) \eta^{g(n)} \left(1 - \frac{\mathbb{E}_\theta R_n}{\Delta g(n)} \right),
\end{aligned}
$$

where the first inequality of (12) is a consequence of Formula (1), the second inequality of (12) and inequality (14) come from Markov's inequality, Inequality (13) is a consequence of (11), and Inequality (15) results from the fact that $\Delta_\ell \geq \Delta$ for all ℓ.

Now, let us conclude. If $\frac{\mathbb{E}_\theta R_n}{\log n} \xrightarrow[n \to +\infty]{} 0$, we set $g(n) = \frac{2\mathbb{E}_\theta R_n}{\Delta}$, so that $g(n) \leq \min\left(\frac{n}{2}, \frac{-\log n}{2 \log \eta} \right)$ for n large enough. Then, we have:

$$
\mathbb{E}_{\tilde{\theta}} R_n \geq \tilde{\Delta}_k \frac{n - g(n)}{2} \eta^{g(n)} \geq \tilde{\Delta}_k \frac{n}{4} \eta^{\frac{-\log n}{2 \log \eta}} = \tilde{\Delta}_k \frac{\sqrt{n}}{4}.
$$

In particular, $\frac{\mathbb{E}_{\tilde{\theta}} R_n}{\log n} \xrightarrow[n \to +\infty]{} +\infty$, and the result follows.

■

5.2 Hannan Consistency

We will prove that there exists a Hannan consistent policy such that there can not be a logarithmic lower bound for every environment θ of Θ. To this aim, we make use of general UCB policies again (cf. Section 2.1). Let us first give sufficient conditions on the f_k for UCB policy to be Hannan consistent.

Proposition 10 *Assume that $f_k(n) = o(n)$ for all $k \in \{1,\dots,K\}$. Assume also that there exist $\gamma > \frac{1}{2}$ and $N \geq 3$ such that $f_k(n) \geq \gamma \log\log n$ for all $k \in \{1,\dots,K\}$ and for all $n \geq N$. Then* UCB *is Hannan consistent.*

Proof Fix an arm k such that $\Delta_k > 0$ and choose $\beta \in (0,1)$ such that $2\beta\gamma > 1$. By means of Lemma 2, we have for n large enough:

$$\mathbb{E}_\theta[T_k(n)] \leq u + 2\sum_{t=u+1}^{n} \left(1 + \frac{\log t}{\log(\frac{1}{\beta})}\right) e^{-2\beta\gamma\log\log t},$$

where $u = \left\lceil \frac{4f_k(n)}{\Delta_k^2} \right\rceil$.

Consequently, we have:

$$\mathbb{E}_\theta[T_k(n)] \leq u + 2\sum_{t=2}^{n} \left(\frac{1}{(\log t)^{2\beta\gamma}} + \frac{1}{\log(\frac{1}{\beta})} \frac{1}{(\log t)^{2\beta\gamma-1}} \right). \tag{16}$$

Sums of the form $\sum_{t=2}^{n} \frac{1}{(\log t)^c}$ with $c > 0$ are equivalent to $\frac{n}{(\log n)^c}$ as n goes to infinity. Indeed, on the one hand we have

$$\sum_{t=3}^{n} \frac{1}{(\log t)^c} \leq \int_2^n \frac{dx}{(\log x)^c} \leq \sum_{t=2}^{n} \frac{1}{(\log t)^c},$$

so that $\sum_{t=2}^{n} \frac{1}{(\log t)^c} \sim \int_2^n \frac{dx}{(\log x)^c}$. On the other hand, we have

$$\int_2^n \frac{dx}{(\log x)^c} = \left[\frac{x}{(\log x)^c}\right]_2^n + c\int_2^n \frac{dx}{(\log x)^{c+1}}.$$

As both integrals are divergent we have $\int_2^n \frac{dx}{(\log x)^{c+1}} = o\left(\int_2^n \frac{dx}{(\log x)^c}\right)$, so that $\int_2^n \frac{dx}{(\log x)^c} \sim \frac{n}{(\log n)^c}$.

Combining the fact that $\sum_{t=2}^{n} \frac{1}{(\log t)^c} \sim \frac{n}{(\log n)^c}$ with Equation (16), we get the existence of a constant $C > 0$ such that

$$\mathbb{E}_\theta[T_k(n)] \leq \left\lceil \frac{4f_k(n)}{\Delta^2} \right\rceil + \frac{Cn}{(\log n)^{2\beta\gamma-1}}.$$

Since $f_k(n) = o(n)$ and $2\beta\gamma - 1 > 0$, the latter inequality shows that $\mathbb{E}_\theta[T_k(n)] = o(n)$. The result follows. ■

We are now in the position to prove the main result of this section.

Theorem 11 *If Θ has the Dirac/Bernoulli property, there exist Hannan consistent policies for which the expected regret can not be lower bounded by a logarithmic function in all environments θ.*

Proof If $f_1(n) = f_2(n) = \log\log n$ for $n \geq 3$, UCB is Hannan consistent by Proposition 10. According to Lemma 1, the expected regret is then of order $\log\log n$ in environments of the form (δ_a, δ_b), $a \neq b$. Hence the conclusion on the non-existence of logarithmic lower bounds. ■

Thus we have obtained a lower bound of order $\log\log n$. This order is critical regarding the methods we used. Yet, we do not know if this order is optimal.

Acknowledgments

This work has been supported by the French National Research Agency (ANR) through the COSINUS program (ANR-08-COSI-004: EXPLO-RA project).

References

R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Mathematics*, 27:1054–1078, 1995.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, volume 1, pages 267–281. Springer Verlag, 1973.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

D. Bergemann and J. Valimaki. Bandit problems. In *The New Palgrave Dictionary of Economics, 2nd ed.* Macmillan Press, 2008.

S. Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, France, 2010.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems 21*, pages 201–208. 2009.

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

P.A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Uncertainty in Artificial Intelligence*, 2007.

S. Gelly and Y. Wang. Exploration exploitation in go: UCT for Monte-Carlo go. In *Online Trading between Exploration and Exploitation Workshop, Twentieth Annual Conference on Neural Information Processing Systems (NIPS 2006)*, 2006.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the Twenty-Third Annual Conference on Learning Theory (COLT)*, 2010.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 681–690, 2008.

R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704. 2005.

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

D. Lamberton, G. Pagès, and P. Tarrès. When can the two-armed bandit algorithm be trusted? *Annals of Applied Probability*, 14(3):1424–1454, 2004.

C.L. Mallows. Some comments on cp. *Technometrics*, pages 661–675, 1973.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.