Deviations of stochastic bandit regret

Antoine Salomon[1] and Jean-Yves Audibert[1,2]

[1] Imagine
École des Ponts ParisTech
Université Paris Est
salomona@imagine.enpc.fr
audibert@imagine.enpc.fr
[2] Sierra, CNRS/ENS/INRIA, Paris, France

Abstract. This paper studies the deviations of the regret in a stochastic multi-armed bandit problem. When the total number of plays n is known beforehand by the agent, Audibert et al. (2009) exhibit a policy such that with probability at least $1 - 1/n$, the regret of the policy is of order $\log n$. They have also shown that such a property is not shared by the popular UCB1 policy of Auer et al. (2002). This work first answers an open question: it extends this negative result to any anytime policy. The second contribution of this paper is to design anytime robust policies for specific multi-armed bandit problems in which some restrictions are put on the set of possible distributions of the different arms.

1 Introduction

Bandit problems illustrate the fundamental difficulty of sequential decision making in the face of uncertainty: a decision maker must choose between following what seems to be the best choice in view of the past ("exploitation") or testing ("exploration") some alternative, hoping to discover a choice that beats the current empirical best choice. More precisely, in the stochastic multi-armed bandit problem, at each stage, an agent (or decision maker) chooses one action (or arm), and receives a reward from it. The agent aims at maximizing his rewards. Since he does not know the process generating the rewards, he does not know the best arm, that is the one having the highest expected reward. He thus incurs a regret, that is the difference between the cumulative reward he would have got by always drawing the best arm and the cumulative reward he actually got. The name "bandit" comes from imagining a gambler in a casino playing with K slot machines, where at each round, the gambler pulls the arm of any of the machines and gets a payoff as a result.

The multi-armed bandit problem is the simplest setting where one encounters the exploration-exploitation dilemma. It has a wide range of applications including advertisement [BSS09], [DK09], economics [BV08], [LPT04], games [GW06] and optimization [Kle05], [CM07], [KSU08], [BMSS09]. It can be a central building block of larger systems, like in evolutionary programming [Hol92] and reinforcement learning [SB98], in particular in large state space Markovian

Decision Problems [KS06]. Most of these applications require that the policy of the forecaster works well *for any time*. For instance, in tree search using bandit policies at each node, the number of times the bandit policy will be applied at each node is not known beforehand (except for the root node in some cases), and the bandit policy should thus provide consistently low regret whatever the total number of rounds is.

Most previous works on the stochastic multi-armed bandit [Rob52], [LR85], [Agr95], [ACBF02] (among others) focused on the expected regret, and showed that after n rounds, the expected regret is of order $\log n$. So far, the analysis of the upper tail of the regret was only addressed in [AMS09]. The two main results there about the deviations of the regret are the following. First, after n rounds, for large enough constant $C > 0$, the probability that the regret of UCB1 (and also its variant taking into account the empirical variance) exceeds $C \log n$ is upper bounded by $1/(\log n)^{C'}$ for some constant C' depending on the distributions of the arms and on C (but not on n). Besides, for most bandit problems, this upper bound is tight to the extent that the probability is also lower bounded by a quantity of the same form. Second, a new upper confidence bound policy was proposed: it requires to know the total number of rounds in advance and uses this knowledge to design a policy which essentially explores in the first rounds and then exploits the information gathered in the exploration phase. Its regret has the advantage of being more concentrated to the extent that with probability at least $1 - 1/n$, the regret is of order $\log n$. The problem left open by [AMS09] is whether it is possible to design an anytime robust policy, that is a policy such that for any n, with probability at least $1 - 1/n$, the regret is of order $\log n$. In this paper, we answer negatively to this question when the reward distributions of all arms are just assumed to be uniformly bounded, say all rewards are in $[0, 1]$ for instance (Corollary 7). We then study which kind of restrictions on the set of probabilities defining the bandit problem allows to answer positively. One of our positive results is the following: if the agent knows the value of the expected reward of the best arm (but does not know which arm is the best one), the agent can use this information to design an anytime robust policy (Theorem 12).

The paper is organised as follows: in Section 2, we formally describe the problem we address and give the corresponding definitions and properties. In Section 3, we present our main impossibility result. In Section 4, we provide restrictions under which it is possible to design anytime robust policies. Section 5 is devoted to the proof of our main result. All other proofs are available at `http://hal.archives-ouvertes.fr/hal-00579607/en/`.

2 Problem setup and definitions

In the stochastic multi-armed bandit problem with $K \geq 2$ arms, at each time step $t = 1, 2, \ldots$, an agent has to choose an arm I_t in the set $\{1, \ldots, K\}$ and obtains a reward drawn from ν_{I_t} independently from the past (actions and observations). The environment is thus parameterized by a K-tuple of probability distributions

$\theta = (\nu_1, \dots, \nu_K)$. The agent aims at maximizing his rewards. He does not know θ but knows that it belongs to some set Θ. We assume for simplicity that $\Theta \subset \bar{\Theta}$, where $\bar{\Theta}$ denotes the set of all K-tuple of probability distributions on $[0,1]$. We thus assume that the rewards are in $[0,1]$.

For each arm k and all times $t \geq 1$, let $T_k(t) = \sum_{s=1}^{t} \mathbb{1}_{I_s=k}$ denote the number of times arm k was pulled from round 1 to round t, and by $X_{k,1}, X_{k,2}, \dots, X_{k,T_k(t)}$ the sequence of associated rewards. For an environment parameterized by $\theta = (\nu_1, \dots, \nu_K)$, let $\mathbb{P}_\theta$ denote the distribution on the probability space such that for any $k \in \{1, \dots, K\}$, the random variables $X_{k,1}, X_{k,2}, \dots$ are i.i.d. realizations of ν_k, and such that these K infinite sequence 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. Let $\mu^* = \max_{k \in \{1,\dots,K\}} \mu_k$ and fix an arm k^* such that $\mu_{k^*} = \mu^*$, that is k^* has the best expected reward. The suboptimality of arm k is measured by $\Delta_k = \mu^* - \mu_k$. The agent aims at minimizing its regret defined as the difference between the cumulative reward he would have got by always drawing the best arm and the cumulative reward he actually got. At time $n \geq 1$, its regret is thus

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

The expectation of this regret has a simple expression in terms of the suboptimalities of the arms and the expected sampling times of the arms at time n. Precisely, we have

$$\begin{aligned}
\mathbb{E}_\theta \hat{R}_n &= n\mu^* - \sum_{t=1}^{n} \mathbb{E}_\theta(\mu_{I_t}) = n\mu^* - \mathbb{E}_\theta\left(\sum_{k=1}^{K} T_k(n)\mu_k \right) \\
&= \mu^* \sum_{k=1}^{K} \mathbb{E}_\theta[T_k(n)] - \sum_{k=1}^{K} \mu_k \mathbb{E}_\theta[T_k(n)] = \sum_{k=1}^{K} \Delta_k \mathbb{E}_\theta[T_k(n)].
\end{aligned}$$

Other notions of regret exists in the literature: the quantity $\sum_{k=1}^{K} \Delta_k T_k(n)$ is called the pseudo regret and may be more practical to study, and the quantity $\max_k \sum_{t=1}^{n} X_{k,t} - \sum_{t=1}^{n} X_{I_t, T_{I_t}(t)}$ defines the regret in adverserial settings. Results and ideas we want to convey here are more suited to definition (1), and taking another definition of the regret would only bring some more technical intricacies.

Our main interest is the study of the *deviations* of the regret $\hat{R}_n$, i.e. the value of $\mathbb{P}_\theta(\hat{R}_n \geq x)$ when x is larger and of order of $\mathbb{E}_\theta \hat{R}_n$. If a policy has small deviations, it means that the regret is small with high probability and in particular, if the policy is used on some real data, it is very likely to be small on this specific dataset. Naturally, small deviations imply small expected regret since we have

$$\mathbb{E}_\theta \hat{R}_n \leq \mathbb{E}_\theta \max(\hat{R}_n, 0) = \int_0^{+\infty} \mathbb{P}_\theta\left(\hat{R}_n \geq x\right) dx.$$

To a lesser extent it is also interesting to study the deviations of the sampling times $T_n(k)$, as this shows the ability of a policy to match the best arm. Moreover our analysis is mostly based on results on the deviations of the sampling times, which then enables to derive results on the regret. We thus define below the notion of being f-upper tailed for both quantities.
Define $\mathbb{R}_+^* = \{x \in \mathbb{R} : x > 0\}$, and let $\Delta = \min_{k \neq k^*} \Delta_k$ be the gap between the best arm and second best arm.

Definition 1 (f-$\mathcal{T}$ and f-$\mathcal{R}$) *Consider a mapping $f : \mathbb{R} \to \mathbb{R}_+^*$. A policy has f-upper tailed sampling Times (in short, we will say that the policy is f-$\mathcal{T}$) if and only if*

$$\exists C, \tilde{C} > 0,\ \forall \theta \in \Theta \text{ such that } \Delta \neq 0,$$
$$\forall n \geq 2,\ \forall k \neq k^*,\ \mathbb{P}_\theta \left(T_k(n) \geq C \frac{\log n}{\Delta_k^2} \right) \leq \frac{\tilde{C}}{f(n)}.$$

A policy has f-upper tailed Regret (in short, f-$\mathcal{R}$) if and only if

$$\exists C, \tilde{C} > 0,\ \forall \theta \in \Theta \text{ such that } \Delta \neq 0,\ \forall n \geq 2,\ \mathbb{P}_\theta \left(\hat{R}_n \geq C \frac{\log n}{\Delta} \right) \leq \frac{\tilde{C}}{f(n)}.$$

We will sometimes prefer to denote $f(n)$-$\mathcal{T}$ (resp. $f(n)$-$\mathcal{R}$) instead of f-$\mathcal{T}$ (resp. f-$\mathcal{R}$) for readability. Note also that, for sake of simplicity, we leave aside the degenerated case of Δ being null (i.e. when there are at least two optimal arms).

In this definition, we considered that the number K of arms is fixed, meaning that C and $\tilde{C}$ may depend on K. The thresholds considered on $T_k(n)$ and $\hat{R}_n$ directly come from known tight upper bounds on the expectation of these quantities for several policies. To illustrate this, let us recall the definition and properties of the popular UCB1 policy. Let $\hat{X}_{k,s} = \frac{1}{s} \sum_{t=1}^{s} X_{k,t}$ be the empirical mean of arm k after s pulls. In UCB1, the agent plays each arm once, and then (from $t \geq K+1$), he plays

$$I_t \in \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \left\{ \hat{X}_{k,T_k(t-1)} + \sqrt{\frac{2 \log t}{T_k(t-1)}} \right\}. \tag{2}$$

While the first term in the bracket ensures the exploitation of the knowledge gathered during steps 1 to $t-1$, the second one ensures the exploration of the less sampled arms. For this policy, [ACBF02] proved:

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

[LR85] showed that these results cannot be improved up to numerical constants. [AMS09] proved that UCB1 is $\log^3$-$\mathcal{T}$ and $\log^3$-$\mathcal{R}$ where $\log^3$ is the function $x \mapsto [\log(x)]^3$. Besides, they also study the case when $2 \log t$ is replaced by $\rho \log t$

in (2) with $\rho > 0$, and proved that this modified UCB1 is $\log^{2\rho-1}$-$\mathcal{T}$ and $\log^{2\rho-1}$-$\mathcal{R}$ for $\rho > 1/2$, and that $\rho = \frac{1}{2}$ is actually a critical value, since for $\rho < 1/2$, the policy does not even have a logarithmic regret guarantee in expectation. Another variant of UCB1 proposed by Audibert et al. is to replace $2\log t$ by $2\log n$ in (2) when we want to have low and concentrated regret at a fixed given time n. We refer to it as UCB-H as its implementation requires the knowledge of the horizon n of the game. The behaviour of UCB-H on the time interval $[1, n]$ is significantly different to the one of UCB1, as UCB-H will explore much more at the beginning of the interval, and thus avoids exploiting the suboptimal arms on the early rounds. Audibert et al. showed that UCB-H is n-$\mathcal{T}$ and n-$\mathcal{R}$ (as it will be recalled in Theorem 8).

We now introduce the weak notion of f-upper tailed as this notion will be used to get our strongest impossibility results.

Definition 2 (f-w$\mathcal{T}$ and f-w$\mathcal{R}$) *Consider a mapping $f : \mathbb{R} \to \mathbb{R}_+^*$. A policy has weak f-upper tailed sampling Times (in short, we will say that the policy is f-w$\mathcal{T}$) if and only if*

$\forall \theta \in \Theta$ *such that* $\Delta \neq 0$,

$$\exists C, \tilde{C} > 0,\ \forall n \geq 2,\ \forall k \neq k^*,\ \mathbb{P}_\theta \left(T_k(n) \geq C \frac{\log n}{\Delta_k^2} \right) \leq \frac{\tilde{C}}{f(n)}.$$

A policy has weak f-upper tailed Regret (in short, f-w$\mathcal{R}$) if and only if

$$\forall \theta \in \Theta \text{ such that } \Delta \neq 0,\ \exists C, \tilde{C} > 0,\ \forall n \geq 2,\ \mathbb{P}_\theta \left(\hat{R}_n \geq C \frac{\log n}{\Delta} \right) \leq \frac{\tilde{C}}{f(n)}.$$

The only difference between f-$\mathcal{T}$ and f-w$\mathcal{T}$ (and between f-$\mathcal{R}$ and f-w$\mathcal{R}$) is the interchange of "$\forall\theta$" and "$\exists C, \tilde{C}$". Consequently, a policy that is f-$\mathcal{T}$ (respectively f-$\mathcal{R}$) is f-$\mathcal{T}$ (respectively f-w$\mathcal{R}$). Let us detail the links between the f-$\mathcal{T}$, f-$\mathcal{R}$, f-w$\mathcal{T}$ and f-w$\mathcal{R}$.

Proposition 3 *Assume that there exists $\alpha, \beta > 0$ such that $f(n) \leq \alpha n^\beta$ for any $n \geq 2$. We have*

$$f\text{-}\mathcal{T} \Rightarrow f\text{-}\mathcal{R} \Rightarrow f\text{-w}\mathcal{R} \Leftrightarrow f\text{-w}\mathcal{T}.$$

The proof of this proposition is technical but rather straightforward. Note that we do not have f-$\mathcal{R} \Rightarrow f$-$\mathcal{T}$, because the agent may not regret having pulled a suboptimal arm if the latter has delivered good rewards. Note also that f is required to be at most polynomial: if not some rare events such as unlikely deviations of rewards towards their actual mean can not be neglected, and none of the implications hold in general (except, of course, f-$\mathcal{R} \Rightarrow f$-w$\mathcal{R}$ and f-$\mathcal{T} \Rightarrow f$-w$\mathcal{T}$).

3 Impossibility result

From now on, we mostly deal with anytime policies (i.e. policies that do not have the knowledge of the horizon n) and the word policy (or algorithm) implicitly refers to anytime policy.
In the previous section, we have mentioned that for any $\rho > 1/2$, there is a variant of UCB1 (obtained by changing $2\log t$ into $\rho \log t$ in (2)) which is $\log^{2\rho-1}$-$\mathcal{T}$. This means that, for any $\alpha > 0$, there exists a $\log^{\alpha}$-$\mathcal{T}$ policy, and a hence $\log^{\alpha}$-$\mathcal{R}$ policy. The following result shows that it is impossible to find an algorithm that would have better deviation properties than these UCB policies. For many usual settings (e.g., when Θ is the set $\bar{\Theta}$ of all K-tuples of measures on $[0,1]$), with not so small probability, the agent gets stuck drawing a suboptimal arm he believes best. Precisely, this situation arises when simultaneously:

(a) an arm k delivers payoffs according to a same distribution ν_k in two distinct environments θ and $\tilde{\theta}$,
(b) arm k is optimal in θ but suboptimal in $\tilde{\theta}$,
(c) in environment $\tilde{\theta}$, other arms may behave as in environment θ, i.e. with positive probability other arms deliver payoffs that are likely in both environments.

If the agent suspects that arm k delivers payoffs according to ν_k, he does not know if he has to pull arm k again (in case the environment is θ) or to pull the optimal arm of $\tilde{\theta}$. The other arms can help to point out the difference between θ and $\tilde{\theta}$, but then they have to be chosen often enough. This is in fact this kind of situation that has to be taken into account when balancing a policy between exploitation and exploration.

Our main result is the formalization of the leads given above. In particular, we give a rigorous description of conditions (a), (b) and (c). Let us first recall the following results, which are needed in the formalization of condition (c). One may look at [Rud86], p.121 for details (among others). Those who are not familiar with measure theory can skip to the non-formal explanation just after the results.

Theorem 4 (Lebesgue-Radon-Nikodym theorem) *Let μ_1 and μ_2 be σ-finite measures on a given measurable space. There exists a μ_2-integrable function $\frac{d\mu_1}{d\mu_2}$ and a σ-finite measure m such that m and μ_2 are singular*[3] *and*

$$\mu_1 = \frac{d\mu_1}{d\mu_2} \cdot \mu_2 + m.$$

The density $\frac{d\mu_1}{d\mu_2}$ is unique up to a μ_2-negligible event.

[3] Two measures m_1 and m_2 on a measurable space $(\Omega, \mathcal{F})$ are singular if and only if there exists two disjoint measurable sets A_1 and A_2 such that $A_1 \cup A_2 = \Omega$, $m_1(A_2) = 0$ and $m_2(A_1) = 0$.

We adopt the convention that $\frac{d\mu_1}{d\mu_2} = +\infty$ on the complementary of the support of μ_2.

Lemma 5 *We have*

- $\mu_1\big(\frac{d\mu_1}{d\mu_2} = 0\big) = 0$.
- $\mu_2\big(\frac{d\mu_1}{d\mu_2} > 0\big) > 0 \Leftrightarrow \mu_1\big(\frac{d\mu_2}{d\mu_1} > 0\big) > 0$.

Proof. The first point is a clear consequence of the decomposition $\mu_1 = \frac{d\mu_1}{d\mu_2} \cdot \mu_2 + m$ and of the convention mentioned above. For the second point, one can write by uniqueness of the decomposition:

$$\mu_2\left(\frac{d\mu_1}{d\mu_2} > 0\right) = 0 \Leftrightarrow \frac{d\mu_1}{d\mu_2} = 0 \ \mu_2 - a.s. \Leftrightarrow \mu_1 = m \Leftrightarrow \mu_1 \text{ and } \mu_2 \text{ are singular.}$$

And by symmetry of the roles of μ_1 and μ_2:

$$\mu_2\left(\frac{d\mu_1}{d\mu_2} > 0\right) > 0 \Leftrightarrow \mu_1 \text{ and } \mu_2 \text{ are not singular} \Leftrightarrow \mu_1\left(\frac{d\mu_2}{d\mu_1} > 0\right) > 0.$$

Let us explain what these results has to do with condition (c).
One may be able to distinguish environment θ from $\tilde{\theta}$ if a certain arm ℓ delivers a payoff that is infinitely more likely in $\tilde{\theta}$ than in θ. This is for instance the case if $X_{\ell,t}$ is in the support of $\tilde{\nu}_\ell$ and not in the support of ν_ℓ, but our condition is more general. If the agent observes a payoff x from arm ℓ, the quantity $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(x)$ represents how much the observation of x is more likely in environment θ than in $\tilde{\theta}$. If ν_k and $\tilde{\nu}_k$ admit density functions (say, respectively, f and $\tilde{f}$) with respect to a common measure, then $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(x) = \frac{f(x)}{\tilde{f}(x)}$. Thus the agent will almost never make a mistake if he removes θ from possible environments when $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(x) = 0$. This may happen even if x is in both supports of ν_ℓ and $\tilde{\nu}_\ell$, for example if x is an atom of $\tilde{\nu}_\ell$ and not of ν_ℓ (i.e. $\tilde{\nu}_\ell(x) > 0$ and $\nu_\ell(x)$=0). On the contrary, if $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(x) > 0$ both environments θ and $\tilde{\theta}$ are likely and arm ℓ's behaviour is both consistent with θ and $\tilde{\theta}$.

Now let us state the impossibility result. Here and throughout the paper we find it more convenient to denote $f \gg_{+\infty} g$ rather than the usual notation $g = o(f)$, which has the following meaning:

$$\forall \varepsilon > 0,\ \exists N \geq 0,\ \forall n \geq N,\ g(n) \leq \varepsilon f(n).$$

Theorem 6 *Let $f : \mathbb{N} \to \mathbb{R}_+^*$ be greater than any $\log^\alpha$, that is $f \gg_{+\infty} \log^\alpha$ for any $\alpha > 0$. Assume that there exists θ, $\tilde{\theta} \in \Theta$, and $k \in \{1, \dots, K\}$ such that:*

(a) $\nu_k = \tilde{\nu}_k$,
(b) *k is the index of the best arm in θ but not in $\tilde{\theta}$,*
(c) $\forall \ell \neq k,\ \mathbb{P}_{\tilde{\theta}}\big(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\big) > 0$.

Then there is no f-w$\mathcal{T}$ policy, and hence no f-$\mathcal{R}$ policy.

Let us give some hints of the proof (see Section 5 for details). The main idea is to consider a policy that would be f-w$\mathcal{T}$, and in particular that would "work well" in environment θ in the sense given by the definition of f-w$\mathcal{T}$. The proof exhibits a time N at which arm k, optimal in environment θ and thus often drawn with high $\mathbb{P}_\theta$-probability, is drawn too many times (more than the logarithmic threshold $C\log(N)/\Delta_k^2$) with not so small $\mathbb{P}_{\tilde{\theta}}$-probability, which shows the nonexistence of such a policy. More precisely, let n be large enough and consider a time N of order $\log n$ and above the threshold. If the policy is f-w$\mathcal{T}$, at time N, sampling times of suboptimal arms are of order $\log N$ at most, with $\mathbb{P}_\theta$-probability at least $1-\tilde{C}/f(N)$. In this case, at time N, the draws are concentrated on arm k. So $T_k(N)$ is of order N, which is more than the threshold. This event holds with high $\mathbb{P}_\theta$-probability. Now, from (a) and (c), we exhibit constants that are characteristic of the ability of arms $\ell \neq k$ to "behave as if in θ": for some $0 < a, \eta < 1$, there is a subset ξ of this event such that $\mathbb{P}_\theta(\xi) \geq a^T$ for $T = \sum_{\ell \neq k} T_\ell(N)$ and for which $\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\tilde{\theta}}}$ is lower bounded by η^T. The event ξ on which the arm k is sampled N times at least has therefore a $\mathbb{P}_{\tilde{\theta}}$-probability of order $(\eta a)^T$ at least. This concludes this sketchy proof since T is of order $\log N$, thus $(\eta a)^T$ is of order $\log^{\log(\eta a)} n$ at least.

Note that the conditions given in Theorem 6 are not very restrictive. The impossibility holds for very basic settings, and may hold even if the agent has great knowledge of the possible environments. For instance, the setting

$$K = 2 \text{ and } \Theta = \left\{ \left(Ber\left(\frac{1}{4}\right), \delta_{\frac{1}{2}} \right), \left(Ber\left(\frac{3}{4}\right), \delta_{\frac{1}{2}} \right) \right\},$$

where $Ber(p)$ denotes the Bernoulli distribution of parameter p and δ_x the Dirac measure on x, satisfies the three conditions of the theorem.
Nevertheless, the main interest of the result regarding the previous literature is the following corollary.

Corollary 7 *If Θ is the whole set $\bar{\Theta}$ of all K-tuples of measures on $[0,1]$, then there is no f-$\mathcal{R}$ policy, where f is any function such that $f \gg_{+\infty} \log^\alpha$ for all $\alpha > 0$.*

This corollary should be read in conjunction with the following result for UCB-H which, for a given n, plays at time $t \geq K+1$,

$$I_t \in \operatorname*{argmax}_{k \in \{1,\dots,K\}} \left\{ \hat{X}_{k,T_k(t-1)} + \sqrt{\frac{2\log n}{T_k(t-1)}} \right\}.$$

Theorem 8 *For any $\beta > 0$,* UCB-H *is n^β-$\mathcal{R}$.*

For $\rho > 1$, Theorem 8 can easily be extended to the policy UCB-H(ρ) which starts by drawing each arm once, and then at time $t \geq K+1$, plays

$$I_t \in \operatorname*{argmax}_{k \in \{1,\dots,K\}} \left\{ \hat{X}_{k,T_k(t-1)} + \sqrt{\frac{\rho\log n}{T_k(t-1)}} \right\}. \tag{3}$$

Naturally, we have $n^\beta \gg_{n\to+\infty} \log^\alpha(n)$ for all $\alpha, \beta > 0$ but this does not contradict our theorem, since UCB-H(ρ) is not an *anytime* policy. UCB-H will work fine if the horizon n is known in advance, but may perform poorly at other rounds. In particular and as any policy, in view of Corollary 7, it cannot achieve anytime polynomial regret concentration.

Corollary 7 should also be read in conjunction with the following result for the policy UCB1(ρ) which starts by drawing each arm once, and then at time $t \geq K+1$, plays

$$I_t \in \operatorname*{argmax}_{k\in\{1,\dots,K\}} \left\{ \hat{X}_{k,T_k(t-1)} + \sqrt{\frac{\rho \log t}{T_k(t-1)}} \right\}. \tag{4}$$

Theorem 9 *For any $\rho > 1/2$,* UCB1(ρ) *is* $\log^{2\rho-1}$*-$\mathcal{R}$.*

Thus, any improvements of existing algorithms which would for instance involve estimations of variance (see [AMS09]), of Δ_k, or of many characteristics of the distributions cannot beat the variants of UCB1 regarding deviations.

4 Positive results

The intuition behind Theorem 6 suggests that, if one of the three conditions (a), (b), (c) does not hold, a robust policy would consist in the following: at each round and for each arm k, compute a distance between the empirical distribution of arm k and the set of distribution ν_k that makes arm k optimal in a given environment θ. As this distance decreases with our belief that k is the optimal arm, the policy consists in taking the k minimizing the distance. Thus, the agent chooses an arm that fits better a winning distribution ν_k. He cannot get stuck pulling a suboptimal arm because there are no environments $\tilde{\theta}$ with $\nu_k = \tilde{\nu}_k$ in which k would be suboptimal. More precisely, if there exists such an environment $\tilde{\theta}$, the agent is able to distinguish θ from $\tilde{\theta}$: during the first rounds, he pulls every arm and at least one of them will never behave as if in θ if the current environment is $\tilde{\theta}$. Thus, in $\tilde{\theta}$, he is able to remove θ from the set of possible environments Θ (remember that Θ is a parameter of the problem which is known by the agent).

Nevertheless such a policy cannot work in general, notably because of the three following limitations:

- If $\tilde{\theta}$ is the current environment and even if the agent has identified θ as impossible (i.e. $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) = 0$), there still could be other environments θ' that are arbitrary close to θ in which arm k is optimal and which the agent is not able to distinguish from $\tilde{\theta}$. This means that the agent may pull arm k too often because distribution $\tilde{\nu}_k = \nu_k$ is too close to a distribution ν'_k that makes arm k the optimal arm.
- The ability to identify environments as impossible relies on the fact that the event $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) > 0$ is almost sure under $\mathbb{P}_\theta$ (see Lemma 5). If the set

of all environments Θ is uncountable, such a criterion can lead to exclude the actual environment. For instance, assume an agent has to distinguish a distribution among all Dirac measures δ_x ($x \in [0,1]$) and the uniform probability λ over $[0,1]$. Whatever the payoff x observed by the agent, he will always exclude λ from the possible distributions, as x is always infinitely more likely under δ_x than under λ:

$$\forall x \in [0,1],\ \frac{d\lambda}{d\delta_x}(x) = 0.$$

– On the other hand, the agent could legitimately consider an environment θ as unlikely if, for $\varepsilon > 0$ small enough, there exists $\tilde{\theta}$ such that $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) \leq \varepsilon$. Criterion (c) only considers as unlikely an environment θ when there exists $\tilde{\theta}$ such that $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) = 0$.

Despite these limitations, we give in this section sufficient conditions on Θ for such a policy to be robust. This is equivalent to finding conditions on Θ under which the converse of Theorem 6 holds, i.e. under which the fact one of the conditions (a), (b) or (c) does not hold implies the existence of a robust policy. This can also be expressed as finding which kind of knowledge of the environment enables to design anytime robust policies.

We estimate distributions of each arm by means of their empirical cumulative distribution functions, and distance between two c.d.f. is measured by the norm $\|.\|_\infty$, defined by $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$ where f is any function $[0,1] \to \mathbb{R}$. The empirical c.d.f of arm k after having been pulled t times is denoted $\hat{F}_{k,t}$. The way we choose an arm at each round is based on confidence areas around $\hat{F}_{k,T_k(n-1)}$. We choose the greater confidence level (GCL) such that there is still an arm k and a winning distribution ν_k such that F_{ν_k}, the c.d.f. of ν_k, is in the area of $\hat{F}_{k,T_k(n-1)}$. We then select the corresponding arm k. By means of Massart's inequality (1990), this leads to the c.d.f. based algorithm described in Figure 1. Let Θ_k denote the set $\{\theta \in \Theta | k \text{ is the optimal arm in } \theta\}$, i.e. the set of environments that makes k the index of the optimal arm.

Proceed as follows:

– Draw each arm once.
– Remove each $\theta \in \Theta$ such that there exists $\tilde{\theta} \in \Theta$ and $\ell \in \{1,\dots,K\}$ with $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) = 0$.
– Then at each round t, play an arm

$$I_t \in \operatorname*{argmin}_{k\in\{1,\dots,K\}} T_k(t-1) \inf_{\theta\in\Theta_k} \left\|\hat{F}_{k,T_k(t-1)} - F_{\nu_k}\right\|_\infty^2.$$

Fig. 1. A c.d.f.-based algorithm: GCL.

4.1 Θ is finite

When Θ is finite the limitations presented above do not really matter, so that the converse of Theorem 6 is true and our algorithm is robust.

Theorem 10 *Assume that Θ is finite and that for all $\theta = (\nu_1, \dots, \nu_K)$, $\tilde{\theta} = (\tilde{\nu}_1, \dots, \tilde{\nu}_K) \in \Theta$, and all $k \in \{1, \dots, K\}$, at least one of the following holds:*

- $\nu_k \neq \tilde{\nu}_k$,
- *k is suboptimal in θ, or is optimal in $\tilde{\theta}$.*
- $\exists \ell \neq k,\ \mathbb{P}_{\tilde{\theta}}\left(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\right) = 0.$

Then GCL *is n^β-$\mathcal{T}$ (and hence n^β-$\mathcal{R}$) for all $\beta > 0$.*

4.2 Bernoulli laws

We assume that any ν_k ($k \in \{1, \dots, K\}$, $\theta \in \Theta$) is a Bernoulli law, and denote by μ_k its parameter. We also assume that there exists $\gamma \in (0, 1)$ such that $\mu_k \in [\gamma, 1]$ for all k and all θ.[4] Moreover we may denote arbitrary environments $\theta, \tilde{\theta}$ by $\theta = (\mu_1, \dots, \mu_K)$ and $\tilde{\theta} = (\tilde{\mu}_1, \dots, \tilde{\mu}_K)$.
In this case $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(1) = \frac{\mu_l}{\tilde{\mu}_l} > 0$, so that for any θ, $\tilde{\theta} \in \Theta$ and any $l \in \{1, \dots, K\}$ one has

$$\mathbb{P}_{\tilde{\theta}}\left(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\right) \geq \mathbb{P}_{\tilde{\theta}}(X_{\ell,1} = 1) = \tilde{\mu}_l > 0.$$

Therefore condition (c) of Theorem 6 holds, and the impossibility result only relies on conditions (a) and (b). Our algorithm can be made simpler: there is no need to try to exclude unlikely environments and computing the empirical c.d.f. is equivalent to computing the empirical mean (see Figure 2). The theorem and its converse are expressed as follows. We will refer to our policy as GCL-B as it looks for the environment matching the observations at the Greatest Confidence Level, in the case of Bernoulli distributions.

Proceed as follows:

- Draw each arm once.
- Then at each round t, play an arm

$$I_t \in \operatorname*{argmin}_{k \in \{1, \dots, K\}} T_k(t-1) \inf_{\theta \in \Theta_k} \left(\mu_k - \hat{X}_{k, T_k(t-1)}\right)^2.$$

Fig. 2. A c.d.f.-based algorithm in case of Bernoulli laws: GCL-B.

[4] The result also holds if all parameters μ_k are in a given interval $[0, \gamma]$, $\gamma \in (0, 1)$.

Theorem 11 *For any $\theta \in \Theta$ and any $k \in \{1, \dots, K\}$, let us set*

$$d_k = \inf_{\tilde{\theta} \in \Theta_k} |\mu_k - \tilde{\mu}_k|.$$

GCL-B *is such that:* $\forall \beta > 0,\ \exists C, \tilde{C} > 0, \forall \theta \in \Theta,\ \forall n \geq 2,$

$$\forall k \in \{1, \dots, K\},\ \mathbb{P}_\theta \left(T_k(n) \geq \frac{C \log n}{d_k^2} \right) \leq \frac{\tilde{C}}{n^\beta}.$$

Let $f : \mathbb{N}^* \to \mathbb{R}_+^*$ be greater than any $\log^\alpha$, that is $\forall \alpha > 0,\ f \gg_{+\infty} \log^\alpha$. If there exists k such that

(a') $$\inf_{\theta \in \Theta \setminus \Theta_k} d_k = \inf_{\substack{\theta \in \Theta_k \\ \tilde{\theta} \in \Theta \setminus \Theta_k}} |\mu_k - \tilde{\mu}_k| = 0,$$

then there is no policy such that:

$$\exists C, \tilde{C} > 0, \forall \theta \in \Theta,\ \forall n \geq 2,\ \forall k \neq k^*,\ \mathbb{P}_\theta \left(T_k(n) \geq C \log n \right) \leq \frac{\tilde{C}}{f(n)}.$$

Note that we do not adopt the former definitions of robustness (f-$\mathcal{R}$ and f-$\mathcal{T}$), because the significant term here is d_k (and not Δ_k)[5], which represents the distance between Θ_k and $\Theta \setminus \Theta_k$. Indeed robustness lies on the ability to distinguish environments, and this ability is all the more stronger as the distance between the parameters of these environments is greater. Provided that the density $\frac{d\nu}{d\tilde{\nu}}$ is uniformly bounded away from zero, the theorem holds for any parametric model, with d_k being defined with a norm on the space of parameters (instead of $|.|$). Note also that the second part of the theorem is a bit weaker than Theorem 6, because of the interchange of "$\forall \theta$" and "$\exists C, \tilde{C}$". The reason for this is that condition (a) is replaced by a weaker assumption: ν_k does not equal $\tilde{\nu}_k$, but condition (a') means that such ν_k and $\tilde{\nu}_k$ can be chosen arbitrarily close.

4.3 μ^* is known

This section shows that the impossibility result also breaks down if μ^* is known by the agent. This situation is formalized as μ^* being constant over Θ. Conditions (a) and (b) of Theorem 6 do not hold: if a distribution ν_k makes arm k optimal in an environment θ, it is still optimal in any environment $\tilde{\theta}$ such that $\tilde{\nu}_k = \nu_k$. In this case, our algorithm can be made simpler (see Figure 3). At each round we choose the greatest confidence level such that at least one empirical mean $\hat{X}_{k,T_k(t-1)}$ has μ^* in its confidence interval, and select the corresponding arm k. This is similar to the previous algorithm, deviations being evaluated according to Hoeffding's inequality instead of Massart's one. We will refer to this policy as GCL*.

Theorem 12 *When μ^* is known,* GCL* *is n^β-T (and hence n^β-R) for all $\beta > 0$.*

[5] There is no need to leave aside the case of $d_k = 0$: with the convention $\frac{1}{0} = +\infty$, the corresponding event has zero probability.

Proceed as follows:

- Draw each arm once.
- Then at each round t, play an arm

$$I_t \in \operatorname*{argmin}_{k\in\{1,\dots,K\}} T_k(t-1)\left(\mu^* - \hat{X}_{k,T_k(t-1)}\right)^2.$$

Fig. 3. GCL*: a variant of c.d.f.-based algorithm when μ^* is known.

5 Proof of Theorem 6

Let us first notice that we can remove the Δ_k^2 denominator in the the definition of f-w$\mathcal{T}$ without loss of generality. This would not be possible for the f-$\mathcal{T}$ definition owing to the different position of "$\forall\theta$" with respect to "$\exists C, \tilde{C}$".
Thus, a policy is f-w$\mathcal{T}$ if and only if

$\forall\theta \in \Theta$ such that $\Delta \neq 0$,

$$\exists C, \tilde{C} > 0,\ \forall n \geq 2,\ \forall k \neq k^*,\ \mathbb{P}_\theta\left(T_k(n) \geq C\log n\right) \leq \frac{\tilde{C}}{f(n)}.$$

Let us assume that the policy has the f-upper tailed property in θ, i.e., there exists $C, \tilde{C} > 0$

$$\forall N \geq 2,\ \forall \ell \neq k,\ \mathbb{P}_\theta\big(T_\ell(N) \geq C\log N\big) \leq \frac{\tilde{C}}{f(N)}. \tag{5}$$

Let us show that this implies that the policy cannot have also the f-upper tailed property in $\tilde{\theta}$. To prove the latter, it is enough to show that for any $C', \tilde{C}' > 0$

$$\exists n \geq 2,\ \mathbb{P}_{\tilde{\theta}}\big(T_k(n) \geq C'\log n\big) > \frac{\tilde{C}'}{f(n)}. \tag{6}$$

since k is suboptimal in environment $\tilde{\theta}$. Note that proving (6) for $C' = C$ is sufficient. Indeed if (6) holds for $C' = C$, it a fortiori holds for $C' < C$. Besides, when $C' > C$, (5) holds for C replaced by C', and we are thus brought back to the situation when $C = C'$. So we only need to lower bound $\mathbb{P}_{\tilde{\theta}}\big(T_k(n) \geq C\log n\big)$.

From Lemma 5, $\mathbb{P}_{\tilde{\theta}}\big(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\big) > 0$ is equivalent to $\mathbb{P}_\theta\big(\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) > 0\big) > 0$. By independence of $X_{1,1}, \dots, X_{K,1}$ under $\mathbb{P}_\theta$, condition (c) in the theorem may be written as

$$\mathbb{P}_\theta\left(\prod_{\ell\neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) > 0\right) > 0.$$

Since $\left\{\prod_{\ell\neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) > 0\right\} = \cup_{m\geq 2}\left\{\prod_{\ell\neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \frac{1}{m}\right\}$, this readily implies that

$$\exists \eta \in (0,1),\ \mathbb{P}_\theta\left(\prod_{\ell\neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \eta\right) > 0.$$

Let $a = \mathbb{P}_\theta\Big(\prod_{\ell\neq k} \frac{d\tilde\nu_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \eta\Big)$.

Let us take n large enough such that $N = \lfloor 4C\log n\rfloor$ satisfies $N < n$, $C\log N < \frac{N}{2K}$ and $f(n)\eta^t\big(a^t - \frac{(K-1)\tilde C}{f(N)}\big) > \tilde C'$ for $t = \lfloor C\log N\rfloor$. For any $\tilde C'$, such a n does exist since $f \gg_{+\infty} \log^\alpha$ for any $\alpha > 0$.

The idea is that if until round N, arms $\ell \neq k$ have a behaviour that is typical of θ, then the arm k (which is suboptimal in $\tilde\theta$) may be pulled about $C\log n$ times at round N. Precisely, we prove that $\forall \ell \neq k$, $\mathbb{P}_\theta\big(T_\ell(N) \geq C\log N\big) \leq \frac{\tilde C}{f(N)}$ implies $\mathbb{P}_{\tilde\theta}\big(T_k(n) \geq C'\log n\big) > \frac{\tilde C'}{f(n)}$. Let $A_t = \cap_{s=1\dots t}\Big\{\prod_{\ell\neq k}\frac{d\tilde\nu_\ell}{d\nu_\ell}(X_{\ell,s}) \geq \eta\Big\}$. By independence and by definition of a, we have $\mathbb{P}_\theta(A_t) = a^t$. We also have

$$\begin{aligned}
\mathbb{P}_{\tilde\theta}\big(T_k(n) \geq C\log n\big) &\geq \mathbb{P}_{\tilde\theta}\left(T_k(N) \geq \frac{N}{2}\right)\\
&\geq \mathbb{P}_{\tilde\theta}\left(\bigcap_{\ell\neq k}\left\{T_\ell(N) \leq \frac{N}{2K}\right\}\right)\\
&\geq \mathbb{P}_{\tilde\theta}\left(\bigcap_{\ell\neq k}\Big\{T_\ell(N) < C\log N\Big\}\right)\\
&\geq \mathbb{P}_{\tilde\theta}\left(A_t \cap \left\{\bigcap_{\ell\neq k}\Big\{T_\ell(N) < C\log N\Big\}\right\}\right).
\end{aligned}$$

Introduce $B_N = \bigcap_{\ell\neq k}\big\{T_\ell(N) < C\log N\big\}$, and the function q such that

$$\mathbb{1}_{A_t\cap B_N} = q\big((X_{\ell,s})_{\ell\neq k,\ s=1..t}, (X_{k,s})_{s=1..N}\big).$$

Since $\tilde\nu_k = \nu_k$, by definition of A_t and by standard properties of density functions $\frac{d\tilde\nu_\ell}{d\nu_\ell}$, we have

$$\begin{aligned}
&\mathbb{P}_{\tilde\theta}\left(A_t \cap \left\{\bigcap_{\ell\neq k}\{T_\ell(N) < C\log N\}\right\}\right)\\
&= \int q\big((x_{\ell,s})_{\ell\neq k,\ s=1..t}, (x_{k,s})_{s=1..N}\big)\prod_{\substack{\ell\neq k\\ s=1..t}} d\tilde\nu_\ell(x_{\ell,s})\prod_{s=1..N} d\tilde\nu_k(x_{k,s})\\
&\geq \eta^t\int q\big((x_{\ell,s})_{\ell\neq k,\ s=1..t}, (x_{k,s})_{s=1..N}\big)\prod_{\substack{\ell\neq k\\ s=1..t}} d\nu_\ell(x_{\ell,s})\prod_{s=1..N} d\nu_k(x_{k,s})\\
&= \eta^t\mathbb{P}_\theta\left(A_t \cap \left\{\bigcap_{\ell\neq k}\{T_\ell(N) < C\log N\}\right\}\right) \geq \eta^t\left(a^t - \frac{(K-1)\tilde C}{f(N)}\right) > \frac{\tilde C'}{f(n)},
\end{aligned}$$

where the one before last inequality relies on a union bound with (5) and $\mathbb{P}_\theta(A_t) = a^t$, and the last inequality uses the definition of n. We have thus proved that (6) holds, and thus the policy cannot have the f-upper tailed property simultaneously in environment θ and $\tilde\theta$.

References

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

[Agr95] 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.

[AMS09] 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.

[BMSS09] 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.

[BSS09] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms: extended abstract. In *Proceedings of the tenth ACM conference on Electronic commerce*, pages 79–88. ACM, 2009.

[BV08] D. Bergemann and J. Valimaki. Bandit problems. 2008. In The New Palgrave Dictionary of Economics, 2nd ed. Macmillan Press.

[CM07] P.A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Uncertainty in Artificial Intelligence*, 2007.

[DK09] N.R. Devanur and S.M. Kakade. The price of truthfulness for pay-per-click auctions. In *Proceedings of the tenth ACM conference on Electronic commerce*, pages 99–106. ACM, 2009.

[GW06] 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.

[Hol92] J.H. Holland. *Adaptation in natural and artificial systems*. MIT press Cambridge, MA, 1992.

[Kle05] R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704. 2005.

[KS06] L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML-2006)*, pages 282–293, 2006.

[KSU08] 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.

[LPT04] 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.

[LR85] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[Rob52] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

[Rud86] W. Rudin. *Real and complex analysis (3rd)*. New York: McGraw-Hill Inc, 1986.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.