Robustness of Anytime Bandit Policies

Antoine Salomon, Jean-Yves Audibert

Robustness of anytime bandit policies

Antoine Salomon
Imagine
École des Ponts ParisTech
Université Paris Est
salomona@imagine.enpc.fr

Jean-Yves Audibert
Imagine, Université Paris Est
&
Sierra, CNRS/ENS/INRIA, Paris, France
audibert@imagine.enpc.fr

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 empirically 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 get by always drawing the best arm and the cumulative reward he actually gets. 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 (Babaioff et al., 2009, Devanur and Kakade, 2009), economics (Bergemann and Valimaki, 2008, Lamberton et al., 2004), games (Gelly and Wang, 2006) and optimization (Kleinberg, 2005, Coquelin and Munos, 2007, Kleinberg et al., 2008, Bubeck et al., 2009). It can be a central building block of larger systems, like in evolutionary programming (Holland, 1992) and reinforcement learning (Sutton and Barto, 1998), in particular in large state space Markovian Decision Problems (Kocsis and Szepesvári, 2006). 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 (Robbins, 1952, Lai and Robbins, 1985, Agrawal, 1995, Auer et al., 2002, 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 Audibert et al. (2009). The two main results there about the deviation 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). 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 Audibert et al. (2009) is whether it is possible to design an anytime robust policy, that is a policy for which for any n, with probability at least $1-1/n$, its 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 3.4). 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 4.3).

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 $\theta = (\nu_1, \ldots, \nu_K)$. 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 rounds 1 to t, and by $X_{k,1}, X_{k,2}, \ldots, X_{k,T_k(t)}$ the sequence of associated rewards. For an environment parameterized by $\theta = (\nu_1, \ldots, \nu_K)$, let $\mathbb{P}_\theta$ denote the distribution on the probability space such that for any $k \in \{1, \ldots, K\}$, the random variables $X_{k,1}, X_{k,2}, \ldots$ 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. Introduce $\mu^* = \max_{k\in\{1,\ldots,K\}} \mu_k$ and $k^* \in \operatorname{argmax}_{k\in\{1,\ldots,K\}} \mu_k$, 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 get by always drawing the best arm and the cumulative reward he actually gets. At time $n \geq 1$, its regret is thus

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

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

$$\mathbb{E}_\theta \hat{R}_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}_\theta[T_k(n)]. \tag{1}$$

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 in the order of $\mathbb{E}_\theta \hat{R}_n$. If a policy has small deviations, it means that the risks involved by its decisions are smaller, and also, as our simulations will demonstrate it ***** to be done *****, that the expectation of the regret tends to be smaller. This can also be explained by the formula:

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

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 $\Delta = \min_{k\neq k^*} \Delta_k$ the gap between the best arm and second best arm. Note that the case $\Delta = 0$ is degenerated as the sampling times of $T_k(n)$ for $k \neq k^*$ such that $\mu_k = \mu_{k^*}$ will in general no longer be logarithmic in n, and the definition of the regret is then ambiguous since k^* is not unique.

Definition 1 (f-T and f-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-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-R) if and only if

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

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 $\widehat{X}_{k,s} = \frac{1}{s} \sum_{t=1}^{s} X_{k,t}$ 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

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

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, Auer et al. (2002) proved:

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

Lai and Robbins (1985) showed that these results cannot be improved up to numerical constants. Audibert et al. (2009) proved that UCB1 is $\log^3$-T and $\log^3$-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 $\log^{2\rho-1}$-T and $\log^{2\rho-1}$-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. Up to a change in the former definitions (n fixed, no "$\forall n \geq 2$"), Audibert et al. showed that UCB-H is Id-T and Id-R where Id is the identity function.

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-wT and f-wR) *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-wT) if and only if*

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

A policy has weak f-upper tailed Regret (in short, f-wR) 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-T and f-wT (and between f-R and f-wR) is the interchange of "$\forall \theta$" and "$\exists C, \tilde{C}$". Consequently, a policy that is f-T (respectively f-R) is f-wT (respectively f-wR). Let us detail the links between the f-T, f-R, f-wT and f-wR.

Proposition 2.1 *Assume that there exists $\alpha, \beta > 0$ such that $f(n) \leq \alpha n^\beta$ for any $n \geq 2$. We have*

$$f\text{-}T \Rightarrow f\text{-}R \Rightarrow f\text{-}wR \Leftrightarrow f\text{-}wT.$$

3 Impossibility result

In the previous section, we have mentioned that for any $\alpha > 0$, there is a variant of UCB1 (obtained by changing $2 \log t$ into $\frac{1+\alpha}{2} \log t$ in (2)) which is $\log^\alpha$-T, and hence $\log^\alpha$-R. The following result shows that it is impossible to find a policy that could be more robust than these policies. For many usual settings (e.g., when Θ is the set $\bar{\Theta}$ of all K-tuples of measures on $[0, 1]$), the agent is too easily 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.

The forecaster has to choose arm k often enough, in case the current environment were θ. As arm k delivers payoffs according to the same law in both environments, these payoffs do not help to distinguish $\tilde{\theta}$ from θ at all. The other arms can help to point out the difference, but they are not chosen often enough. This is in fact this kind of situation that have to be taken into account when balancing a policy between exploitation and exploration.

Before stating our main result, which formalizes the leads given above, let us detail the third condition. To this aim, let us remind the following result.

Theorem 3.1 (Lebesgue-Radon-Nikodym theorem) *Let μ_1 and μ_2 be σ-finite measures. 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*[1] *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 μ_2-negligible event.

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

Lemma 3.2 *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:

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

■

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 an payoff x from arm ℓ, the quantity $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(x)$ represents how much the observation of x makes environment θ more likely than $\tilde{\theta}$. 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. 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 $\tilde{\theta}$.

Theorem 3.3 *Let $f : \mathbb{N} \to \mathbb{R}_+^*$ be greater than order $\log^\alpha$, that is for any $\alpha > 0$, $f \gg_{+\infty} \log^\alpha$. 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 $\tilde{\theta}$,

[1] Two measures μ_1 and μ_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 $A_1 \cup A_2 = \Omega$, $\mu_1(A_2) = 0$ and $\mu_2(A_1) = 0$.

(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-wT policy, and hence no f-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-wT, and in particular that would "work well" in environment θ in the sense given by the definition of f-wT. The proof exhibits a time 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) 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-wT, 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 constant 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 of order N times at least has therefore a 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 condition given in theorem 3.3 are not very restrictive. The impossibility holds for very basic settings, and may hold even if the agent has great knowledges of the possible environments. For instance, the setting

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

satisfies the tree conditions of the theorem.
Nevertheless, the main interest of the result regarding the previous literature is the following corollary.

Corollary 3.4 *If Θ is the whole set $\bar{\Theta}$ of all K-tuples of measures on $[0,1]$, then there is no f-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 of the following result for UCB-H which, for a given n, plays at time $t \geq K+1$,

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

Theorem 3.5 *For any $\beta > 0$,* UCB-H *satisfies*

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

Of course, $n^\beta \gg_{n \to +\infty} \log^\alpha(n)$ for all $\alpha, \beta > 0$ but this does not contradict our theorem, since we are dealing with *anytime* policies. UCB-H will work fine if n is known in advance, but may do terrible at other rounds. In particular and as any policy, it can not achieve anytime polynomial regret concentration.

Corollary 3.4 should also be read in conjunction of the following result for the policy UCB1(ρ) which plays at time $t \geq K+1$,

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

Theorem 3.6 *For any $\rho > 1/2$,* UCB1(ρ) *is $\log^{2\rho-1}$-R.*

Thus, any improvements of existing algorithms which would for instance involve estimations of variance (see Audibert et al. (2009)), of Δ_k, or of many characteristics of the distributions cannot beat the variants of UCB1 regarding deviations. One may at best improve constants $C, \tilde{C}$, and this is equivalent to changing f into f^β for a given $\beta > 1$. Nevertheless, one can not improve f, i.e. find a better one, $\tilde{f}$, such that $f/\tilde{f} \to_{+\infty} 0$.

Let us denote $\Theta_k = \{\theta \in \Theta | k \text{ is the optimal arm in } \theta\}$.

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

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

Figure 1: c.d.f.-based algorithm.

4 Positive results

The intuition behind Theorem 3.3 suggests that, if one of the three conditions 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. Thus, the agent chooses an arm that fits better a winning distribution ν_k. He can not 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 $\tilde{\theta}$ because, during the first rounds, he pulls every arms and at least one of them will not behave as if in θ if the current environment is $\tilde{\theta}$.
Nevertheless, such a policy cannot work in general:

- If $\tilde{\theta}$ is the current environment and even if the agent has identified θ as impossible, there still could be other environments that are arbitrary close to θ in which arm k is optimal and which the agent is not able to distinguish from $\tilde{\theta}$.
- 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 3.2). If the set of all environments Θ is not discountable, 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 contrary, the agent could legitimately consider an environment θ as unlikely if, for $\varepsilon > 0$ small enough, there exists $\tilde{\theta}$ such that $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) \leq \varepsilon$.[2] The former criterion only consider as unlikely an environment θ when there exists $\tilde{\theta}$ such that $\frac{d\nu_k}{d\tilde{\nu}_k}(X_{k,1}) = 0$.

In this section we give sufficient conditions on Θ for such a policy to be robust, and this is equivalent to finding conditions under which the converse of Theorem 3.3 holds. We estimate distributions of each arm by means of their empirical cumulative distribution functions, and distance between two c.d.f. is measured thanks to the norm $\|.\|_\infty$, defined by $\|F\|_\infty = \sup_{[0,1]} |F|$. 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 such that there is still an arm k and a winning distribution ν_k such that F_{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 a c.d.f. based algorithm described in Figure 1.

[2] Note that an algorithm that includes this ability would be hard to balance.

Let us denote $\Theta_k = \{\theta \in \Theta | k \text{ is the optimal arm in } \theta\}$.

Proceed as follows:

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

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

where $\hat{X}_{k,t}$ is the empirical mean of arm of k after having been pulled t times.

Figure 2: c.d.f.-based algorithm in case of Bernoulli laws.

4.1 Θ is finite

When Θ is finite, none of the three limitations presented above holds, so that the converse of Theorem 3.3 is true and our algorithm is robust.

Theorem 4.1 *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 $k \in \{1, \dots, K\}$, 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 the c.d.f. based algorithm is Id^β-T (and hence Id^β-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 (whose parameter is μ_k), and that there exists $\gamma \in (0, 1)$ such that $\mu_k \in [\gamma, 1]$ for all k and all θ.[3] 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, the event $\left\{\forall \tilde{\theta} \in \Theta,\ \forall \ell \in \{1, \cdots, K\},\ \frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\right\}$ is $\mathbb{P}_\theta$-a.s. for all θ so that the impossibility result only relies on conditions (a) and (b) of Theorem 3.3. This theorem can be modified to cover any settings, and our algorithm can be made simpler (see Figure 2).

Theorem 4.2 *For any $\theta \in \Theta$ and any $k \in \{1, \dots, K\}$, let us set*

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

Then c.d.f.-based algorithm is such that:

$$\forall \beta > 0,\ \exists C, \tilde{C} > 0, \forall \theta \in \Theta,\ \forall n \geq 1,\ \forall k \in \{1, \dots, K\},\ \mathbb{P}_\theta\left(T_k(n) \geq \frac{C \log n}{d_k^2}\right) \leq \frac{\tilde{C}}{n^\beta}.$$

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

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

then there is no policy such that:

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

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

Proceed as follows:

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

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

where $\hat{X}_{k,t}$ is the empirical mean of arm of k after having been pulled t times.

Figure 3: Variant of c.d.f.-based algorithm when μ^* is known.

Note that we do not adopt the former definitions of robustness ($f-R$ and $f-T$), because the significant term here is d_k (and not Δ_k), 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 impossibility theorem 3.3, 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. The first and second conditions of Theorem 3.3 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 greater 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 thanks to Hoeffding's inequality instead of Massart's one.

Theorem 4.3 *When μ^* is known, the variant of the c.d.f.-based algorithm is Id^β for all $\beta > 0$.*

5 Proofs

5.1 Proof of Proposition 2.1

f-T $\Rightarrow$ f-R: When a policy is f-T, by a union bound, the event

$$\xi_1 = \left\{ \exists k \in \{1,\dots,K\},\ T_k(n) \geq C\frac{\log n}{\Delta_k^2} \right\}$$

occurs with probability at most $\frac{K\tilde{C}}{f(n)}$. Introduce $S_{k,s} = \sum_{t=1}^{s}(X_{k,t} - \mu_k)$. Since we have

$$\sum_{t=1}^{n} X_{I_t,T_{I_t}(t)} = \sum_{k=1}^{K} S_{k,T_k(n)} + \sum_{k=1}^{K} T_k(n)\mu_k,$$

we have

$$\hat{R}_n = S_{k^*,n} - S_{k^*,T_{k^*}(n)} - \sum_{k\neq k^*} S_{k,T_k(n)} + \sum_{k\neq k^*} \Delta_k T_k(n). \tag{3}$$

Let $T = \sum_{k\neq k^*} T_k(n) = n - T_{k^*}(n)$, $t^* = \sum_{k\neq k^*} C\frac{\log n}{\Delta_k^2}$, and $W = \max_{0\leq s\leq t^*}(S_{k^*,n} - S_{k^*,n-s})$. Since $S_{k^*,n} - S_{k^*,T_{k^*}(n)} \leq W$ on the complement ξ_1^c of ξ_1, we have

$$\hat{R}_n \leq n\mathbb{1}_{\xi_1} + W - \sum_{k\neq k^*} S_{k,T_k(n)} + \sum_{k\neq k^*} \Delta_k T_k(n). \tag{4}$$

Consider the events

$$\xi_2 = \left\{ W > \sum_{k \neq k^*} \sqrt{\frac{C\beta}{2}} \frac{\log n}{\Delta_k} \right\},$$

$$\xi_{3,k} = \left\{ \max_{1 \leq s \leq \frac{C \log n}{\Delta_k^2}} (-S_{k,s}) > \sqrt{\frac{C\beta}{2}} \frac{\log n}{\Delta_k} \right\},$$

and

$$\xi = \xi_1 \cup \xi_2 \underset{k \neq k^*}{\cup} \xi_{3,k}.$$

From Hoeffding's maximal inequality, we have

$$\mathbb{P}_\theta(\xi_2) \leq \exp\left(- \frac{2\big(\sum_{k \neq k^*} \sqrt{C\beta/2} \frac{\log n}{\Delta_k}\big)^2}{\sum_{k \neq k^*} (C \log n)/\Delta_k^2} \right) \leq \exp\big(- \beta \log n \big) = \frac{1}{n^\beta} \leq \frac{\alpha}{f(n)}.$$

We also use Hoeffding's maximal inequality to control $\mathbb{P}_\theta(\xi_{3,k})$:

$$\mathbb{P}_\theta(\xi_{3,k}) \leq \exp\left(- \frac{2\big(\sqrt{C\beta/2} \frac{\log n}{\Delta_k}\big)^2}{(C \log n)/\Delta_k^2} \right) = \frac{1}{n^\beta} \leq \frac{\alpha}{f(n)}.$$

By gathering the previous results using a union bound, we have $\mathbb{P}(\xi) \leq \frac{2\alpha + \tilde{C}}{f(n)}$. Besides on the complement of ξ, by using (4), we have

$$\hat{R}_n < \sum_{k \neq k^*} \sqrt{\frac{C\beta}{2}} \frac{\log n}{\Delta_k} + \sum_{k \neq k^*} \sqrt{\frac{C\beta}{2}} \frac{\log n}{\Delta_k} + \sum_{k \neq k^*} \frac{C \log n}{\Delta_k}.$$

We have thus proved that

$$\forall \theta \in \Theta, \forall n \geq 1,\ \mathbb{P}_\theta \left(\hat{R}_n \geq (C + \sqrt{2C\beta}) \frac{\log n}{\Delta} \right) \leq \frac{\tilde{C} + 2\alpha}{f(n)},$$

hence the policy is f-R.

f-wT $\Rightarrow$ f-wR: it is exactly the same proof as for f-T $\Rightarrow$ f-R since the core of the argument is independent of the position of "$\forall \theta$" with respect to "$\exists C, \tilde{C}$".

f-wR $\Rightarrow$ f-wT: let us prove the contrapositive. So we assume

$$\exists \theta \in \Theta \text{ such that } \Delta \neq 0,\ \forall C', \tilde{C}' > 0,\ \exists n \geq 1,\ \exists k \neq k^*,\ \mathbb{P}_\theta \left(T_k(n) \geq C' \frac{\log n}{\Delta_k^2} \right) > \frac{\tilde{C}'}{f(n)}. \tag{5}$$

It is enough to prove that for this θ, we have

$$\forall C > 9K/\Delta,\ \forall \tilde{C} > \alpha,\ \exists n \geq 1, \mathbb{P}_\theta \left(\hat{R}_n \geq C \frac{\log n}{\Delta} \right) > \frac{\tilde{C}}{f(n)}.$$

To achieve this, we consider $C' = (\beta + 2)C/\Delta$ and $\tilde{C}' = \max\big(2\tilde{C}, \max_{m \leq K} f(m)\big)$ in (5) and let $k' \neq k^*$ be such that the event

$$\xi' = \left\{ T_{k'}(n) \geq C' \frac{\log n}{\Delta_{k'}^2} \right\}$$

holds with probability greater than $\tilde{C}'/f(n) = 2\tilde{C}/f(n)$. From (5) and using $\tilde{C}' \geq \max_{m \leq K} f(m)$, we necessarily have $n \geq K$. Let $L = \log\big(\frac{f(n)}{\tilde{C}} nK\big)$ and

$$\xi'' = \left\{ \forall k \neq k^*,\ \forall s \in \{1, \dots, n\},\ |S_{k,s}| \leq \sqrt{\frac{sL}{2}} \right\} \bigcap \left\{ \forall s \in \{1, \dots, n\},\ |S_{k^*,n} - S_{k^*,n-s}| \leq \sqrt{\frac{sL}{2}} \right\}.$$

By Hoeffding's inequality and a union bound, this event holds with probability at least $1 - \tilde{C}/f(n)$. As a consequence, we have $\mathbb{P}(\xi' \cap \xi'') > \tilde{C}/f(n)$. We now prove that on the event $\xi' \cap \xi''$, we have

$$\hat{R}_n \geq C \frac{\log n}{\Delta}.$$

First note that for any $a > 0$ the function $s \mapsto as - \sqrt{2sL}$ is decreasing on $\left[0, \frac{L}{2a^2}\right]$ and increasing on $\left[\frac{L}{2a^2}, +\infty\right)$, and that

$$T_{k'}(n) \geq C' \frac{\log n}{\Delta_{k'}^2} \geq \frac{CL}{\Delta_{k'}^2},$$

since $\frac{f(n)}{C} nK \leq \frac{\alpha n^\beta}{\alpha} n^2 = n^{\beta+2} \leq n^{C'/C}$. Then, by using (3) and $T_{k^*}(n) = n - \sum_{k \neq k^*} T_k(n)$, we have

$$\begin{aligned}
\hat{R}_n &\geq -|S_{k^*,n} - S_{k^*,T_{k^*}(n)}| - \sum_{k \neq k^*} |S_{k,T_k(n)}| + \sum_{k \neq k^*} \Delta_k T_k(n) \\
&\geq -\sqrt{\frac{L \sum_{k \neq k^*} T_k(n)}{2}} - \sum_{k \neq k^*} \sqrt{\frac{L T_k(n)}{2}} + \sum_{k \neq k^*} \Delta_k T_k(n) \\
&\geq \sum_{k \neq k^*} \left(\Delta_k T_k(n) - \sqrt{2 T_k(n) L} \right) \\
&\geq \frac{\Delta_{k'} T_{k'}(n)}{2} + \left(\frac{\Delta_{k'} T_{k'}(n)}{2} - \sqrt{2 L T_{k'}(n)} \right) + \sum_{k \neq k^*, k \neq k'} \min_{s \geq 1} \left(\Delta_k s - \sqrt{2Ls} \right) \\
&\geq C' \frac{\log n}{2\Delta_{k'}} + \left(\frac{C}{2} - \sqrt{2C} \right) \frac{L}{\Delta_{k'}} - \sum_{k \neq k^*, k \neq k'} \frac{L}{2\Delta_k} \\
&\geq C' \frac{\log n}{2\Delta_{k'}} + \frac{C}{6} \frac{L}{\Delta_{k'}} - \frac{KL}{2\Delta} \geq C' \frac{\log n}{2\Delta_{k'}} \geq C \frac{\log n}{\Delta},
\end{aligned}$$

which ends the proof of the contrapositive.

5.2 Proof of Theorem 3.3

Let us first notice that a policy is f-wT if and only if

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

This means that we can remove the Δ_ℓ denominator without altering the definition of f-wT. Note that this would not be possible for the f-T definition owing to the different position of "$\forall \theta$" with respect to "$\exists C, \tilde{C}$".

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

$$\forall N \geq 2,\ \forall \ell \neq k,\ \mathbb{P}_\theta\big(T_\ell(N) \geq C \log N\big) \leq \frac{\tilde{C}}{f(N)}. \tag{6}$$

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 $C', \tilde{C}' > 0$

$$\exists n \geq 2,\ \mathbb{P}_{\tilde{\theta}}\big(T_k(n) \geq C' \log n\big) > \frac{\tilde{C}'}{f(n)}. \tag{7}$$

since k is suboptimal in environment $\tilde{\theta}$. Note that proving (7) for $C' = C$ is sufficient. Indeed if (7) holds for $C' = C$, it a fortiori holds for $C' < C$. Besides, when $C' > C$, (6) 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 3.2, $\mathbb{P}_{\tilde{\theta}}\big(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\big) > 0$ is equivalent to $\mathbb{P}_\theta\big(\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) > 0\big) > 0$. By independence of $X_{1,1}, \ldots, X_{K,1}$ 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 $a = \mathbb{P}_\theta\left(\prod_{\ell\neq k}\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \eta\right)$.

Let us take n large enough such that $N = \lfloor 4C\log n\rfloor$ satisfies $N < n$, $C\log N < \frac{N}{2}$ 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 $A_t = \cap_{s=1\ldots t}\left\{\prod_{\ell\neq k}\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,s}) \geq \eta\right\}$. By independence and definition of a, we have $\mathbb{P}_\theta(A_t) = a^t$. We have

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

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 $\tilde{\nu}_k = \nu_k$ and by definition of A_t, 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 step relies on a union bound with (6) and $\mathbb{P}_\theta(A_t) = a^t$, and the last inequality uses the definition of n. We have thus proved that (7) holds, and thus the policy cannot have the f-upper tailed property simultaneously in environment θ and $\tilde{\theta}$.

5.3 Proof of theorem 4.1

Let θ be in Θ. Consider the event

$$\xi = \left\{ \forall k \in \{1, \dots, K\}, T \in \{1, \dots, n\}, T\|\hat{F}_{k,T} - F_{\nu_k}\|_\infty^2 < \frac{\beta+1}{2}\log n \right\}.$$

From Massart's inequality (see Massart (1990)) applied nK times corresponding to the different times and arms and a union bound to combine the inequalities, we have

$$\mathbb{P}_\theta(\xi) \geq 1 - nK(2e^{-(\beta+1)\log n}) = 1 - \frac{2K}{n^\beta}.$$

We show that on the event ξ, inequalities $T_k(n) \leq \frac{2(\beta+1)\log n}{\delta_k^2} + 1$ hold for any $k \neq k^*$, where $\delta_k = \min_{\tilde{\theta} \in \Theta_k} \|F_{\nu_k} - F_{\tilde{\nu}_k}\|_\infty$. Note that $\delta_k > 0$: if not, it would mean that k is suboptimal in θ and optimal in an other environment $\tilde{\theta}$, with $\nu_k = \tilde{\nu}_k$. In this case, by hypothesis there exists $\ell \neq k$ such that $\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) = 0$ $\mathbb{P}_\theta$-a.s. Thus $\tilde{\theta}$ is almost surely removed during the first rounds of the policy and, as Θ is finite, all of these problematic $\tilde{\theta}$ are removed almost surely. Note also that θ cannot be removed: it is readily seen that $\mathbb{P}_\theta\left(\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0\right) = 1$ for all $\tilde{\theta} \in \Theta$ and, still because Θ is finite, it is almost sure that $\frac{d\nu_\ell}{d\tilde{\nu}_\ell}(X_{\ell,1}) > 0$ for all $\tilde{\theta} \in \Theta$. A last consequence of the finiteness of Θ is that terms δ_k are uniformly bounded away from zero over Θ, and so are the terms Δ_k, so that the inequalities we are going to prove easily lead to the conclusion of the proof.

Assume by contradiction that there exists $k \neq k^*$ such that $T_k(n) > \frac{2(\beta+1)\log n}{\delta_k^2} + 1$. Then there exists $t \leq n$ such that $I_t = k$ and $T_k(t-1) > \frac{2(\beta+1)\log n}{\delta_k^2}$.
As arm k is chosen at round t, we have:

$$T_{k^*}(t-1) \inf_{\tilde{\theta} \in \Theta_{k^*}} \|\hat{F}_{k^*,T_k^*(t-1)} - F_{\tilde{\nu}_{k^*}}\|_\infty^2 \geq T_k(t-1) \inf_{\tilde{\theta} \in \Theta_k} \|\hat{F}_{k,T_k(t-1)} - F_{\tilde{\nu}_k}\|_\infty^2$$

On the one hand, we have:

$$\frac{\beta+1}{2}\log n > T_{k^*}(t-1) \inf_{\tilde{\theta} \in \Theta_{k^*}} \|\hat{F}_{k^*,T_k^*(t-1)} - F_{\tilde{\nu}_{k^*}}\|_\infty^2,$$

and on the other hand

$$\begin{aligned}
\sqrt{T_k(t-1)} \inf_{\tilde{\theta} \in \Theta_k} \|\hat{F}_{k,T_k(t-1)} - F_{\tilde{\nu}_k}\|_\infty &\geq \sqrt{T_k(t-1)}\left(\delta_k - \|\hat{F}_{k,T_k(t-1)} - F_{\nu_k}\|_\infty\right) \\
&\geq \sqrt{T_k(t-1)}\left(\delta_k - \sqrt{\frac{(\beta+1)\log n}{2T_k(t-1)}}\right) \\
&= \sqrt{T_k(t-1)}\delta_k - \sqrt{\frac{\beta+1}{2}\log n}.
\end{aligned}$$

By combining the former inequalities, we get:

$$\sqrt{\frac{\beta+1}{2}\log n} > \sqrt{T_k(t-1)}\delta_k - \sqrt{\frac{\beta+1}{2}\log n}$$

and

$$T_k(t-1) < \frac{2(\beta+1)\log n}{\delta_k^2},$$

which is the contradiction expected.

5.4 Proof of Theorem 4.2

The proof of the first part of the theorem is the same as the previous section, except that one has to substitute δ_k by d_k and that the d_k ($k \neq k^*$) are not necessarily non negative. Indeed, the distance $\|\hat{F}_{k,T} - F_{\nu_k}\|_\infty$ equals $|\hat{X}_{k,T} - \mu_k|$ in the context of Bernoulli laws.

The proof of the second part is similar to the one of Theorem 3.3: we assume by contradiction that there exists a policy such that

$$\exists C, \tilde{C} > 0, \forall \theta \in \Theta,\ \forall n \geq 1,\ \forall k \neq k^*,\ \mathbb{P}_\theta \left(T_k(n) \geq C \log n \right) \leq \frac{\tilde{C}}{f(n)}.$$

The main difference is that we cannot fix $\theta, \tilde{\theta}$ such that $\theta \in \Theta_k$, $\tilde{\theta} \in \Theta \setminus \Theta_k$ and $\mu_k = \tilde{\mu}_k$. The hypothesis only allows us to take μ_k and $\tilde{\mu}_k$ arbitrarily close. This means that we are allowed to consider two sequences $(\theta^n)_{n\geq 1}$ and $(\tilde{\theta}^n)_{n\geq 1}$ such that, for all $n \geq 1$ (with obvious notations):

- $\theta^n \in \Theta_k,\ \tilde{\theta}^n \in \Theta \setminus \Theta_k$,
- $\tilde{\mu}_k^n \geq 2^{-\frac{1}{N}} \mu_k^n$,
- $1 - \tilde{\mu}_k^n \geq 2^{-\frac{1}{N}} (1 - \mu_k^n)$,

where $N = \lfloor 4C \log n \rfloor$.

On the other hand, the hypothesis readily implies that

$$\forall \theta, \tilde{\theta} \in \Theta,\ \forall \ell \in \{1, \cdots, K\},\ \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(1) = \frac{\tilde{\mu}_l}{\mu_l} \geq \gamma$$

and

$$\begin{aligned}
\mathbb{P}_\theta \left(\prod_{\ell \neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \gamma^{K-1} \right) &\geq \mathbb{P}_\theta \left(\bigcap_{\ell \neq k} \left\{ \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \gamma \right\} \right) = \prod_{\ell \neq k} \mathbb{P}_\theta \left(\frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,1}) \geq \gamma \right) \\
&\geq \prod_{\ell \neq k} \mathbb{P}_\theta \left(X_{\ell,1} = 1 \right) = \prod_{\ell \neq k} \mu_l \geq \gamma^{K-1}.
\end{aligned}$$

Let us denote $a = \gamma^{K-1}$ and $A_t = \bigcap_{s=1}^t \left\{ \prod_{\ell \neq k} \frac{d\tilde{\nu}_\ell}{d\nu_\ell}(X_{\ell,s}) \geq a \right\}$. By independence, we have $\mathbb{P}_\theta(A_t) = a^t$.

To find a contradiction, we set $t = \lfloor C \log N \rfloor$ and we adapt the reasoning of the former proof. If n is chosen large enough, one has $N < n$ and $C \log N < \frac{N}{2}$, and then:

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

Let us denote $B_N = \bigcap_{\ell \neq k} \left\{ T_\ell(N) < C \log N \right\}$. B_N is measurable w.r.t. $X_{k,1}, \ldots, X_{k,N}$ and $X_{\ell,1}, \ldots, X_{\ell,t}$ $(\ell \neq k)$, and A_t is measurable w.r.t. $X_{\ell,1}, \ldots, X_{\ell,t}$ $(\ell \neq k)$, so that we may write

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

By properties of $\tilde{\nu}_k^n$ and ν_k^n and by definition of A_t we have

$$\begin{aligned}
&\mathbb{P}_{\tilde{\theta}^n}\left(A_t \cap \left\{\bigcap_{\ell\neq k}\{T_\ell(N) < C\log N\}\right\}\right)\\
&= \int c_{t,N}\left((x_{\ell,s})_{\ell\neq k,\ s=1..t}, (x_{k,s})_{s=1..N}\right) \prod_{\substack{\ell\neq k\\ s=1..t}} d\tilde{\nu}_\ell^n(x_{\ell,s}) \prod_{s=1..N} d\tilde{\nu}_k^n(x_{k,s})\\
&\geq \int c_{t,N}\left((x_{\ell,s})_{\ell\neq k,\ s=1..t}, (x_{k,s})_{s=1..N}\right) a^t \prod_{\substack{\ell\neq k\\ s=1..t}} d\nu_\ell^n(x_{\ell,s}) \prod_{s=1..N}\left(2^{-\frac{1}{N}} d\nu_k^n(x_{k,s})\right)\\
&= \frac{a^t}{2}\mathbb{P}_{\theta^n}\left(A_t \cap \left\{\bigcap_{\ell\neq k}\{T_\ell(N) < C\log N\}\right\}\right)\\
&\geq \frac{a^t}{2}\left(a^t - \frac{(K-1)\tilde{C}}{f(N)}\right).
\end{aligned}$$

By straightforward calculations, one can then show that $f(n)\mathbb{P}_{\tilde{\theta}^n}\left(T_k(n) \geq C\log n\right) \xrightarrow[N\to+\infty]{} +\infty$, which is the contradiction expected.

5.5 Proof of Theorem 4.3

The proof is similar the one of Theorem 4.1, except that we use Hoeffding's inequality rather than Massart's one. Consider the event

$$\xi = \left\{\forall k \in \{1,\dots,K\}, s \in \{1,\dots,n\}, s(\hat{X}_{k,s} - \mu_k)^2 < \frac{\beta+1}{2}\log n\right\}.$$

From Hoeffding's inequality applied $2nK$ times corresponding to the different times and arms and a union bound to combine the inequalities, we have $\mathbb{P}(\xi) \geq 1 - 2nKe^{-(\beta+1)\log n} = 1 - \frac{2K}{n^\beta}$. We will prove by contradiction that on the event ξ, we have $T_k(n) \leq 1 + \frac{2(\beta+1)\log n}{\Delta_k^2}$ for all k. For this, consider k such that $T_k(n) > \frac{2(\beta+1)\log n}{\Delta_k^2} + 1$. Then there exists $t \leq n$ such that $I_t = k$ and $T_k(t-1) > \frac{2(\beta+1)\log n}{\Delta_k^2}$. Since the arm k is chosen at time t, it means that

$$T_k(t-1)\big(\hat{X}_{k,T_k(t-1)} - \mu^*\big)^2 \leq T_{k^*}(t-1)\big(\hat{X}_{k^*,T_{k^*}(t-1)} - \mu^*\big)^2.$$

On the one hand, we have:

$$\frac{\beta+1}{2}\log n > T_{k^*}(t-1)\big(\hat{X}_{k^*,T_{k^*}(t-1)} - \mu^*\big)^2,$$

and on the other hand

$$\begin{aligned}
\sqrt{T_k(t-1)}\big|\hat{X}_{k,T_k(t-1)} - \mu^*\big| &\geq \sqrt{T_k(t-1)}\left(\Delta_k - \big|\hat{X}_{k,T_k(t-1)} - \mu_k\big|\right)\\
&\geq \sqrt{T_k(t-1)}\left(\Delta_k - \sqrt{\frac{(\beta+1)\log n}{2T_k(t-1)}}\right)\\
&= \sqrt{T_k(t-1)}\Delta_k - \sqrt{\frac{\beta+1}{2}\log n}.
\end{aligned}$$

The former inequalities leads to

$$\sqrt{\frac{\beta+1}{2}\log n} > \sqrt{T_k(t-1)}\Delta_k - \sqrt{\frac{\beta+1}{2}\log n} \Rightarrow T_k(t-1) < \frac{2(\beta+1)\log n}{\Delta_k^2}.$$

Thus there is a contradiction, meaning that there is no k such that $T_k(n) > \frac{2(\beta+1)\log n}{\Delta_k^2} + 1$.

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.

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. ISSN 0304-3975.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125.

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.

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

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.

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

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.

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

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.

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.

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.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. ISSN 0091-1798.

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

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