On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems

Aurélien Garivier, Eric Moulines

On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems

Aurélien Garivier AURELIEN.GARIVIER@TELECOM-PARISTECH.FR
Institut Telecom / Telecom ParisTech / Laboratoire LTCI / CNRS UMR 5141
46 rue Barrault, 75634 Paris Cedex 13

Eric Moulines ERIC.MOULINES@TELECOM-PARISTECH.FR
Institut Telecom / Telecom ParisTech / Laboratoire LTCI / CNRS UMR 5141
46 rue Barrault, 75634 Paris Cedex 13

Editor: unknown

Abstract

Multi-armed bandit problems are considered as a paradigm of the trade-off between exploring the environment to find profitable actions and exploiting what is already known. In the stationary case, the distributions of the rewards do not change in time, *Upper-Confidence Bound* (UCB) policies, proposed in Agrawal (1995) and later analyzed in Auer et al. (2002), have been shown to be rate optimal.

A challenging variant of the MABP is the non-stationary bandit problem where the gambler must decide which arm to play while facing the possibility of a changing environment. In this paper, we consider the situation where the distributions of rewards remain constant over epochs and change at unknown time instants. We analyze two algorithms: the *discounted UCB* and the *sliding-window UCB*. We establish for these two algorithms an upper-bound for the expected regret by upper-bounding the expectation of the number of times a suboptimal arm is played. For that purpose, we derive a Hoeffding type inequality for self normalized deviations with a random number of summands. We establish a lower-bound for the regret in presence of abrupt changes in the arms reward distributions. We show that the discounted UCB and the sliding-window UCB both match the lower-bound up to a logarithmic factor.

Keywords: Multi-armed bandit, reinforcement learning, deviation inequalities, non-stationary environment

1. Introduction

Multi-armed bandit (MAB) problems, modelling allocation issues under uncertainty, are fundamental to stochastic decision theory. The archetypal MAB problem may be stated as follows: there is a bandit with K independent arms. At each time step, the player can play only one arm and receive a reward. In the stationary case, the distribution of the rewards are initially unknown, but are assumed to remain constant during all games. The player iteratively plays one action (pulls an arm) per round, observes the associated reward, and decides on the action for the next iteration. The goal of a MAB algorithm is to minimize the expected regret over T rounds, which is defined as the expectation of the difference between the total reward obtained by playing the best arm and the total reward obtained by using the algorithm (or *policy*). The minimization of the regret is achieved by balancing *exploitation*, the use of acquired information, with *exploration*, acquiring new information. If the player always plays the arm which he currently believes to be the best, he might miss to identify another arm having an actually higher expected reward. On the other hand, if the gambler explores too often the environment to find profitable actions, he will fail to accumulate as many rewards as he could. For several algorithms in the literature (e.g. Lai and Robbins (1985); Agrawal (1995)), as the number of plays T goes to infinity, the expected total reward asymptotically approaches that

of playing a policy with the highest expected reward, and the regret grows as the logarithm of T. More recently, finite-time bounds for the regret have been derived (see Auer et al. (2002); Audibert et al. (2007)).

Though the stationary formulation of the MABP allows to address exploration versus exploitation challenges in a intuitive and elegant way, it may fail to be adequate to model an evolving environment where the reward distributions undergo changes in time. As an example, in the cognitive medium radio access problem Lai et al. (2007), a user wishes to opportunistically exploit the availability of an empty channel in a multiple channels system; the reward is the availability of the channel, whose distribution is unknown to the user. Another application is real-time optimization of websites by targetting relevant content at individuals, and maximize the general interest by learning and serving the most popular content (such situations have been considered in the recent Exploration versus Exploitation (EvE) PASCAL challenge by Hartland et al. (2006), see also Koulouriotis and Xanthopoulos (2008) and the references therein). These examples illustrate the limitations of the stationary MAB models. The probability that a given channel is available is likely to change in time. The news stories a visitor of a website is most likely to be interested in vary in time.

To model such situations, we need to consider non-stationary MAB problems, where distributions of rewards may change in time. We show in the following that, as expected, policies tailored for the stationary case fail to track changes of the best arm. In this paper, we consider a particular non-stationary case where the distributions of the rewards undergo abrupt changes. We derive a lower-bound for the regret of any policy, and we analyze two algorithms: the Discounted UCB (Upper Confidence Bound) proposed by Koczis and Szepesvári and the Sliding Window UCB we introduce. We show that they are almost rate-optimal, as their regret almost matches a lower-bound.

1.1 The stationary MAB problem

At each time s, the player chooses an arm $I_s \in \{1, \dots, K\}$ to play according to a (deterministic or random) policy π based on the sequence of past plays and rewards, and obtains a reward $X_s(I_s)$[1]. The rewards $\{X_s(i)\}_{s \geq 1}$ for each arm $i \in \{1, \dots, K\}$ are modeled by a sequence of independent and indentically distributed (i.i.d.) random variables from a distribution unknown to the player. We denote by $\mu(i)$ the expectation of the reward $X_1(i)$.

The optimal (oracle) policy π^* consists in always playing the arm $i^* \in \{1, \dots, K\}$ with largest expected reward

$$\mu(*) = \max_{1 \leq i \leq K} \mu(i) \ , \quad i^* = \arg\max_{1 \leq i \leq K} \mu(i) \ .$$

The performance of a policy π is measured in terms of *regret* in the first T plays, which is defined as the expected difference between the total rewards collected by the optimal policy π^* (playing at each time instant the arm i^* with the highest expected reward) and the total rewards collected by the policy π.

Denote by $N_t(i) = \sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}$ the number of times arm i has been played in the t first games. The expected regret after T plays may be expressed as:

$$\mathbb{E}_\pi \left[\sum_{t=1}^{T} \{\mu(*) - \mu(I_t)\} \right] = \sum_{i \neq i^*} \{\mu(*) - \mu(i)\} \, \mathbb{E}_\pi \left[N_T(i) \right] ,$$

where $\mathbb{E}_\pi$ the expectation under policy π.

1. Note that we use here the convention that the reward after at time s if the i-th arm is played is supposed to be $X_s(i)$ and not the $N_s(i)$-th reward in the sequence of rewards for arm i, where $N_s(i)$ denotes the number of time the arm i has been played up to time s; while this convention makes no difference in the stationary case, because the distribution of the rewards are independent, it is meaningful in the non-stationary case, since the distribution of the arm *may change* even if the arm has not been played. These models can be seen as a special instance of the so-called *restless* bandit, proposed by Whittle (1988).

Obviously, bounding the expected regret after T plays essentially amounts to controlling the expected number of times a sub-optimal arm is played. In their seminal paper, Lai and Robbins (1985) consider stationary MAB problem, in which the distribution of rewards was taken from a one-dimensional parametric family (each being associated with a different value of the parameter, unknown to the player). They have proposed a policy achieving a logarithmic regret. Furthermore, they have established a lower-bound for the regret for policy satisfying an appropriately defined consistency condition, and show that their policy was asymptotically efficient. Later, the non-parametric context has been considered; several algorithms have been proposed, among which *softmax action selection* policies and *Upper-Confidence Bound* (UCB) policies.

Softmax methods are randomized policies where, at time t, the arm I_t is chosen at random by the player according to some probability distribution giving more weight to arms which have so-far performed well. The greedy action is given the highest selection probability, but all the others are ranked and weighted according to their accumulated rewards. The most common softmax action selection method uses a Gibbs, or Boltzman distribution. A prototypal example of softmax action selection is the so-called EXP3 policy (for *Exponential-weight algorithm for Exploration and Exploitation*), which has been introduced by Freund and Schapire (1997) for solving a worst-case sequential allocation problem and thouroughly examined as an instance of "prediction with limited feedback" problem in Chapter 6 of Cesa-Bianchi and Lugosi (2006) (see also Auer et al. (2002/03); Cesa-Bianchi and Lugosi (1999)).

UCB methods are deterministic policies extending the algorithm proposed by Lai and Robbins (1985) to a non-parametric context; they have been introduced and analyzed by Agrawal (1995). They consist in playing during the t-th round the arm i that maximizes the upper bound of a confidence interval for expected reward $\mu(i)$, which is constructed from the past observed rewards. The most popular, called UCB-1, relies on the upper-bound $\bar{X}_t(i) + c_t(i)$, where $\bar{X}_t(i) = (N_t(i))^{-1} \sum_{s=1}^{t} X_s(i) \mathbb{1}_{\{I_s=i\}}$ denotes the empirical mean, and $c_t(i)$ is a *padding function*. A standard choice is $c_t(i) = B\sqrt{\xi \log(t)/N_t(i)}$, where B is an upper-bound on the rewards and $\xi > 0$ is some appropriate constant. UCB-1 is defined in Algorithm 1.

Algorithm 1 UCB-1

for t from 1 to K, play arm $I_t = t$;
for t from $K+1$ to T, play arm

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(i) + c_t(i).$$

UCB-1 belongs to the family of "follow the perturbed leader" algorithms, and has proven to retain the optimal logarithmic rate (but with suboptimal constant). A finite-time analysis of this algorithm has been given in Auer et al. (2002); Auer (2002); Auer et al. (2002/03). Other types of padding functions are considered in Audibert et al. (2007).

1.2 The non-stationary MAB problem

In the non-stationary context, the rewards $\{X_s(i)\}_{s \ge 1}$ for arm i are modeled by a sequence of independent random variables from potentially different distributions (unknown to the user) which may vary across time. For each $s > 0$, we denote by $\mu_s(i)$ the expectation of the reward $X_s(i)$ for arm i. Likewise, let i_t^* be the arm with highest expected reward, denoted $\mu_t(*)$, at time t. The regret of a policy π is now defined as the expected difference between the total rewards collected by the optimal policy π^* (playing at each time instant the arm i_t^*) and the total rewards collected by the policy π. Note that, in this paper, the non-stationary regret is not defined with respect to the best arm on average, but with respect to a strategy tracking the best

arm at each step (this notion of regret is similar to the "regret against arbitrary strategies" introduced in Section 8 of Auer et al. (2002/03) for the non-stochastic bandit problem).

In this paper, we consider *abruptly changing environments*: the distributions of rewards remain constant during periods and change at unknown time instants called *breakpoints*. In the following, we denote by Υ_T the number of abrupt changes in the reward distributions that occur before time T. Another type of non-stationary MAB, where the distribution of rewards changes continuously, are considered in Slivkins and Upfal (2008).

Standard soft-max and UCB policies are not appropriate for abruptly changing environments:as stressed in Hartland et al. (2006), "empirical evidence shows that their Exploration versus Exploitation trade-off is not appropriate for abruptly changing environments". To address this problem, several methods have been proposed.

In the family of softmax action selection policies, Auer et al. (2002/03) and Cesa-Bianchi et al. (2006, 2008) have proposed an adaptation referred to as *EXP3.S* of the Fixed-Share algorithm, a computationally efficient variant of EXP3 called introduced by Herbster and Warmuth (1998) (see also (Cesa-Bianchi and Lugosi, 2006) and the references therein). Theorem 8.1 and Corollary 8.3 in Auer et al. (2002/03) state that when EXP3.S is tuned properly (which requires in particular that Υ_T is known in advance), the expected regret is upper-bounded as

$$\mathbb{E}_\pi\left[R_T\right] \leq 2\sqrt{\mathrm{e}-1}\sqrt{KT(\Upsilon_T \log(KT) + \mathrm{e})}\ .$$

Compared to the stationary case, such an upper-bound may seem deceiving: the rate $O(\sqrt{T\log T})$ is much larger than the $O(\log T)$ achievable in absence of changes. But actually, we prove in Section 4 that no policy can achieve an average regret smaller than $O(\sqrt{T})$ in the non-stationary case. Hence, EXP3.S matches the best achievable rate up to a factor $\sqrt{\log T}$. Moreover, by construction this algorithm can as well be used in an adversarial setup.

On the other hand, in the family of UCB policies, several attempts have been made; see for examples Slivkins and Upfal (2008) and Kocsis and Szepesvári (2006). In particular, Kocsis and Szepesvári (2006) have proposed an adaptation of the UCB policies that relies on a discount factor $\gamma \in (0,1)$. This policy constructs an UCB $\bar{X}_t(\gamma,i)+c_t(\gamma,i)$ for the instantaneous expected reward, where the discounted empirical average is given by

$$\bar{X}_t(\gamma,i) = \frac{1}{N_t(\gamma,i)}\sum_{s=1}^{t}\gamma^{t-s}X_s(i)\mathbb{1}_{\{I_s=i\}}\ , \quad N_t(\gamma,i) = \sum_{s=1}^{t}\gamma^{t-s}\mathbb{1}_{\{I_s=i\}},$$

and the discounted padding function is defined as

$$c_t(\gamma,i) = 2B\sqrt{\frac{\xi\log n_t(\gamma)}{N_t(\gamma,i)}}\ , \quad n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma,i)\ ,$$

for an appropriate parameter ξ. Using these notations, discounted-UCB (D-UCB) is defined in Algorithm 2. Remark that for $\gamma = 1$, D-UCB boils down to the standard UCB-1 algorithm.

In order to estimate the instantaneous expected reward, the D-UCB policy averages past rewards with a discount factor giving more weight to recent observations. We propose in this paper a more abrupt variant of UCB where averages are computed on a fixed-size horizon. At time t, instead of averaging the rewards over all past with a discount factor, *sliding-window UCB* relies on a local empirical average of the observed

Algorithm 2 Discounted UCB

for t from 1 to K, play arm $I_t = t$;
for t from $K+1$ to T, play arm

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(\gamma, i) + c_t(\gamma, i).$$

rewards, using only the τ last plays. Specifically, this algorithm constructs an UCB $\bar{X}_t(\tau, i) + c_t(\tau, i)$ for the instantaneous expected reward; the local empirical average is given by

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} X_s(i) \mathbb{1}_{\{I_s = i\}} \ , \quad N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{I_s = i\}},$$

and the padding function is defined as

$$c_t(\tau, i) = B \sqrt{\frac{\xi \log(t \wedge \tau)}{N_t(\tau, i)}} \ ,$$

where $t \wedge \tau$ denotes the minimum of t and τ, and ξ is a some appropriate constant. The policy defined in Algorithm 3 will be called in the sequel *Sliding-Window UCB* (SW-UCB).

Algorithm 3 Sliding-Window UCB

for t from 1 to K, play arm $I_t = t$;
for t from $K+1$ to T, play arm

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(\tau, i) + c_t(\tau, i),$$

In this paper, we investigate the behaviors of the discounted-UCB and of the sliding-window-UCB in an abruptly changing environment, and prove that they are almost rate-optimal in a minimax sense. In Section 2, we derive a finite-time upper-bound on the regret of D-UCB. In Section 3, we propose a similar analysis for the SW-UCB policy. We establish that it achieves the slightly better regret. In Section 4, we establish a lower-bound on the regret of any policy in an abruptly changing environment. As a by-product, we show that any policy (like UCB-1) that achieves a logarithmic regret in the stationary case cannot reach a regret of order smaller than $T/\log T$ in presence of breakpoints. The upper-bounds obtained in Sections 2 and 3 are based on a novel deviation inequality for self-normalized averages with random number of summands which is stated and proved in Section A. A maximal inequality, of independent interest, is also derived in Section B. Two simple Monte-Carlo experiments are presented to support our findings in Section 5.

2. Analysis of Discounted UCB

In this section, we analyze the behavior of D-UCB in an abruptly changing environment. Let Υ_T denote the number of breakpoints before time T, and let $\tilde{N}_T(i)$ denote the number of times arm i was played when it was not the best arm during the T first rounds:

$$\tilde{N}_T(i) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*\}}.$$

Denote by $\Delta\mu_T(i)$ the minimum of the difference of expected reward of the best arm $\mu_t(*)$ and the expected reward $\mu_t(i)$ of the i-th arm for all times $t \in \{1, \dots, T\}$ such that arm i is not the leading arm ($i \neq i_t^*$),

$$\Delta\mu_T(i) = \min\left\{t \in \{1, \dots, T\}, i \neq i_t^*, \mu_t(*) - \mu_t(i)\right\} \ . \tag{1}$$

We denote by $\mathbb{P}_\gamma$ and $\mathbb{E}_\gamma$ the probability distribution and expectation under policy D-UCB with discount factor γ. The next theorem computes a bound for the expected number of times in T rounds that the arm i is played, when this arm is suboptimal.

Theorem 1 *Let $\xi > 1/2$ and $\gamma \in (0,1)$.For any arm $i \in \{1, \dots, K\}$,*

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] \leq \mathrm{B}(\gamma)T(1-\gamma)\log\frac{1}{1-\gamma} + \mathrm{C}(\gamma)\frac{\Upsilon_T}{1-\gamma}\log\frac{1}{1-\gamma} \ , \tag{2}$$

where

$$\mathrm{B}(\gamma) = \frac{16B^2\xi}{\gamma^{1/(1-\gamma)}(\Delta\mu_T(i))^2}\frac{\lceil T(1-\gamma)\rceil}{T(1-\gamma)} + \frac{2\left\lceil -\log(1-\gamma)/\log(1+4\sqrt{1-1/2\xi})\right\rceil}{-\log(1-\gamma)\left(1-\gamma^{1/(1-\gamma)}\right)}$$

and

$$\mathrm{C}(\gamma) = \frac{\gamma-1}{\log(1-\gamma)\log\gamma} \times \log\left((1-\gamma)\xi\log n_K(\gamma)\right) \ . \tag{3}$$

Remark 2 *When γ goes to 1 we have $\mathrm{C}(\gamma) \to 1$ and*

$$\mathrm{B}(\gamma) \to \frac{16\,\mathrm{e}\,B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{(1-\mathrm{e}^{-1})\log\left(1+4\sqrt{1-1/2\xi}\right)} \ .$$

Proof The proof is adapted from the finite-type analysis of Auer et al. (2002). There are however two main differences. First, because the expected reward changes, the discounted empirical mean $\bar{X}_t(\gamma, i)$ is now a *biased* estimator of the expected reward $\mu_t(i)$. The second difference stems from the deviation inequality itself: instead of using a Chernoff-Hoeffding bound, we use a novel tailored-made control on a self-normalized mean of the rewards with a random number of summands. The proof is in 5 steps:

Step 1 We upper-bound the number of times the suboptimal arm i is played as follows:

$$\begin{aligned}\tilde{N}_T(i) &= 1 + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*\}} \\ &= 1 + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) < A(\gamma)\}} + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\}} \ ,\end{aligned}$$

where

$$A(\gamma) = \frac{16B^2\xi\log n_T(\gamma)}{(\Delta\mu_T(i))^2} \ . \tag{4}$$

Using Corollary 26 (stated and proved in the Appendix), we may upper-bound the first sum in the RHS as:

$$\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) < A(\gamma)\}} \leq \lceil T(1-\gamma)\rceil A(\gamma)\gamma^{-1/(1-\gamma)} \ .$$

In the sequel, for any positive m, we denote by $\mathcal{T}(\gamma)$ the set of all indices $t \in \{K+1, \dots, T\}$ such that $\mu_s(j) = \mu_t(j)$ for all $j \in \{1, \dots, K\}$ and all $t - D(\gamma) < s \leq t$, where

$$D(\gamma) = \frac{\log\left((1-\gamma)\xi \log n_K(\gamma)\right)}{\log \gamma} \, .$$

During a number of rounds (that depends on γ) following a breakpoint, the estimates of the expected rewards can be poor. Because of this, the D-UCB policy may play constantly the suboptimal arm i, which leads to the following bound:

$$\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\}} \leq \Upsilon_T D(\gamma) + \sum_{t \in \mathcal{T}(\gamma)} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\}} \, .$$

Putting everything together, we obtain:

$$\tilde{N}_T(i) \leq 1 + \lceil T(1-\gamma) \rceil A(\gamma) \gamma^{-1/(1-\gamma)} + \Upsilon_T D(\gamma) + \sum_{t \in \mathcal{T}(\gamma)} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\}} \, . \tag{5}$$

Step 2 Now, for $t \in \mathcal{T}(\gamma)$ the event $\{I_t = i \neq i_t^*, N_t(\gamma, i) \geq A(\gamma)\}$ may be decomposed as follows:

$$\begin{aligned} \{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\} \subseteq \left\{\bar{X}_t(\gamma,i) > \mu_t(i) + c_t(\gamma,i)\right\} \cup \left\{\bar{X}_t(\gamma,*) < \mu_t(*) - c_t(\gamma,*)\right\} \\ \cup \left\{\mu_t(*) - \mu_t(i) < 2c_t(\gamma,i), N_t(\gamma,i) \geq A(\gamma)\right\} . \end{aligned} \tag{6}$$

In words, playing the suboptimal arm i at time t may occur in three cases: if $\mu_t(i)$ is substantially over-estimated, if $\mu_t(*)$ is substantially under-estimated, or if $\mu_t(i)$ and $\mu_t(*)$ are close from each other. But for the choice of $A(\gamma)$ given in Equation (4), we have

$$c_t(\gamma,i) \leq 2B\sqrt{\frac{\xi \log n_t(\gamma)}{A(\gamma)}} \leq \frac{\Delta\mu_T(i)}{2} \, ,$$

so that the event $\{\mu_t(*) - \mu_t(i) < 2c_t(\gamma,i), N_t(\gamma,i) \geq A(\gamma)\}$ never occurs.

In Steps 3 and 4 we upper-bound the probability of the two first events of the RHS of (6). We show that for $t \in \mathcal{T}(\gamma)$, that is at least $D(\gamma)$ rounds after a breakpoint, the expected rewards of all arms are well estimated with high probability. For all $j \in \{1, \dots, K\}$, consider the following events

$$\mathcal{E}_t(\gamma, j) = \left\{\bar{X}_t(\gamma,i) > \mu_t(j) + c_t(\gamma,j)\right\}$$

The idea is the following: we upper-bound the probability of $\mathcal{E}_t(\gamma,j)$ by separately considering the fluctuations of $\bar{X}_t(\gamma,j)$ around $M_t(\gamma,j)/N_t(\gamma,j)$, and the 'bias' $M_t(\gamma,j)/N_t(\gamma,j) - \mu_t(j)$, where

$$M_t(\gamma,j) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{I_s = j\}} \mu_s(j) \, .$$

Step 3 Let us first consider the bias. First note that $M_t(\gamma,j)/N_t(\gamma,j)$, as a convex combination of elements $\mu_s(j) \in [0,B]$, belongs to interval $[0,B]$. Hence, $|M_t(\gamma,j)/N_t(\gamma,j) - \mu_t(j)| \leq B$. Second, for $t \in \mathcal{T}(\gamma)$,

$$\begin{aligned}|M_t(\gamma,j) - \mu_t(j)N_t(\gamma)| &= \left|\sum_{s=1}^{t-D(\gamma)} \gamma^{t-s}\left(\mu_s(j)-\mu_t(j)\right)\mathbb{1}_{\{I_s=j\}}\right| \\ &\leq \sum_{s=1}^{t-D(\gamma)} \gamma^{t-s}\left|\mu_s(j)-\mu_t(j)\right|\mathbb{1}_{\{I_s=j\}} \leq B\sum_{s=1}^{t-D(\gamma)} \gamma^{t-s}\mathbb{1}_{\{I_s=j\}} = B\gamma^{D(\gamma)}N_{t-D(\gamma)}(\gamma,j).\end{aligned}$$

As $N_{t-D(\gamma)}(\gamma,j) \leq (1-\gamma)^{-1}$, we get $|M_t(\gamma,j)/N_t(\gamma,j) - \mu_t(j)| \leq B\gamma^{D(\gamma)}(1-\gamma)^{-1}$. Altogether,

$$\left|\frac{M_t(\gamma,j)}{N_t(\gamma,j)} - \mu_t(j)\right| \leq B\left(1 \wedge \gamma^{D(\gamma)}(1-\gamma)^{-1}\right).$$

Hence, using the elementary inequality $1 \wedge x \leq \sqrt{x}$ and the definition of $D(\gamma)$, we obtain for $t \in \mathcal{T}(\gamma)$:

$$\left|\frac{M_t(\gamma,j)}{N_t(\gamma,j)} - \mu_t(j)\right| \leq B\sqrt{\frac{\gamma^{D(\gamma)}}{(1-\gamma)N_t(\gamma,i)}} \leq B\sqrt{\frac{\xi\log n_K(\gamma)}{N_t(\gamma,j)}} \leq \frac{1}{2}c_t(\gamma,j)\ .$$

In words: $D(\gamma)$ rounds after a breakpoint, the 'bias' is smaller than the half of the padding function. The other half of the padding function is used to control the fluctuations. In fact, for $t \in \mathcal{T}(\gamma)$:

$$\begin{aligned}\mathbb{P}_\gamma\left(\mathcal{E}_t(\gamma,j)\right) &\leq \mathbb{P}_\gamma\left(\bar{X}_t(\gamma,j) > \mu_t(j) + B\sqrt{\frac{\xi\log n_t(\gamma)}{N_t(\gamma,j)}} + \left|\frac{M_t(\gamma,j)}{N_t(\gamma,j)} - \mu_t(j)\right|\right) \\ &\leq \mathbb{P}_\gamma\left(\bar{X}_t(\gamma,j) - \frac{M_t(\gamma,j)}{N_t(\gamma,j)} > B\sqrt{\frac{\xi\log n_t(\gamma)}{N_t(\gamma,j)}}\right).\end{aligned}$$

Step 4 Denote the discounted total reward obtained with arm j by

$$S_t(\gamma,j) = \sum_{s=1}^{t}\gamma^{t-s}\mathbb{1}_{\{I_s=j\}}X_s(j) = N_t(\gamma,j)\bar{X}_t(\gamma,j)\ .$$

Using Theorem 18 and the fact that $N_t(\gamma,j) \geq N_t(\gamma^2,j)$, the previous inequality rewrites:

$$\begin{aligned}\mathbb{P}_\gamma\left(\mathcal{E}_t(\gamma,j)\right) &\leq \mathbb{P}_\gamma\left(\frac{S_t(\gamma,j)-M_t(\gamma,j)}{\sqrt{N_t(\gamma^2,j)}} > B\sqrt{\frac{\xi N_t(\gamma,j)\log n_t(\gamma)}{N_t(\gamma^2,j)}}\right) \\ &\leq \mathbb{P}_\gamma\left(\frac{S_t(\gamma,j)-M_t(\gamma,j)}{\sqrt{N_t(\gamma^2,j)}} > B\sqrt{\xi\log n_t(\gamma)}\right) \\ &\leq \left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil \exp\left(-2\xi\log n_t(\gamma)\left(1-\frac{\eta^2}{16}\right)\right) \\ &= \left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil n_t(\gamma)^{-2\xi\left(1-\frac{\eta^2}{16}\right)}\ .\end{aligned}$$

Step 5 Hence, we finally obtain from Equation (5) :

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] \leq 1 + \lceil T(1-\gamma)\rceil A(\gamma)\gamma^{-1/(1-\gamma)} + D(\gamma)\Upsilon_T + 2\sum_{t\in\mathcal{T}(\gamma)}\left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil n_t(\gamma)^{-2\xi\left(1-\frac{\eta^2}{16}\right)}.$$

When $\Upsilon_T \neq 0$, γ is taken strictly smaller than 1 (see Remark 3). As $\xi > \frac{1}{2}$, we take $\eta = 4\sqrt{1-1/2\xi}$, so that $2\xi\left(1-\eta^2/16\right) = 1$. For that choice, with $\tau = (1-\gamma)^{-1}$,

$$\begin{aligned}\sum_{t\in\mathcal{T}(\gamma)}\left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil n_t(\gamma)^{-2\xi\left(1-\frac{\eta^2}{16}\right)} &\leq \tau - K + \sum_{t=\tau}^{T}\left\lceil\frac{\log n_\tau(\gamma)}{\log(1+\eta)}\right\rceil n_\tau(\gamma)^{-1}\\ &\leq \tau - K + \left\lceil\frac{\log n_\tau(\gamma)}{\log(1+\eta)}\right\rceil\frac{n}{n_\tau(\gamma)}\\ &\leq \tau - K + \left\lceil\frac{\log\frac{1}{1-\gamma}}{\log(1+\eta)}\right\rceil\frac{T(1-\gamma)}{1-\gamma^{1/(1-\gamma)}},\end{aligned}$$

we obtain the statement of the Theorem. ■

Remark 3 *If horizon T and the growth rate of the number of breakpoints Υ_T are known in advance, the discount factor γ can be chosen so as to minimize the RHS in Equation 2. Taking $\gamma = 1-(4B)^{-1}\sqrt{\Upsilon_T/T}$ yields:*

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] = O\left(\sqrt{T\Upsilon_T}\log T\right).$$

Assuming that $\Upsilon_T = O(T^\beta)$ for some $\beta \in [0,1)$, the regret is upper-bounded as $O\left(T^{(1+\beta)/2}\log T\right)$. In particular, if $\beta = 0$, the number of breakpoints Υ_T is upper-bounded by Υ independently of T, taking $\gamma = 1-(4B)^{-1}\sqrt{\Upsilon/T}$, the regret is bounded by $O\left(\sqrt{\Upsilon T}\log T\right)$. Thus, D-UCB matches the lower-bound of Theorem 13 up to a factor $\log T$.

Remark 4 *On the other hand, if the breakpoints have a positive density over time (say, if $\Upsilon_T \leq rT$ for a small positive constant r), then γ has to remain lower-bounded independently of T; Theorem 1 gives a linear, non-trivial bound on the regret and permits to calibrate the discount factor γ as a function of the density of the breakpoint: taking $\gamma = 1-\sqrt{r}/(4B)$ we get an upper-bound with a dominant term in $O\left(-T\sqrt{r}\log r\right)$.*

Remark 5 *Theorem 22 shows that for $\xi > 1/2$ and $t \in \mathcal{T}(\gamma)$, with high probability $\bar{X}_t(\gamma,i)$ is actually* never *larger than $\mu_t(i) + c_t(\gamma,i)$.*

Remark 6 *If the growth rate of Υ_T is known in advance, but not the horizon T, then we can use the "doubling trick" to set the value of γ. Namely, for t and k such that $2^k \leq t < 2^{k+1}$, take $\gamma = 1-(4B)^{-1}(2^k)^{(\beta-1)/2}$.*

3. Sliding window UCB

In this section, we analyze the performance of SW-UCB in an abruptly changing environment. We denote by $\mathbb{P}_\tau$ and $\mathbb{E}_\tau$ the probability distribution and expectation under policy SW-UCB with window size τ.

Theorem 7 *Let $\xi > 1/2$. For any integer τ and any arm $i \in \{1, \dots, K\}$,*

$$\mathbb{E}_\tau \left[\tilde{N}_T(i) \right] \leq \mathrm{C}(\tau) \frac{T \log \tau}{\tau} + \tau \Upsilon_T + \log^2(\tau) \,, \tag{7}$$

where

$$\mathrm{C}(\tau) = \frac{4B^2\xi}{(\Delta\mu_T(i))^2} \frac{\lceil T/\tau \rceil}{T/\tau} + \frac{2}{\log \tau} \left\lceil \frac{\log(\tau)}{\log(1 + 4\sqrt{1 - (2\xi)^{-1}})} \right\rceil \,.$$

Remark 8 *As τ goes to infinity*

$$\mathrm{C}(\tau) \to \frac{4B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{\log(1 + 4\sqrt{1 - (2\xi)^{-1}})}.$$

Proof We follow the lines of the proof of Theorem 1. The main difference is that for $t \in \mathcal{T}(\tau)$ defined here as the set of all indices $t \in \{K+1, \dots, T\}$ such that $\mu_s(j) = \mu_t(j)$ for all $j \in \{1, \dots, K\}$ and all $t - \tau < s \leq t$, the bias exactly vanishes; consequently, Step 3 can be bypassed.

Step 1 Let $A(\tau) = 4B^2\xi \log \tau (\Delta\mu_T(i))^{-2}$; using Lemma 25, we have:

$$\begin{aligned}
\tilde{N}_T(i) &= 1 + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*\}} \\
&\leq 1 + \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i, N_t(\tau,i) < A(\tau)\}} + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\tau,i) \geq A(\tau)\}} \\
&\leq 1 + \lceil T/\tau \rceil A(\tau) + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\tau,i) \geq A(\tau)\}} \\
&\leq 1 + \lceil T/\tau \rceil A(\tau) + \Upsilon_T \tau + \sum_{t \in \mathcal{T}(\tau)} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\tau,i) \geq A(\tau)\}}
\end{aligned} \tag{8}$$

Step 2 For $t \in \mathcal{T}(\tau)$ we have

$$\begin{aligned}
\{I_t = i, N_t(\tau,i) \geq A(\tau)\} \subset \left\{ \bar{X}_t(\tau,i) > \mu_t(i) + c_t(\tau,i) \right\} \cup \left\{ \bar{X}_t(\tau,*) < \mu_t(*) - c_t(\tau,*) \right\} \\
\cup \left\{ \mu_t(*) - \mu_t(i) < 2c_t(\tau,i), N_t(\tau,i) \geq A(\tau) \right\}.
\end{aligned} \tag{9}$$

On the event $\{N_t(\tau,i) \geq A(\tau)\}$, we have

$$c_t(\tau,i) = B\sqrt{\frac{\xi \log(t \wedge \tau)}{N_t(\tau,i)}} \leq B\sqrt{\frac{\xi \log \tau}{A(\tau)}} = B\sqrt{\frac{\xi \log(\tau) \left(\Delta\mu_T(i)\right)^2}{4B^2\xi \log \tau}} \leq \frac{\Delta\mu_T(i)}{2},$$

so that the event $\{\mu_t(*) - \mu_t(i) < 2c_t(\tau,i), N_t(\tau,i) \geq A(\tau)\}$ has $\mathbb{P}_\tau$-probability 0.

Steps 3-4 Now, for $t \in \mathcal{T}(\tau)$ and for all $j \in \{1, \dots, K\}$, Corollary 21 applies and yields:

$$\begin{aligned}\mathbb{P}\left(\bar{X}_t(\tau,j) > \mu_t(j) + c_t(\tau,j)\right) &\leq \mathbb{P}\left(\bar{X}_t(\tau,j) > \mu_t(j) + B\sqrt{\frac{\xi \log(t\wedge\tau)}{N_t(\tau,j)}}\right) \\ &\leq \left\lceil \frac{\log(t\wedge\tau)}{\log(1+\eta)} \right\rceil \exp\left(-2\xi\log(t\wedge\tau)\left(1-\frac{\eta^2}{16}\right)\right) \\ &= \left\lceil \frac{\log(t\wedge\tau)}{\log(1+\eta)} \right\rceil (t\wedge\tau)^{-2\xi\left(1-\eta^2/16\right)}, \end{aligned} \tag{10}$$

and similarly

$$\mathbb{P}\left(\bar{X}_t(\tau,j) < \mu_t(j) - c_t(\tau,j)\right) \leq \left\lceil \frac{\log(t\wedge\tau)}{\log(1+\eta)} \right\rceil (t\wedge\tau)^{-2\xi\left(1-\eta^2/16\right)}. \tag{11}$$

Steps 5 In the following we take $\eta = 4\sqrt{1-\frac{1}{2\xi}}$, so that we have $2\xi\left(1-\eta^2/16\right) = 1$. Thus, using Equations (9),(10) and (11), Inequality (8) yields

$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] \leq 1 + \lceil T/\tau \rceil A(\tau) + \tau\Upsilon_T + 2\sum_{t=1}^{T} \frac{\left\lceil \frac{\log(t\wedge\tau)}{\log(1+\eta)} \right\rceil}{(t\wedge\tau)}.$$

The results follows, noting that

$$\sum_{t=K+1}^{T} \frac{\log(t\wedge\tau)}{t\wedge\tau} \leq \sum_{t=2}^{\tau} \frac{\log t}{t} + \sum_{t=1}^{T} \frac{\log\tau}{\tau} \leq \frac{1}{2}\log^2(\tau) + \frac{T\log\tau}{\tau}.$$

■

Remark 9 *If the horizon T and the growth rate of the number of breakpoints Υ_T are known in advance, the window size τ can be chosen so as to minimize the RHS in Equation* (7)*. Taking $\tau = 2B\sqrt{T\log(T)/\Upsilon_T}$ yields*

$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] = O\left(\sqrt{\Upsilon_T T \log T}\right).$$

Assuming that $\Upsilon_T = O(T^\beta)$ for some $\beta \in [0,1)$, the average regret is upper-bounded as $O\left(T^{(1+\beta)/2}\sqrt{\log T}\right)$. In particular, if $\beta = 0$, the number of breakpoints Υ_T is upper-bounded by Υ independently of T, then with $\tau = 2B\sqrt{T\log(T)/\Upsilon}$ the upper-bound is $O\left(\sqrt{\Upsilon T\log T}\right)$. Thus, SW-UCB matches the lower-bound of Theorem 13 up to a factor $\sqrt{\log T}$, slightly better than the D-UCB.

Remark 10 *On the other hand, if the breakpoints have a positive density over time, then τ has to remain lower-bounded independently of T. For instance, if $\Upsilon_T \leq rT$ for some (small) positive rate r, and for the choice $\tau = 2B\sqrt{-\log r/r}$, Theorem 7 gives*

$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] = O\left(T\sqrt{-r\log(r)}\right).$$

Remark 11 *If there is no breakpoint ($\Upsilon_T = 0$), the best choice is obviously to take the window as a large as possible, that is $\tau = T$. Then the procedure is exactly standard UCB. A slight modification of the preceeding proof for $\xi = \frac{1}{2} + \epsilon$ with arbitrary small ϵ yields*

$$\mathbb{E}_{\mathrm{UCB}}\left[\tilde{N}_T(i)\right] \leq \frac{2B^2}{(\Delta\mu(i))^2}\log T\,(1+O(1))\,.$$

We recover the same kind of bounds that are usually obtained in the analysis of UCB, see for instance Auer et al. (2002), with a better constant.

Remark 12 *The computational complexity of SW-UCB is, as for D-UCB, linear in time and does not involve τ. However, SW-UCB requires to store the last τ actions and rewards at each time t in order to efficiently update $N_t(\tau, i)$ and $\bar{X}_t(\tau, i)$.*

4. A lower-bound on the regret in abruptly changing environment

In this section, we consider a particular non-stationary bandit problem where the distributions of rewards on each arm are piecewise constant and have two breakpoints. Given any policy π, we derive a lower-bound on the number of times a sub-optimal arm is played (and thus, on the regret) in at least one such game. Quite intuitively, the less explorative a policy is, the longer it may keep a suboptimal policy after a breakpoint. Theorem 13 gives a precise content to this statement.

As in the previous section, K denotes the number of arms, and the rewards are assumed to be bounded in $[0, B]$. Consider any deterministic policy π of choosing the arms $I_1, \ldots, I_T$ played at each time depending to the past rewards

$$G_t \triangleq X_t(I_t),$$

and recall that I_t is measurable with respect to the sigma-field $\sigma(G_1, \ldots, G_t)$ of the past observed rewards. Denote by $N_{s:t}(i)$ the number of times arm i is played between times s and t

$$N_{s:t}(i) = \sum_{u=s}^{t} \mathbb{1}_{\{I_u = i\}},$$

and $N_T(i) = N_{1:T}(i)$. For $1 \leq i \leq K$, let P_i be the probability distribution of the outcomes of arm i, and let $\mu(i)$ denote its expectation. Assume that $\mu(1) > \mu(i)$ for all $2 \leq i \leq K$. Denote by $\mathbb{P}_\pi$ the distribution of rewards under policy π, that is:

$$d\mathbb{P}_\pi(g_{1:T}|I_{1:T}) = \prod_{t=1}^{T} dP_{i_t}(g_t).$$

For any random variable W measurable with respect to $\sigma(G_1, \ldots, G_T)$, denote by $\mathbb{E}_\pi[W]$ its expectation under distribution $\mathbb{P}_\pi$.

In the sequel, we divide the period $\{1, \ldots, T\}$ into epochs of size $\tau \in \{1, \ldots, T\}$, and we modify the distribution of the rewards so that on one of those periods, arm K becomes the one with highest expected reward. Specifically: let Q be a distribution of rewards with expectation $\nu > \mu(1)$, let $\delta = \nu - \mu(1)$ and let $\alpha = \mathrm{D}(P_K; Q)$ be the Kullback-Leibler divergence between P_K and Q. For all $1 \leq j \leq M = \left\lfloor \frac{T}{\tau} \right\rfloor$, we

consider the modification $\mathbb{P}_\pi^j$ of $\mathbb{P}_\pi$ such that on the j-th period of size τ, the distribution of rewards of the K-th arm is changed to ν. That is, for every sequence of rewards $g_{1:T}$,

$$\frac{d\mathbb{P}_\pi^j}{d\mathbb{P}_\pi}(g_{1:T}|I_{1:T}) = \prod_{t=1+(j-1)\tau, I_t=K}^{j\tau} \frac{dQ}{dP_K}(g_t) \ .$$

Besides, let

$$N^j(i) = N_{1+(j-1)\tau:j\tau}(i)$$

be the number of times arm i is played in the j-th period. For any random variable W in $\sigma(G_1, \ldots, G_T)$, denote by $\mathbb{E}_\pi^j[W]$ its expectation under distribution $\mathbb{P}_\pi^j$. Now, denote by $\mathbb{P}_\pi^*$ the distribution of rewards when j is chosen uniformly at random in the set $\{1, \ldots, M\}$ - in other words, $\mathbb{P}_\pi^*$ is the (uniform) mixture of the $(\mathbb{P}_\pi^j)_{1\leq j\leq M}$, and denote by $\mathbb{E}_\pi^*[\cdot]$ the expectation under $\mathbb{P}_\pi^*$:

$$\mathbb{E}_\pi^*[W] = \frac{1}{M}\sum_{j=1}^M \mathbb{E}_\pi^j[W].$$

In the following, we lower-bound the expect regret of any policy π under $\mathbb{P}_\pi^*$ in terms of its regret under $\mathbb{P}_\pi$.

Theorem 13 *For any policy π and any horizon T such that $64/(9\alpha) \leq \mathbb{E}_\pi[N_T(K)] \leq T/(4\alpha)$,*

$$\mathbb{E}_\pi^*[R_T] \geq C(\mu)\frac{T}{\mathbb{E}_\pi[R_T]},$$

where

$$C(\mu) = \frac{32\delta(\mu(1)-\mu(K))}{27\alpha} \ .$$

Proof The main ingredients of this reasoning are inspired by the proof of Theorem 5.1 in Auer et al. (2002/03), see also Kulkarni and Lugosi (2000). First, note that the Küllback-Leibler divergence $\mathrm{D}(\mathbb{P}_\pi, \mathbb{P}_\pi^j)$ is:

$$\begin{aligned}
\mathrm{D}(\mathbb{P}_\pi, \mathbb{P}_\pi^j) &= \sum_{t=1}^T \mathrm{D}\left(\mathbb{P}_\pi\left(G_t|G_{1:t-1}\right); \mathbb{P}_\pi^j\left(G_t|G_{1:t-1}\right)\right) \\
&= \sum_{t=1+(j-1)\tau}^{j\tau} \mathbb{P}_\pi\left(I_t = K\right)\mathrm{D}(P_K; Q) \\
&= \alpha\mathbb{E}_\pi\left[N_{1+(j-1)\tau:j\tau}(K)\right] \ .
\end{aligned}$$

Hence, by Lemma A.1 in Auer et al. (2002/03),

$$\mathbb{E}_\pi^j[N^j(K)] \leq \mathbb{E}_\pi[N^j(K)] + \frac{\tau}{2}\sqrt{\mathrm{D}(\mathbb{P}_\pi, \mathbb{P}_\pi^j)} = \mathbb{E}_\pi[N^j(K)] + \frac{\tau}{2}\sqrt{\alpha\mathbb{E}_\pi[N^j(K)]} \ .$$

Consequently, since $\sum_{j=1}^M N^j(K) \leq N_T(K)$,

$$\sum_{j=1}^M \mathbb{E}_\pi^j[N^j(K)] \leq \mathbb{E}[N_T(K)] + \frac{\tau}{2}\sum_{j=1}^M \sqrt{\alpha\mathbb{E}_\pi[N^j(K)]} \leq \mathbb{E}_\pi[N_T(K)] + \frac{\tau}{2}\sqrt{\alpha M\mathbb{E}_\pi[N_T(K)]} \ .$$

Thus, there exists $1 \leq j \leq M$ such that

$$\begin{aligned}\mathbb{E}_\pi^*[N^j(K)] &\leq \frac{1}{M}\mathbb{E}_\pi[N_T(K)] + \frac{\tau}{2M}\sqrt{\alpha M \mathbb{E}_\pi[N_T(K)]} \\ &\leq \frac{\tau}{T-\tau}\mathbb{E}_\pi[N_T(K)] + \frac{1}{2}\sqrt{\alpha\frac{\tau^3}{T-\tau}\mathbb{E}_\pi[N_T(K)]} \ .\end{aligned}$$

Now, the expectation under $\mathbb{P}_\pi^*$ of the regret R_T is lower-bounded by:

$$\mathbb{E}_\pi^*[R_T] \geq \delta\left(\tau - \mathbb{E}_\pi^*[N_T(K)]\right) \geq \delta\left(\tau - \frac{\tau}{T-\tau}\mathbb{E}_\pi[N_T(K)] - \frac{1}{2}\sqrt{\alpha\frac{\tau^3}{T-\tau}\mathbb{E}_\pi[N_T(K)]}\right) \ .$$

Maximizing the right hand side of the previous inequality by choosing $\tau = 16T/(9\alpha\mathbb{E}_\pi[N(K)])$ or equivalently $M = 9\alpha/(16\mathbb{E}_\pi[N(K)])$ leads to the lower-bound:

$$\mathbb{E}_\pi^*[R_T] \geq \frac{32\delta}{27\alpha}\left(1 - \frac{\alpha\mathbb{E}_\pi[N_T(K)]}{T}\right)^2\left(1 - \frac{16}{9\alpha\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[N_T(K)]} \ .$$

To conclude, simply note that $N_T(K) \leq \mathbb{E}_\pi[R_T]/(\mu(1) - \mu(K))$. We obtain:

$$\mathbb{E}_\pi^*[R_T] \geq \frac{32\delta(\mu(1)-\mu(K))}{27\alpha}\left(1 - \frac{\alpha\mathbb{E}_\pi[N_T(K)]}{T}\right)^2\left(1 - \frac{16}{9\alpha\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[R_T]} \ ,$$

which directly leads to the statement of the Theorem. ■

The following corollary states that no policy can have a non-stationary regret of order smaller than $\sqrt{T}$. It appears here as a consequence of Theorem 13, although it can also be proved directly.

Corollary 14 *For any policy π and any positive horizon T,*

$$\max\{\mathbb{E}_\pi(R_T), \mathbb{E}_\pi^*(R_T)\} \geq \sqrt{C(\mu)T} \ .$$

Proof If $\mathbb{E}_\pi[N_T(K)] \leq 16/(9\alpha)$, or if $\mathbb{E}_\pi[N_T(K)] \geq T/\alpha$, the result is obvious. Otherwise, Theorem 13 implies that:

$$\max\{\mathbb{E}_\pi(R_T), \mathbb{E}_\pi^*(R_T)\} \geq \max\{\mathbb{E}_\pi(R_T), C(\mu)\frac{T}{\mathbb{E}_\pi(R_T)}\} \geq \sqrt{C(\mu)T} \ .$$

■

Remark 15 *To keep simple notations, Theorem 13 is stated and proved here for deterministic policy. It is easily verified that the same results also holds for randomized strategies (such as EXP3-P, see Auer et al. (2002/03)).*

Remark 16 *In words, Theorem 13 states that for any policy not playing each arm often enough, there is necessarily a time where a breakpoint is not seen after a long period. For instance, as standard UCB satisfies $\mathbb{E}_\pi[N(K)] = \Theta(\log T)$, then*

$$\mathbb{E}_\pi^*[R_T] \quad \geq \quad c\frac{T}{\log T}$$

for some positive c depending on the reward distribution.

Remark 17 *This result is to be compared with standard minimax lower-bounds on the regret. On one hand, a* fixed-game lower-bound *in $O(\log T)$ was proved in Lai and Robbins (1985) for the stationary case, when the distributions of rewards are fixed and T is allowed to go to infinity. On the other hand, a finite-time* minimax lower-bound *for individual sequences in $O(\sqrt{T})$ is proved in Auer et al. (2002/03). In this bound, for each horizon T the worst case among all possible reward distributions is considered, which explains the discrepancy. This result is obtained by letting the distance between distributions of rewards tend to 0 (typically, as $1/\sqrt{T}$). In Theorem 13, no assumption is made on the distributions of rewards P_i and Q, their distance actually remains lower-bounded independently of T. In fact, in the case considered here minimax regret and fixed-game minimal regret appear to have the same order of magnitude.*

5. Simulations

We consider here two settings. In the first example, there are $K = 3$ arms and the time horizon is set to $T = 10^4$. The agent goal is to minimize the expected regret. The rewards of arm $i \in \{1, \dots, K\}$ at time t are independent Bernoulli random variables with success probability $p_t(i)$, with $p_t(1) = 0.5$, $p_t(2) = 0.3$ and for $t \in \{1, \dots, T\}$:

$$p_t(3) = \begin{cases} 0.4 & \text{for} \quad t < 3000 \text{ or } t \geq 5000, \\ 0.9 & \text{for} \quad 3000 \leq t < 5000. \end{cases}$$

As one may notice, the optimal policy for this bandit task is to select arm 1 before the first breakpoint ($t = 3000$) and after the second breakpoint ($t = 5000$). In the left panel of Figure 1, we represent the evolution of two criteria in function of t: the number of times policy 1 has been played, and the cumulated regret (bottom plot). These two measures are obviously related, but they are not completely equivalent as sub-optimal arms can yield relatively high rewards. We compare the UCB-1 algorithm with $\xi = \frac{1}{2}$, the EXP3.S algorithm described in Auer et al. (2002/03) with the tuned parameters given in Corollary 8.3 (with the notations of this paper $\alpha = T^{-1}$ and $\gamma = \sqrt{K(\Upsilon_T \log(KT) + \mathrm{e})/[(\mathrm{e}-1)T]}$ with $\Upsilon_T = 2$), the D-UCB algorithm with $\xi = 1/2$ and $\gamma = 1 - 1/4\sqrt{T}$ and the SW-UCB with $\xi = 1/2$ and $\tau = 4\sqrt{n \log T}$. The parameters are tuned to obtain roughly optimal performance for the chosen horizon T and the number of breakpoints.

As can be seen in Figure 1 (and as consistently observed over the simulations), D-UCB performs almost as well as SW-UCB. Both of them waste significantly less time than EXP3.S and UCB-1 to detect the breakpoints, and quickly concentrate their pulls on the optimal arm. Observe that policy UCB-1, initially the best, reacts very fast to the first breakpoint ($t = 3000$), as the confidence interval for arm 3 at this step is very loose. On the contrary, it takes a very long time after the second breakpoint ($t = 5000$) for UCB-1 to play arm 1 again.

In the second example, there are $K = 2$ arms, the rewards are still Bernoulli random variables with parameters $p_t(i)$ but are in persistent, continuous evolution. Arm 2 is taken as a reference ($p_t(2) = 1/2$ for all t), and the parameter of arm 1 evolves periodically as: $p_t(1) = 0.5 + 0.4 \cos(6\pi R t/T)$. Hence, the best arm to pull evolves cyclically and the transitions are smooth (regularly, the two arms are statistically indistinguishable). The middle plot in the right panel of Figure 1 represents the cumulative frequency of arm 1 pulls: D-UCB, SW-UCB and, to a lesser extent, EXP3.S track the cycles, while UCB-1 fails to identify the best current arm. Below, the evolutions of the cumulative regrets under the four policies are shown: in this continuously evolving environment, the performance of D-UCB and SW-UCB are almost equivalent while UCB-1 and the Exp3.S algorithms accumulate larger regrets.

6. Conclusion and perspectives

This paper theoretically establishes that the UCB policies can also be successfully adapted to cope with non-stationary environments. The upper bound of the SW-UCB in abruptly changing environment matches the upper bounds of the Exp3.S algorithm (i.e. $O(\sqrt{T\log(T)})$), showing that UCB policies can be at least as good as the softmax methods. In practice, numerical experiments also support this finding. For the two examples considered in this paper, the D-UCB and SW-UCB policies outperform the optimality tuned version of the Exp3.S algorithm.

The focus of this paper is on abruptly changing environment, but it is believed that the theoretical tools developed to handle the non-stationarity can be applied in different contexts. In particular, using a similar bias-variance decomposition of the discounted or windowed-rewards, the analysis of continuously evolving reward distributions can be done (and will be reported in a forthcoming paper). Furthermore, Theorems 18 and 22, dealing with concentration inequality for discounted martingale transforms, are powerful tools of independent interest.

As the previously reported Exp3.S algorithm, the performance of the proposed policy depends on tuning parameters, the discount factor for D-UCB and the window size for SW-UCB. These tuning parameters may be adaptively set, using data-driven approaches, such as the one proposed in Hartland et al. (2006). This is the subject of on-going research.

Appendix A. A Hoeffding-type inequality for self-normalized means with a random number of summands

Let $(X_t)_{t\geq 1}$ be a sequence of non-negative independent bounded random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We denote by B the upper bound, $X_t \in [0, B]$, $\mathbb{P}$-a.s. and by μ_t its expectation $\mu_t = \mathbb{E}[X_t]$. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields of $\mathcal{A}$ such that for each t, $\sigma(X_1 \ldots, X_t) \subset \mathcal{F}_t$ and for $s > t$, X_s is independent from $\mathcal{F}_t$. Consider a previsible sequence $(\epsilon_t)_{t\geq 1}$ of Bernoulli variables (for all $t > 0$, ϵ_t is $\mathcal{F}_{t-1}$-measurable). Denote by ϕ_t the Cramer transform of X_t: for $\lambda \in \mathbb{R}$,

$$\phi_t(\lambda) = \log \mathbb{E}[\exp(\lambda X_t)] \; .$$

For $\gamma \in [0, 1)$, consider the following random variables

$$S_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} X_s \epsilon_s \; , \qquad M_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \mu_s \epsilon_s \; , \qquad N_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s \; . \tag{12}$$

Let also

$$n_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} = \left\{ \begin{array}{cc} \frac{1-\gamma^t}{1-\gamma} & \text{if} \quad \gamma < 1, \\ t & \text{if} \quad \gamma = 1. \end{array} \right.$$

Theorem 18 *For all integers t and all $\delta > 0$,*

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta \right) \leq \left\lceil \frac{\log n_t(\gamma)}{\log(1+\eta)} \right\rceil \exp\left(-\frac{2\delta^2}{B^2} \left(1 - \frac{\eta^2}{16} \right) \right)$$

for all $\eta > 0$.

Remark 19 *Actually, we prove the slightly stronger inequality:*

$$\mathbb{P}\left(\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}>\delta\right)\leq\left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil\times\exp\left(-\frac{8\delta^2}{B^2\left((1+\eta)^{1/4}+(1+\eta)^{-1/4}\right)^2}\right)\ . \tag{13}$$

Proof First observe that we can assume $\epsilon_t=1$, since otherwise $(S_t(\gamma)-M_t(\gamma))/\sqrt{N_t(\gamma^2)}=(S_{t-1}(\gamma)-M_{t-1}(\gamma))/\sqrt{N_{t-1}(\gamma^2)}$ and the result follows from a simple induction. Second, note that for every positive λ and for every $u<t$, since ϵ_{u+1} is predictible, and since X_{u+1} is independent from $\mathcal{F}_u$,

$$\mathbb{E}\left[\exp\left(\lambda X_{u+1}\epsilon_{u+1}\right)|\,\mathcal{F}_u\right]=\exp\left(\phi_{u+1}\left(\lambda\epsilon_{u+1}\right)\right)=\exp\left(\phi_{u+1}\left(\lambda\right)\epsilon_{u+1}\right)\ .$$

Hence, as $S_{u+1}(\gamma)=\gamma S_u(\gamma)+X_{u+1}\epsilon_{u+1}$,

$$\mathbb{E}\left[\exp\left(\lambda S_{u+1}(\gamma)-\sum_{s=1}^{u+1}\phi_s\left(\lambda\gamma^{u+1-s}\right)\epsilon_s\right)\right]=\mathbb{E}\left[\exp\left(\lambda\gamma S_u(\gamma)-\sum_{s=1}^{u}\phi_s\left((\lambda\gamma)\gamma^{u-s}\right)\epsilon_s\right)\right]\ .$$

As $\phi(0)=0$, this proves by induction that

$$\mathbb{E}\left[\exp\left(\lambda S_t(\gamma)-\sum_{s=1}^{t}\phi_s\left(\lambda\gamma^{t-s}\right)\epsilon_s\right)\right]=1\ .$$

It is easily verified (see e.g. (Devroye et al., 1996, Lemma 8.1)) that under the stated assumptions, for all positive λ,

$$\phi_s(\lambda)\leq\lambda\mu_s+B^2\lambda^2/8\ , \tag{14}$$

showing that

$$\mathbb{E}\left[\exp\left(\lambda\left\{S_t(\gamma)-M_t(\gamma)\right\}-\frac{B^2}{8}\lambda^2N_t(\gamma^2)\right)\right]\leq 1.$$

Hence, for any $x>0$, the Markov inequality yields

$$\begin{aligned}\mathbb{P}\Bigg(\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}&>\frac{x}{\lambda\sqrt{N_t(\gamma^2)}}+\frac{\lambda B^2\sqrt{N_t(\gamma^2)}}{8}\Bigg)\\&=\mathbb{P}\left(\exp\left(\lambda\left(S_t(\gamma)-M_t(\gamma)\right)-\frac{B^2}{8}\lambda^2N_t(\gamma^2)\right)\geq\mathrm{e}^x\right)\leq\exp(-x)\ .\end{aligned}$$

Now, take $\eta>0$, let $D=\left\lceil\frac{\log n_t(\gamma)}{\log(1+\eta)}\right\rceil$ and, for every integer $k\in\{1,\ldots,D\}$, define

$$\lambda_k=\sqrt{\frac{8x}{B^2(1+\eta)^{k-\frac{1}{2}}}}.$$

Elementary algebra shows that for all z such that $(1+\eta)^{k-1}\leq z\leq(1+\eta)^k$, we have

$$\sqrt{\frac{(1+\eta)^{k-\frac{1}{2}}}{z}}+\sqrt{\frac{z}{(1+\eta)^{k-\frac{1}{2}}}}\leq(1+\eta)^{1/4}+(1+\eta)^{-1/4} \tag{15}$$

Thus, if $(1+\eta)^{k-1} \leq N_t(\gamma^2) \leq (1+\eta)^k$, then

$$\frac{x}{\lambda_k \sqrt{N_t(\gamma^2)}} + \frac{B^2}{8}\lambda_k \sqrt{N_t(\gamma^2)} = B\sqrt{\frac{x}{8}} \left(\sqrt{\frac{(1+\eta)^{k-\frac{1}{2}}}{N_t(\gamma^2)}} + \sqrt{\frac{N_t(\gamma^2)}{(1+\eta)^{k-\frac{1}{2}}}} \right)$$
$$\leq B\sqrt{\frac{x}{8}} \left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right).$$

Therefore, as $\epsilon_t = 1$ we have $1 \leq N_t(\gamma^2) \leq (1+\eta)^D$ and

$$\left\{ \frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > B\sqrt{\frac{x}{8}} \left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right) \right\}$$
$$\subset \bigcup_{k=1}^{D} \left\{ \frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \frac{x}{\lambda_k \sqrt{N_t(\gamma^2)}} + \frac{\lambda_k B^2 \sqrt{N_t(\gamma^2)}}{8} \right\} .$$

The union bound thus implies that:

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > B\sqrt{\frac{x}{8}} \left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right) \right)$$
$$\leq \sum_{k=1}^{D} \mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \frac{x}{\lambda_k \sqrt{N_t(\gamma^2)}} + \frac{\lambda_k B^2 \sqrt{N_t(\gamma^2)}}{8} \right) \leq D\exp(-x) .$$

For $\delta = B\sqrt{\frac{x}{8}} \left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right)$, this yields

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta \right) \leq D\exp\left(-\frac{8\delta^2}{B^2 \left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right)^2} \right) .$$

The conclusion follows, as it is easy to see that, for all $\eta > 0$,

$$\frac{4}{\left((1+\eta)^{1/4} + (1+\eta)^{-1/4} \right)^2} \geq 1 - \frac{\eta^2}{16} . \tag{16}$$

■

Remark 20 *For example, taking $\eta = 0.3$ in (13) yields*

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta \right) \leq 4\log n_t(\gamma)\, \mathrm{e}^{-\frac{1.99\delta^2}{B^2}} .$$

Classical Hoeffding bounds for deterministic ϵ_s yield an upper-bound in

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta \right) \leq \mathrm{e}^{-\frac{2\delta^2}{B^2}}$$

for all positive t. The factor behind the exponential and the very slightly larger exponent are the price to pay for the presence of random ϵ_s. Theorem 18 is maybe sub-optimal, but it is possible to show that for all $\delta > 0$ and for an appropriate choice of the previsible sequence $(\epsilon_s)_{s\geq 1}$

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta\right) \to 1$$

as t goes to infinity.

If all variables X_t have the same expectation μ, taking $\gamma = 1$ in Theorem 18 immediately leads to the following corollary:

Corollary 21 *For all integers t and τ,*

$$\mathbb{P}\left(\frac{\sum_{s=(t-\tau+1)\wedge 1}^{t}(X_s - \mu)\epsilon_s}{\sqrt{\sum_{s=(t-\tau+1)\wedge 1}^{t}\epsilon_s}} > \delta\right) \leq \left\lceil \frac{\log(t\wedge\tau)}{\log(1+\eta)}\right\rceil \exp\left(-\frac{2\delta^2}{B^2}\left(1-\frac{\eta^2}{16}\right)\right)$$

Appendix B. A maximal inequality for self-normalized means with a random number of summands

In this section, we prove a stronger version of Theorem 18: we upper-bound the probability that, at some time t, the average reward deviates from its expectation. We keep the same notations as in Section A.

Theorem 22 *For all positive integer T and all $\delta > 0$,*

$$\mathbb{P}\left(\sup_{1\leq t\leq T}\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta\right) \leq \left\lceil \frac{\log\left(\gamma^{-2T} n_T(\gamma^2)\right)}{\log(1+\eta)}\right\rceil \exp\left(-\frac{2\delta^2}{B^2}\left(1-\frac{\eta^2}{16}\right)\right) .$$

for all $\eta > 0$.

Remark 23 *Note that if $\gamma < 1$, then*

$$\log\left(\gamma^{-2T} n_T(\gamma^2)\right) \leq \frac{2T(1-\gamma)}{\gamma} + \log\frac{1}{1-\gamma^2} ,$$

while for $\gamma = 1$ we have:

$$\log\left(\gamma^{-2T} n_T(\gamma^2)\right) = \log T.$$

Remark 24 *Classical Hoeffding bounds for deterministic ϵ_s yield an upper-bound in*

$$\mathbb{P}\left(\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta\right) \leq \exp(-2\delta^2)$$

for all positive t. The factor behind the exponential (depending on T and ϵ) and the very slightly larger exponant are the price to pay for uniformity in t. For example, taking $\eta = 0.3$ yields

$$\mathbb{P}\left(\sup_{1\leq t\leq T}\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta\right) \leq \left\lceil 4\log\left(\gamma^{-2T} n_T(\gamma^2)\right)\right\rceil \exp\left(-\frac{1.99\delta^2}{B^2}\right) .$$

Proof For $\lambda > 0$, define

$$Z_t^\lambda = \exp\left(\lambda\gamma^{-t}S_t(\gamma) - \sum_{s=1}^{t}\phi_s\left(\lambda\gamma^{-s}\right)\epsilon_s\right) . \tag{17}$$

Note that

$$\mathbb{E}\left[\exp\left(\lambda\gamma^{-t}X_t\epsilon_t\right)\middle|\,\mathcal{F}_{t-1}\right] = \exp\left(\epsilon_t\phi_t\left(\lambda\gamma^{-t}\right)\right).$$

Since $\gamma^{-t}S_t(\gamma) = \gamma^{-(t-1)}S_{t-1}(\gamma) + \gamma^{-t}X_t\epsilon_t$, we may therefore write

$$\mathbb{E}\left[\exp(\lambda\gamma^{-t}S_t(\gamma))\middle|\,\mathcal{F}_{t-1}\right] = \exp\left(\lambda\gamma^{-(t-1)}S_{t-1}(\gamma)\right)\exp\left(\epsilon_t\phi_t(\lambda\gamma^{-t})\right) ,$$

showing that $\{Z_t^\lambda\}$ is a martingale adapted to the filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$. As already mentionned (see e.g. (Devroye et al., 1996, Lemma 8.1)), under the stated assumptions

$$\phi_t(\lambda) \leq \lambda\mu_t + B^2\lambda^2/8 ,$$

showing that for all $\lambda > 0$,

$$W_t^\lambda = \exp\left(\lambda\gamma^{-t}S_t(\gamma) - \lambda\gamma^{-t}M_t(\gamma) - (B^2/8)\lambda^2\gamma^{-2t}N_t(\gamma^2)\right) \tag{18}$$

is a super-martingale. Hence, for any $x > 0$ we have

$$\mathbb{P}\left(\sup_{1\leq t\leq T} W_t^\lambda \geq \exp(x)\right) \leq \exp(-x) . \tag{19}$$

On the other hand, note that

$$\left\{W_t^\lambda > \exp(x)\right\} = \left\{\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \frac{x\gamma^t}{\lambda\sqrt{N_t(\gamma^2)}} + \frac{B^2}{8}\lambda\gamma^{-t}\sqrt{N_t(\gamma^2)}\right\} . \tag{20}$$

Now, let $D = \left\lceil \frac{\log\left(\gamma^{-2T}n_T(\gamma^2)\right)}{\log(1+\eta)} \right\rceil$ and for every integer $k \in \{1, \ldots, D\}$, define

$$\lambda_k = \sqrt{\frac{8x}{B^2(1+\eta)^{k-\frac{1}{2}}}}.$$

Thus, if $(1+\eta)^{k-1} \leq \gamma^{-2t}N_t(\gamma^2) \leq (1+\eta)^k$, then using Equation (15) yields:

$$\begin{aligned}\frac{x\gamma^t}{\lambda_k\sqrt{N_t(\gamma^2)}} + \frac{B^2}{8}\lambda_k\gamma^{-t}\sqrt{N_t(\gamma^2)} &= B\sqrt{\frac{x}{8}}\left(\sqrt{\frac{(1+\eta)^{k-\frac{1}{2}}}{\gamma^{-2t}N_t(\gamma^2)}} + \sqrt{\frac{\gamma^{-2t}N_t(\gamma^2)}{(1+\eta)^{k-\frac{1}{2}}}}\right) \\ &\leq B\sqrt{\frac{x}{8}}\left((1+\eta)^{1/4} + (1+\eta)^{-1/4}\right),\end{aligned}$$

which proves, using Equation (20), that

$$\begin{aligned}&\left\{\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > B\sqrt{\frac{x}{8}}\left((1+\eta)^{1/4} + (1+\eta)^{-1/4}\right)\right\} \\ &\qquad \subset \bigcup_{k=1}^{D}\left\{\frac{S_t(\gamma) - M_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \frac{x\gamma^t}{\lambda_k\sqrt{N_t(\gamma^2)}} + \frac{B^2}{8}\lambda_k\gamma^{-t}\sqrt{N_t(\gamma^2)}\right\} \subset \left\{W_t^{\lambda_k} > \exp(x)\right\} .\end{aligned}$$

But as

$$\sup_{1\le t\le T}\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}=\sup_{1\le t\le T,\epsilon_t=1}\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}\ ,$$

we assume $\epsilon_t=1$ and thus $1\le N_t(\gamma^2)\le(1+\eta)^D$. Hence, thanks to Equation (19) we obtain:

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{1\le t\le T}\left\{\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}>B\sqrt{\frac{x}{8}}\left((1+\eta)^{1/4}+(1+\eta)^{-1/4}\right)\right\}\right)\\&\quad\le\mathbb{P}\left(\bigcup_{1\le t\le T}\bigcup_{1\le k\le D}\left\{W_t^{\lambda_k}>\exp(x)\right\}\right)=\mathbb{P}\left(\bigcup_{1\le k\le D}\bigcup_{1\le t\le T}\left\{W_t^{\lambda_k}>\exp(x)\right\}\right)\le D\exp(-x)\ .\end{aligned}$$

For

$$\delta=B\sqrt{\frac{x}{8}}\left((1+\eta)^{1/4}+(1+\eta)^{-1/4}\right),$$

and using Equation (16), this yields

$$\begin{aligned}\mathbb{P}\left(\frac{S_t(\gamma)-M_t(\gamma)}{\sqrt{N_t(\gamma^2)}}>\delta\right)&\le D\exp\left(-\frac{8\delta^2}{B^2\left((1+\eta)^{1/4}+(1+\eta)^{-1/4}\right)^2}\right)\\&\le D\exp\left(-\frac{2\delta^2}{B^2}\left(1-\frac{\eta^2}{16}\right)\right)\ .\end{aligned}$$

■

Appendix C. Technical results

Lemma 25 *For any $i\in\{1,\dots,K\}$ and for any positive integer τ, let $N_{t-\tau:t}(1,i)=\sum_{s=t-\tau+1}^{t}\mathbb{1}_{\{I_t=i\}}$. Then for any positive m,*

$$\sum_{t=K+1}^{T}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}\le K\lceil T/\tau\rceil m\ .$$

Proof

$$\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}\le\sum_{j=1}^{\lceil T/\tau\rceil}\sum_{t=(j-1)\tau+1}^{j\tau}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}.$$

For any given $j\in\{1,\dots,\lceil T/\tau\rceil\}$, either $\sum_{t=(j-1)\tau+1}^{j\tau}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}=0$ or there exists an index $t\in\{(j-1)\tau+1,\dots,j\tau\}$ such that $I_t=i$, $N_{t-\tau:t}(1,i)<m$. In this case, we put $t_j=\max\{t\in\{(j-1)\tau+1,\dots,j\tau\}:I_t=i,N_{t-\tau:t}(1,i)<m\}$, the last time this condition is met in the j-th block. Then,

$$\begin{aligned}\sum_{t=(j-1)\tau+1}^{j\tau}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}&=\sum_{t=(j-1)\tau+1}^{t_j}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}\\&\le\sum_{t=t_j-\tau+1}^{t_j}\mathbb{1}_{\{I_t=i,N_{t-\tau:t}(1,i)<m\}}\le\sum_{t=t_j-\tau+1}^{t_j}\mathbb{1}_{\{I_t=i\}}=N_{t_j-\tau:t_j}(1,i)<m.\end{aligned}$$

■

Corollary 26 *For any $i \in \{1, \dots, K\}$, any integers $\tau \geq 1$ and $A > 0$,*

$$\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t=i, N_t(\gamma,i)<A\}} \leq K\lceil T/\tau \rceil A\gamma^{-\tau} \ .$$

Proof Simply note that

$$\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t=i, N_t(\gamma,i)<A\}} \leq \sum_{t=1}^{T} \mathbb{1}_{\{I_t=i, N_{t-\tau:t}(1,i)<\gamma^{-\tau}A\}} \ , \tag{21}$$

and apply the preceeding lemma with $m = \gamma^{-\tau}A$. ■

Acknowledgment

The authors wish to thank Gilles Stoltz and Jean-Yves Audibert for stimulating discussions and for providing us with unpublished material.

References

R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. in Appl. Probab.*, 27(4):1054–1078, 1995. ISSN 0001-8678.

J-Y. Audibert, R. Munos, and A. Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, 2007.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77 (electronic), 2002/03. ISSN 0097-5397.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):397–422, 2002. ISSN 1532-4435.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002. URL `citeseer.ist.psu.edu/auer00finitetime.html`.

Nicolò Cesa-Bianchi and Gábor Lugosi. On prediction of individual sequences. *Ann. Statist.*, 27(6):1865–1895, 1999. ISSN 0090-5364.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Math. Oper. Res.*, 31(3):562–580, 2006. ISSN 0364-765X.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Competing with typical compound actions, 2008.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1, part 2):119–139, 1997. ISSN 0022-0000. Second Annual European Conference on Computational Learning Theory (EuroCOLT '95) (Barcelona, 1995).

C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits, 2006. NIPS-2006 workshop, Online trading between exploration and exploitation, Whistler, Canada.

M. Herbster and M.K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998. ISSN 0885-6125. doi: http://dx.doi.org/10.1023/A:1007424614876.

L. Kocsis and C. Szepesvári. Discounted UCB. 2nd PASCAL Challenges Workshop, Venice, Italy, April 2006.

D. E. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913–922, 2008.

S. R. Kulkarni and G. Lugosi. Finite-time lower bounds for the two-armed bandit problem. *IEEE Trans. Automat. Control*, 45(4):711–714, 2000. ISSN 0018-9286.

L. Lai, H. El Gamal, H. Jiang, and H. V. Poor. Cognitive medium access: Exploration, exploitation and competition, 2007. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0710.1385`.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6(1): 4–22, 1985. ISSN 0196-8858.

A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits, 2008. URL `http://research.microsoft.com/users/slivkins/`. submitted to COLT 2008.

P. Whittle. Restless bandits: activity allocation in a changing world. *J. Appl. Probab.*, Special Vol. 25A: 287–298, 1988. ISSN 0021-9002. A celebration of applied probability.

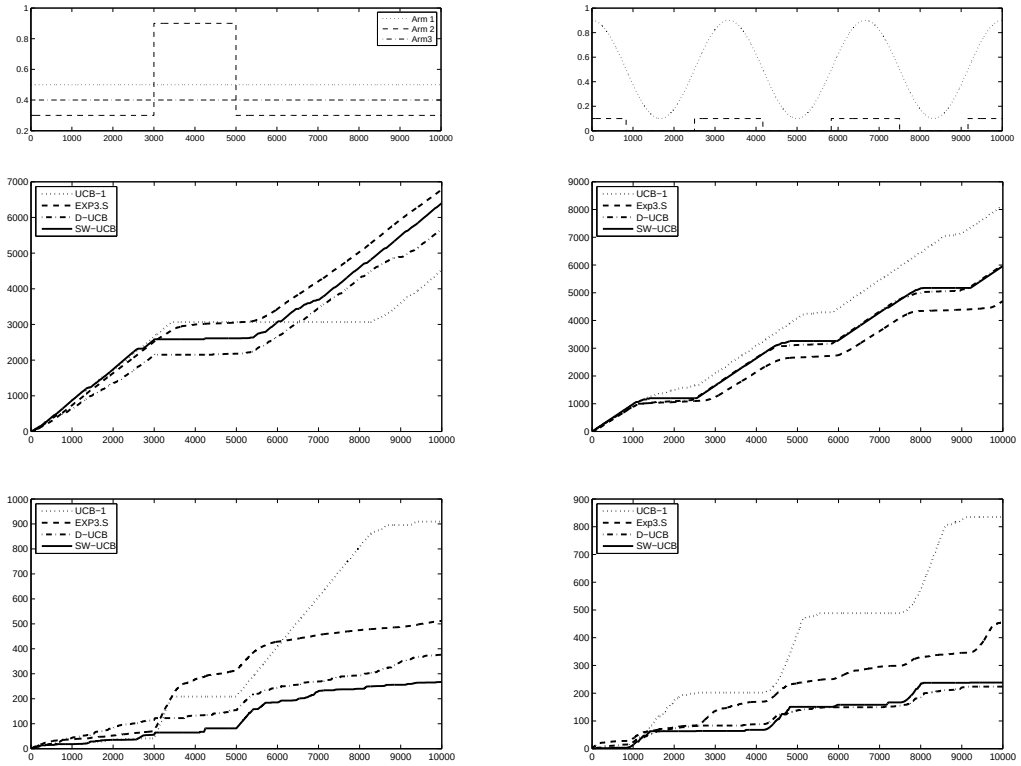

Figure 1: Left panel: Bernoulli MAB problem with two swaps. Upper: evolution of the probability of having a reward 1 for each arm; Middle: cumulative frequency of arm 1 pulls for each policy. Below: cumulative regret of each policy. Right panel: Bernoulli MAB problem with periodic rewards: Upper: evolution of the probability of having a reward 1 for arm 1 (and time intervals when it should be played); Middle: cumulative frequency of arm 1 pulls for each policy. Below: cumulative regret of each policy.