On Upper-Confidence Bound Policies for Switching Bandit Problems

Aurélien Garivier and Eric Moulines

Institut Telecom, Telecom ParisTech, Laboratoire LTCI, CNRS UMR 5141
46 rue Barrault, 75634 Paris Cedex 13

Abstract. Many problems, such as cognitive radio, parameter control of a scanning tunnelling microscope or internet advertisement, can be modelled as non-stationary bandit problems where the distributions of rewards changes abruptly at unknown time instants. In this paper, we analyze two algorithms designed for solving this issue: *discounted UCB* (D-UCB) and *sliding-window UCB* (SW-UCB). We establish an upper-bound for the expected regret by upper-bounding the expectation of the number of times suboptimal arms are played. The proof relies on an interesting 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. Numerical simulations show that D-UCB and SW-UCB perform significantly better than existing soft-max methods like EXP3.S.

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 agent chooses one arm and receives a reward accordingly. In the stationary case, the distribution of the rewards are initially unknown, but are assumed to remain constant during all games. The agent aims at minimizing 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. For several algorithms in the literature (e.g. [20, 1]), as the number of plays T tends 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 and improvements have been derived (see [5, 2, 16]), but those improvements do not address the issue of non-stationarity.

Though the stationary formulation of the MAB 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

J. Kivinen et al. (Eds.): ALT 2011, LNAI 6925, pp. 174–188, 2011.

problem [19], a user wishes to opportunistically exploit the availability of an empty channel in a multiple channel 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 maximizing 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 [14], see also [18] 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, non-stationary MAB problems have been considered (see [17, 14, 22, 24]), where distributions of rewards may change in time. Motivated by the problems cited above, and following a paradigm widely used in the change-point detection literature (see [12, 21] and references therein), we focus on non-stationary environments where the distributions of the rewards undergo abrupt changes. We show in the following that, as expected, policies tailored for the stationary case fail to track changes of the best arm.

Section 2 contains the formal presentation of the non-stationary setting we consider, together with two algorithms adressing this exploration/exploitation dilemma : D-UCB and SW-UCB. D-UCB had been proposed in [17] with empirical evidence of efficiency, but no theoretical analysis. SW-UCB is a new UCB-like algorithm that appears to perform slightly better in switching environments. In Section 3, we provide upper-bounds on the performance of D-UCB and SW-UCB; moreover, we provide a lower-bound on the performance of any algorithm in abruptly changing environments, that almost matches the upper-bounds. 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 the presence of switches. D-UCB is analyzed in Section 4; it relies on a novel deviation inequality for self-normalized averages with random number of summands which is stated in Section 7 together with some technical results. A lower bound on the regret of any algorithm in an abruptly changing environment is given in Section 5. In Section 6, two simple Monte-Carlo experiments are presented to support our findings.

2 Algorithms

In the sequel, we assume that the set of arms is $\{1, \ldots, K\}$, and that the rewards $\{X_t(i)\}_{t \geq 1}$ for arm $i \in \{1, \ldots, K\}$ are modeled by a sequence of independent random variables from potentially different distributions (unknown to the user) which may vary across time but remain bounded by $B > 0$. For each $t > 0$, we denote by $\mu_t(i)$ the expectation of the reward $X_t(i)$ for arm i. Let i_t^* be the arm with highest expected reward at time t (in case of ties, let i_t^* be one of the arms with highest expected rewards). The regret of a policy π is 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 [3] for the non-stochastic bandit problem).

We consider *abruptly changing environments*: the distributions of rewards remain constant during periods and change at unknown time instants called *breakpoints* (which do not depend on the policy of the player or on the sequence of rewards). In the following, we denote by Υ_T the number of breakpoints in the reward distributions that occur before time T. Another type of non-stationary MAB, where the distribution of rewards changes continuously, is considered in [22].

Standard soft-max and UCB policies are not appropriate for abruptly changing environments: as stressed in [14], "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, [3] and [8, 9] have proposed an adaptation of the Fixed-Share algorithm referred to as *EXP3.S* (see [15, 6] and the references therein). Theorem 8.1 and Corollary 8.3 in [3] state that when EXP3.S is tuned properly (which requires in particular that Υ_T is known in advance), the expected regret satisfies $\mathbb{E}_\pi\left[R_T\right] \leq 2\sqrt{\mathrm{e}-1}\sqrt{KT(\Upsilon_T \log(KT)+\mathrm{e})}$. Despite the fact that it holds uniformly over all reward distributions, such an upper-bound may seem deceptive in comparison to the stationary case,: the rate $O(\sqrt{T\log T})$ is much larger than the $O(\log T)$ achievable for a fixed distribution in the absence of changes. But actually, we prove in Section 5 that no policy can always achieve an average fixed-game regret smaller than $O(\sqrt{T})$ in the non-stationary case. Hence, EXP3.S matches the best achievable rate up to a factor $\sqrt{\log T}$. By construction, this algorithm can as well be used in an adversarial setup; but, in a stochastic environment, it is not guaranteed to be optimal (think that, in the stationary case, UCB outperforms EXP3 in the stochastic setup), and specific methods based on probabilistic estimation have to be considered.

In fact, in the family of UCB policies, several attempts have been made; see for examples [22] and [17]. In particular, [17] 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

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

where the discounted exploration bonus is $c_t(\gamma,i) = 2B\sqrt{\xi\log n_t(\gamma)/N_t(\gamma,i)}$, with $n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma,i)$, for an appropriate parameter ξ. Using these notations, discounted-UCB (D-UCB) is defined in Algorithm 1. For $\gamma = 1$, D-UCB boils down to the standard UCB-1 algorithm.

Algorithm 1. Discounted UCB

```
for $t$ from 1 to $K$, play arm $I_t = t$;
for $t$ from $K+1$ to $T$, play arm
    $I_t = \arg\max_{1\le i\le K} \bar{X}_t(\gamma, i) + c_t(\gamma, i).$
```

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 the whole past with a discount factor, *sliding-window UCB* relies on a local empirical average of the observed 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

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

and the exploration bonus is defined as $c_t(\tau, i) = B\sqrt{\xi \log(t \wedge \tau)/(N_t(\tau, i))}$, where $t \wedge \tau$ denotes the minimum of t and τ, and ξ is an appropriate constant. The policy defined in Algorithm 2 is denoted *Sliding-Window UCB* (SW-UCB).

Algorithm 2. Sliding-Window UCB

```
for $t$ from 1 to $K$, play arm $I_t = t$;
for $t$ from $K+1$ to $T$, play arm
    $I_t = \arg\max_{1\le i\le K} \bar{X}_t(\tau, i) + c_t(\tau, i),$
```

3 Regret Bounds

In this section, we provide upper-bounds on the regret of D-UCB and SW-UCB, as well as an almost matching lower-bound on the regret of any algorithm facing an abruptly changing environment.

Let Υ_T denote the number of breakpoints before time T, and let $\tilde{N}_T(i) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*\}}$ denote the number of times arm i was played when it was not the best arm during the T first rounds. Denote by $\Delta\mu_T(i)$ the minimum of the difference of expected reward of the best arm $\mu_t(i_t^*)$ and the expected reward $\mu_t(i)$ of arm i for all times $t \in \{1, \dots, T\}$ such that arm i is not optimal:

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

We denote by $\mathbb{P}_\gamma$ and $\mathbb{E}_\gamma$ the probability distribution and expectation under the policy D-UCB using the discount factor γ. As the expected regret is

$$\mathbb{E}_\gamma\left[R_T\right] = \mathbb{E}_\gamma\left[\sum_{t=1}^{T} \sum_{i:\mu_t(i)<\mu_t(i_t^*)} \left(X_t(i_t^*) - X_t(i)\right) \mathbb{1}_{\{I_t=i\}}\right] \le B \sum_{i=1}^{K} \mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] ,$$

it is sufficient to upper-bound the expected number of times an arm i is played when this arm is suboptimal.

Theorem 1. *Let $\xi \in (1/2, 1)$ and $\gamma \in (1/2, 1)$. For any $T \ge 1$ and for any arm $i \in \{1, \ldots, K\}$:*

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] \le \mathrm{C}_1\, T(1-\gamma)\log\frac{1}{1-\gamma} + \mathrm{C}_2\,\frac{\Upsilon_T}{1-\gamma}\log\frac{1}{1-\gamma} , \tag{2}$$

where

$$\mathrm{C}_1 = \frac{32\sqrt{2}B^2\xi}{\gamma^{1/(1-\gamma)}(\Delta\mu_T(i))^2} + \frac{4}{(1-\frac{1}{\mathrm{e}})\log\left(1+4\sqrt{1-1/2\xi}\right)}$$

and

$$\mathrm{C}_2 = \frac{\gamma-1}{\log(1-\gamma)\log\gamma} \times \log\left((1-\gamma)\xi\log n_K(\gamma)\right) .$$

When γ goes to 1, $\mathrm{C}_2 \to 1$ and

$$\mathrm{C}_1 \to \frac{16\,\mathrm{e}\,B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{(1-\mathrm{e}^{-1})\log\left(1+4\sqrt{1-1/2\xi}\right)} .$$

Algorithm SW-UCB shows a similar behavior, but the absence of infinite memory makes it slightly more suited to abrupt changes of the environment. Denote by $\mathbb{P}_\tau$ and $\mathbb{E}_\tau$ the probability distribution and expectation under policy SW-UCB with window size τ. The following bound holds:

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

$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] \le \mathrm{C}(\tau)\frac{T\log\tau}{\tau} + \tau\Upsilon_T + \log^2(\tau) , \tag{3}$$

where

$$\mathrm{C}(\tau) = \frac{4B^2\xi}{(\Delta\mu_T(i))^2}\frac{\lceil T/\tau\rceil}{T/\tau} + \frac{2}{\log\tau}\left\lceil\frac{\log(\tau)}{\log(1+4\sqrt{1-(2\xi)^{-1}})}\right\rceil$$
$$\to \frac{4B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{\log(1+4\sqrt{1-(2\xi)^{-1}})} \quad \textit{as } \tau \textit{ and } T/\tau \textit{ go to infinity.}$$

3.1 Tuning the Parameters

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. Choosing $\gamma = 1 - (4B)^{-1}\sqrt{\Upsilon_T/T}$ 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 3 stated below, up to a factor $\log T$.

Similary, choosing $\tau = 2B\sqrt{T\log(T)/\Upsilon_T}$ in SW-UCB 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 $O\left(T^{(1+\beta)/2}\sqrt{\log T}\right)$. 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 3 up to a factor $\sqrt{\log T}$, slightly better than the D-UCB.

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 allows to calibrate the discount factor γ as a function of the density of the breakpoint: with $\gamma = 1 - \sqrt{r}/(4B)$ we get an upper-bound with a dominant term in $-\sqrt{r}\log(r)O(T)$.

Concerning SW-UCB, 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 2 gives $\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] = O\left(T\sqrt{-r\log(r)}\right)$. 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 and τ. Namely, for t and k such that $2^k \leq t < 2^{k+1}$, take $\gamma = 1 - (4B)^{-1}(2^k)^{(\beta-1)/2}$.

If there is no breakpoint ($\Upsilon_T = 0$), the best choice is obviously to make the window as 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 $\mathbb{E}_{\text{UCB}}\left[\tilde{N}_T(i)\right] \leq \frac{2B^2}{(\Delta\mu(i))^2}\log(T)\left(1 + o(1)\right)$. This result improves by a constant factor the bound given in Theorem 1 in [5]. In [13], another constant factor is gained by using a different proof.

4 Analysis of D-UCB

Because of space limitations, we present only the analysis of D-UCB, i.e. the proof of Theorem 1. The case of SW-UCB is similar, although slightly more simple because of the absence of bias at a large distance of the breakpoints.

Compared to the standard regret analysis of the stationary case (see e.g. [5]), there are 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, which is stated in Section 7. The proof is in 5 steps:

Step 1. The number of times a suboptimal arm i is played is:

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

where $A(\gamma) = 16 B^2 \xi \log n_T(\gamma) / (\Delta\mu_T(i))^2$. Using Lemma 1 (see Section 7), we may upper-bound the first sum in the RHS as $\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) < A(\gamma)\}} \leq \lceil T(1-\gamma) \rceil A(\gamma) \gamma^{-\frac{1}{1-\gamma}}$. For a number of rounds (which depends on γ) following a breakpoint, the estimates of the expected rewards can be poor for $D(\gamma) = \log\left((1-\gamma)\xi \log n_K(\gamma)\right) / \log(\gamma)$ rounds. For any positive T, we denote by $\mathcal{T}(\gamma)$ the set of all indices $t \in \{K+1, \dots, T\}$ such that for all integers $s \in]t - D(\gamma), t]$, for all $j \in \{1, \dots, K\}$, $\mu_s(j) = \mu_t(j)$. In other words, t is in $\mathcal{T}(\gamma)$ if it does not follow too soon after a state transition. This 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:

$$\tilde{N}_T(i) \leq 1 + \lceil T(1-\gamma) \rceil A(\gamma) \gamma^{-1/(1-\gamma)} + \Upsilon_T D(\gamma) + \sum_{t \in \mathcal{T}(\gamma)} \mathbb{1}_{\{I_t = i \neq i_t^*, N_t(\gamma,i) \geq A(\gamma)\}} \ . \tag{4}$$

Step 2. Let $t \in \mathcal{T}(\gamma)$. If the following three things were true:

$$\begin{cases} \bar{X}_t(\gamma, i) + c_t(\gamma, i) < \mu_t(i) + 2c_t(\gamma, i) \\ \mu_t(i) + 2c_t(\gamma, i) < \mu_t(i_t^*) \\ \mu_t(i_t^*) < \bar{X}_t(\gamma, i_t^*) + c_t(\gamma, i_t^*) \end{cases}$$

then $\bar{X}_t(\gamma, i) + c_t(\gamma, i) < \bar{X}_t(\gamma, i_t^*) + c_t(\gamma, i_t^*)$, and arm i^* would be chosen. Thus,

$$\{I_t = i \neq i_t^*, N_t(\gamma, i) \geq A(\gamma)\} \subseteq \begin{cases} \{\mu_t(i_t^*) - \mu_t(i) \leq 2c_t(\gamma, i), N_t(\gamma, i) \geq A(\gamma)\} \\ \cup \left\{\bar{X}_t(\gamma, i_t^*) \leq \mu_t(i_t^*) - c_t(\gamma, i_t^*)\right\} \\ \cup \left\{\bar{X}_t(\gamma, i) \geq \mu_t(i) + c_t(\gamma, i)\right\} \end{cases} \tag{5}$$

In words, playing the suboptimal arm i at time t may occur in three cases: if $\mu_t(i)$ is substantially over-estimated, if $\mu_t(i_t^*)$ is substantially under-estimated, or if $\mu_t(i)$ and $\mu_t(i_t^*)$ are close to each other. But for the choice of $A(\gamma)$ given above, we have $c_t(\gamma, i) \leq 2B\sqrt{\left(\xi \log n_t(\gamma)\right) / A(\gamma)} \leq \Delta\mu_T(i)/2$, and the event $\{\mu_t(i_t^*) - \mu_t(i) < 2c_t(\gamma, i), N_t(\gamma, i) \geq A(\gamma)\}$ never occurs.

In Steps 3 and 4 we upper-bound the probability of the first two events of the RHS of (5). 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 event $\mathcal{E}_t(\gamma, j) = \left\{\bar{X}_t(\gamma, i) \geq \mu_t(j) + c_t(\gamma, j)\right\}$. 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 $t \in \mathcal{T}(\gamma)$,

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

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

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

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 exploration bonus. The other half of the exploration bonus 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 4 and the fact that $N_t(\gamma, j) \geq N_t(\gamma^2, j)$, we get:

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

Step 5. Hence, we finally obtain from Equation (4) that for all positive η:

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

When $\Upsilon_T \neq 0$, γ is taken strictly smaller than 1. 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 $\tau = (1-\gamma)^{-1}$,

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

and we obtain the statement of the Theorem.

5 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 3 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 $\mathbb{P}_\pi$.

In the sequel, we divide the period $\{1, \ldots, T\}$ into epochs of the same 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 $1 \leq j \leq M = \lfloor \frac{T}{\tau} \rfloor$, 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 $\{1, \ldots, M\}$, i.e. $\mathbb{P}_\pi^*$ 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 $\mathbb{P}_\pi^*$: $\mathbb{E}_\pi^*[W] = M^{-1} \sum_{j=1}^{M} \mathbb{E}_\pi^j[W]$. In the following, we lower-bound the expected regret of any policy π under $\mathbb{P}_\pi^*$ in terms of its regret under $\mathbb{P}_\pi$.

Theorem 3. *For any horizon T such that $64/(9\alpha) \leq \mathbb{E}_\pi[N_T(K)] \leq T/(4\alpha)$ and for any policy π ,*

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

where $C(\mu) = 2\delta(\mu(1) - \mu(K))/(3\alpha)$.

Proof. The main ingredients of this reasoning are inspired by the proof of Theorem 5.1 in [3].First, note that the Kullback-Leibler divergence $\mathrm{D}(\mathbb{P}_\pi; \mathbb{P}_\pi^j)$ is:

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

But $\mathbb{E}_\pi^j[N^j(K)] - \mathbb{E}_\pi[N^j(K)] \leq \tau d_{TV}(\mathbb{P}_\pi^j, \mathbb{P}_\pi) \leq \tau\sqrt{\mathrm{D}(\mathbb{P}_\pi; \mathbb{P}_\pi^j)/2}$ by Pinsker's inequality, and thus $\mathbb{E}_\pi^j[N^j(K)] \leq \mathbb{E}_\pi[N^j(K)] + \tau\sqrt{\alpha\mathbb{E}_\pi[N^j(K)]/2}$. Consequently, since $\sum_{j=1}^{M} N^j(K) \leq N_T(K)$,

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

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

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

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

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

Maximizing the right hand side of the previous inequality by choosing $\tau = 8T/(9\alpha\mathbb{E}_\pi[N_T(K)])$ or equivalently $M = 9\alpha/(8\mathbb{E}_\pi[N_T(K)])$ leads to the lower-bound:

$$\mathbb{E}^*_\pi[R_T] \geq \frac{16\delta}{27\alpha}\left(1 - \frac{\alpha\mathbb{E}_\pi[N_T(K)]}{T}\right)^2\left(1 - \frac{8}{9\alpha\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[N_T(K)]}\ .$$

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

$$\frac{16\delta(\mu(1)-\mu(K))}{27\alpha}\left(1 - \frac{\alpha\mathbb{E}_\pi[N_T(K)]}{T}\right)^2\left(1 - \frac{8}{9\alpha\mathbb{E}_\pi[N_T(K)]}\right)\frac{T}{\mathbb{E}_\pi[R_T]}\ ,$$

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 3, although it can also be proved directly.

Corollary 1. *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 3 implies that:

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

In words, Theorem 3 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 $\mathbb{E}^*_\pi[R_T] \geq cT/\log(T)$ for some positive c depending on the reward distribution. To keep notation simple, Theorem 3 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 [3]).

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 [20] 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 [3]. 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 3, 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.

6 Simulations

The scope of this section is to present two simple, archetypal settings that show the interest of UCB methods in non-stationary stochastic environments. In the first example, there are $K = 3$ arms and the time horizon is set to $T = 10^4$. 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 $t \in \{1, \dots, T\}$, $p_t(3) = 0.4$ for $t < 3000$ or $t \geq 5000$, and $p_t(3) = 0.9$ for $3000 \leq t < 5000$. 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 Figure 1 , we represent the evolution of the cumulated regret. We compare the UCB-1 algorithm with $\xi = \frac{1}{2}$, the EXP3.S algorithm described in [3] 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 $\xi = 0.6$ and $\gamma = 1 - 1/4\sqrt{T}$ and the SW-UCB with $\xi = 0.6$ and $\tau = 4\sqrt{T \log T}$ (chosen according to Section 3).

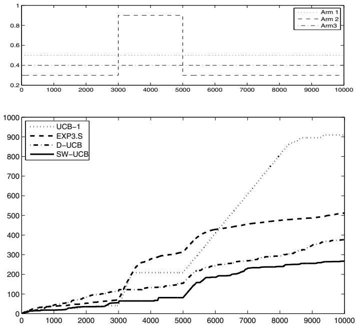

Fig. 1. Bernoulli MAB problem with two swaps. Above: evolution of the reward distributions. Below: cumulative regret of each policy.

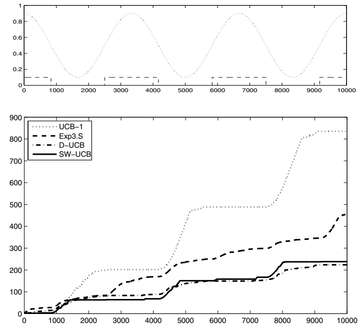

Fig. 2. Bernoulli MAB problem with periodic rewards. Above: evolution of reward distribution of arm 1. Below: cumulative regret of each policy.

As can be seen in Figure 1 (and as can be consistently observed over all 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, we test the behaviour of D-UCB and SW-UCB by investigating their performance in a slowly-varying environment. This environment is made of $K = 2$ arms, the rewards are still Bernoulli random variables with parameters $p_t(i)$ but they 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: $p_t(1) = 0.5 + 0.4 \cos(6\pi Rt/T)$. Hence, the best arm to pull changes cyclically and the transitions are smooth (regularly, the two arms are statistically indistinguishable). In Figure 2 , 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. Continuing the experiment or multiplying the changes only confirms this conclusion.

These modest and yet representative examples suggest that, despite the fact that similar regret bounds are proved for D-UCB, SW-UCB and EXP3.S, the two former methods are significantly more reactive to changes in practice and have a better performance, whether the environment is slowly or abruptly changing. EXP3.S, on the other hand, is expected to be more robust and more adapted to non stochastic (and non-oblivious) environments.

7 Technical Results

We first state a deviation bound for self-normalized discounted average, of independent interest, that proves to be a key ingredient in the analysis of D-UCB. Let $(X_t)_{t\geq 1}$ be a sequence of non-negative independent random variables bounded by B defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and we denote $\mu_t = \mathbb{E}[X_t]$. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields of $\mathcal{A}$ such that for each t, $\sigma(X_1 \dots, X_t) \subset \mathcal{F}_t$ 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). For $\gamma \in [0, 1)$, let $S_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} X_s \epsilon_s$, $M_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \mu_s \epsilon_s N_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s$, and $n_t(\gamma) = \sum_{s=1}^{t} \gamma^{t-s}$.

Theorem 4. *For all integers t and all $\delta, \eta > 0$,*

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

The following lemma is required in the analysis of SW-UCB and D-UCB:

Lemma 1. *Let $i \in \{1, \ldots, K\}$; 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,*

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

Thus, for any $\tau \geq 1$ and $A > 0$, $\sum_{t=K+1}^{T} \mathbb{1}_{\{I_t=i, N_t(\gamma,i)<A\}} \leq \lceil T/\tau \rceil A\gamma^{-\tau}$.

The proof of these results is omitted due to space limitations.

8 Conclusion and Perspectives

This paper theoretically establishes that the UCB policies can be successfully adapted to cope with non-stationary environments. It is shown introducing two breakpoints is enough to move from the $\log(T)$ performance of stationary bandits to the $\sqrt{T \log(T)}$ performance of adversarial bandits. The upper bound of the D-UCB and SW-UCB in an abruptly changing environment are similar to the upper bounds of the EXP3.S algorithm, and numerical experiments show that UCB policies outperform the softmax methods in stochastic environments. The focus of this paper is on an 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, continuously evolving reward distributions can be analysed. Furthermore, the deviation inequality for discounted martingale transforms given in Section 7 is a powerful tool of independent interest.

As the previously reported Exp3.S algorithm, the performance of the proposed policy depends on tuning parameters (discount factor or window). Designing a fully adaptive algorithm, able to actually detect the changes as they occur with no prior knowledge of a typical inter-arrival time, is not an easy task and remains the subject of on-going research. A possibility may be to tune adaptively the parameters by using data-driven approaches, as in [14]. Another possibility is to use margin assumptions on the gap between the distributions before and after the changes, as in [24]: at the price of this extra assumption, one obtains improved bounds without the need for the knowledge of the number of changes.

References

[1] Agrawal, R.: 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)

[2] Audibert, J.Y., Munos, R., Szepesvari, A.: Tuning bandit algorithms in stochastic environments. In: Hutter, M., Servedio, R.A., Takimoto, E. (eds.) ALT 2007. LNCS (LNAI), vol. 4754, pp. 150–165. Springer, Heidelberg (2007)

[3] Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM J. Comput. 32(1), 48–77 (2002)

[4] Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res. 3(Spec. Issue Comput. Learn. Theory), 397–422 (2002)

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

[6] Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, New York (2006)

[7] Cesa-Bianchi, N., Lugosi, G.: On prediction of individual sequences. Ann. Statist. 27(6), 1865–1895 (1999)

[8] Cesa-Bianchi, N., Lugosi, G., Stoltz, G.: Regret minimization under partial monitoring. Math. Oper. Res. 31(3), 562–580 (2006)

[9] Cesa-Bianchi, N., Lugosi, G., Stoltz, G.: Competing with typical compound actions (2008)

[10] Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition. Applications of Mathematics, vol. 31. Springer, New York (1996)

[11] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55(1, part 2), 119–139 (1997); In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904. Springer, Heidelberg (1995)

[12] Fuh, C.D.: Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. Ann. Statist. 32(5), 2305–2339 (2004)

[13] Garivier, A., Cappé, O.: The kl-ucb algorithm for bounded stochastic bandits and beyond. In: Proceedings of the 24rd Annual International Conference on Learning Theory (2011)

[14] Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., Sebag, M.: Multi-armed bandit, dynamic environments and meta-bandits. In: nIPS-2006 Workshop, Online Trading Between Exploration and Exploitation, Whistler, Canada (2006)

[15] Herbster, M., Warmuth, M.: Tracking the best expert. Machine Learning 32(2), 151–178 (1998)

[16] Honda, J., Takemura, A.: An asymptotically optimal bandit algorithm for bounded support models. In: Proceedings of the 23rd Annual International Conference on Learning Theory (2010)

[17] Kocsis, L., Szepesvári, C.: Discounted UCB. In: 2nd PASCAL Challenges Workshop, Venice, Italy (April 2006)

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

[19] Lai, L., El Gamal, H., Jiang, H., Poor, H.V.: Cognitive medium access: Exploration, exploitation and competition (2007)

[20] Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Adv. in Appl. Math. 6(1), 4–22 (1985)

[21] Mei, Y.: Sequential change-point detection when unknown parameters are present in the pre-change distribution. Ann. Statist. 34(1), 92–122 (2006)

[22] Slivkins, A., Upfal, E.: Adapting to a changing environment: the brownian restless bandits. In: Proceedings of the Conference on 21st Conference on Learning Theory, pp. 343–354 (2008)

[23] Whittle, P.: Restless bandits: activity allocation in a changing world. J. Appl. Probab. Special 25A, 287–298 (1988) a celebration of applied probability

[24] Yu, J.Y., Mannor, S.: Piecewise-stationary bandit problems with side observations. In: ICML 2009: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1177–1184. ACM, New York (2009)