Use of variance estimation in the multi-armed bandit problem

Jean Yves Audibert
CERTIS - Ecole des Ponts
19, rue Alfred Nobel - Cité Descartes
77455 Marne-la-Vallée - France
audibert@certis.enpc.fr

Rémi Munos
INRIA Futurs, Grappa
Université Lille 3, France
remi.munos@inria.fr

Csaba Szepesvári
Computer and Automation Research Institute
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
szcsaba@sztaki.hu

Abstract

An important aspect of most decision making problems concerns the appropriate balance between exploitation (acting optimally according to the partial knowledge acquired so far) and exploration of the environment (acting sub-optimally in order to refine the current knowledge and improve future decisions). A typical example of this so-called *exploration versus exploitation dilemma* is the multi-armed bandit problem, for which many strategies have been developed. Here we investigate policies based the choice of the arm having the highest upper-confidence bound, where the bound takes into account the *empirical variance* of the different arms. Such an algorithm was found earlier to outperform its peers in a series of numerical experiments. The main contribution of this paper is the theoretical investigation of this algorithm. Our contribution here is twofold. First, we prove that with probability at least $1-\beta$, the regret after n plays of a variant of the UCB algorithm (called β-UCB) is upper-bounded by a constant, that scales linearly with $\log(1/\beta)$, but which is independent from n. We also analyse a variant which is closer to the algorithm suggested earlier. We prove a logarithmic bound on the expected regret of this algorithm and argue that the bound scales favourably with the variance of the suboptimal arms.

1 Introduction and notations

A *K-armed bandit problem* ($K \geq 2$) is defined by random variables $X_{k,t}$ ($1 \leq k \leq K$, $t \in \mathbb{N}^+$), where each k is the index of an "arm" of the bandit and t represents time. Successive plays of arm k yield rewards $X_{k,1}, X_{k,2}, \ldots$ which are independent and identically distributed (i.i.d.) according to an unknown distribution. Independence also holds for rewards across the different arms, i.e. for any $t \in \mathbb{N}^+$ and $1 \leq k < k' \leq K$, $(X_{k,1}, \ldots, X_{k,t})$ and $(X_{k',1}, \ldots, X_{k',t})$ are independent. Let μ_k and σ_k^2 be respectively the (unknown) expectation and variance of the rewards coming from arm k. For any $k \in \{1, \ldots, K\}$ and $t \in \mathbb{N}$, let $\overline{X}_{k,t}$ and $V_{k,t}$ be their respective empirical estimates:

$$\overline{X}_{k,t} \triangleq \frac{1}{t} \sum_{i=1}^{t} X_{k,i}$$

and

$$V_{k,t} \triangleq \frac{1}{t}\sum_{i=1}^{t}(X_{k,i} - \overline{X}_{k,t})^2,$$

where by convention $\overline{X}_{k,0} \triangleq 0$ and $V_{k,0} \triangleq 0$. An *optimal arm* is an arm having the best expected reward

$$k^* \in \operatorname*{argmax}_{k\in\{1,\dots,K\}} \mu_k.$$

For the sake of simplicity we assume that there is a single optimal arm. The proofs and hence the results hold when there are multiple such arms. We denote quantities related to the optimal arm by putting $*$ in the upper index. In particular, $\mu^* = \max_k \mu_k$. The *expected regret* of an arm k is

$$\Delta_k \triangleq \mu^* - \mu_k.$$

A *policy* is a way of choosing the next arm to play based on the sequence of past plays and obtained rewards. More formally, it is a mapping from $\cup_{t\in\mathbb{N}}\{1,\dots,K\}^t \times \mathbb{R}^t$ into $\{1,\dots,K\}$. Let $T_k(t)$ be the number of times arm k is chosen by the policy during the first t plays. Let I_t denote the arm played by the policy at time t.

We define the *cumulative regret of the policy up to time n* as

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

We also define the *cumulative pseudo-regret*

$$R_n = \sum_{k=1}^{K} T_k(n)\Delta_k.$$

The *expected cumulative regret of the policy up to time n* is

$$\overline{R}_n \triangleq n\mu^* - \mathbb{E}\big[\sum_{t=1}^{n} X_{t,T_{I_t}(t)}\big] = \sum_{k=1}^{K}\mathbb{E}[T_k(n)]\Delta_k.$$

2 The β-UCB policy

2.1 The algorithm

Assume that the rewards are bounded. Then, without loss of generality, we may assume that all the rewards are almost surely in $[0,1]$. Let $0 < \beta < 1$ be some fixed confidence level. Consider the sub-confidence levels β_s defined as

$$\beta_s \triangleq \frac{\beta}{4Ks(s+1)} \tag{1}$$

Let

$$B_{k,s} \triangleq \left(\overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\log(\beta_s^{-1})}{s}} + \frac{16\log(\beta_s^{-1})}{3s}\right) \wedge 1$$

with the convention $1/0 = +\infty$.

β-UCB policy: At time t, play an arm maximizing $B_{k,T_k(t-1)}$.

Let us now explain the choice of $B_{k,T_k(t-1)}$. The quantity essentially comes from the following theorem.

Theorem 1 *Let $X_1,\dots,X_t$ be i.i.d. random variables taking their values in $[0;1]$. Let $\mu = \mathbb{E}X_1$ be their common expected value. Consider the empirical expectation μ_t and standard deviation $\sigma_t \geq 0$ defined respectively as*

$$\mu_t = \frac{\sum_{i=1}^{t} X_i}{t} \quad \text{and} \quad \sigma_t^2 = \frac{\sum_{i=1}^{t}(X_i - \mu_t)^2}{t}.$$

With probability at least $1-\beta$, we have

$$\mu \leq \mu_t + \sigma_t\sqrt{\frac{2\log(3\beta^{-1})}{t}} + \frac{16\log(3\beta^{-1})}{3t}. \tag{2}$$

and

$$\mu \geq \mu_t - \sigma_t\sqrt{\frac{2\log(3\beta^{-1})}{t}} - \frac{16\log(3\beta^{-1})}{3t}. \tag{3}$$

Proof 1 *See Section A.1.*

Note that (2) is useless for $t \leq 5$ since its r.h.s. is larger than 1. We may apply Theorem 1 to the rewards $X_{k,1}, \ldots, X_{k,s}$ and the confidence level $3\beta_s$. Since $\sum_{1 \leq k \leq K; t \geq 1} 3\beta_{k,t} = 3\beta/4$, it gives that with probability at least $1 - 3\beta/4 \geq 1 - \beta$, for any $s \in \mathbb{N}$ and $k \in \{1, \ldots, K\}$, we have $\mu_k \leq B_{k,s}$.It means that with confidence level β, for any time $t \geq 1$, after the first $t-1$ plays, the user of the policy knows that the expected reward of arm k is upper bounded by $B_{k,T_k(t-1)}$. The user of the β-UCB policy chooses his plays only through these upper confidence bounds (UCB).

2.2 Properties of the β-UCB policy

We start with a deviation inequality for the number of plays of non-optimal arms.

Theorem 2 *For any non-optimal arm k (i.e. $\Delta_k > 0$), consider u_k the smallest integer such that*

$$\frac{u_k}{\log[4Ku_k(u_k+1)\beta^{-1}]} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}. \tag{4}$$

With probability at least $1 - \beta$, the β-UCB policy plays any non-optimal arm k at most u_k times.

Proof 2 *See Section A.2.*

This means that with high probability, the number of plays of non-optimal arms is bounded by some quantity independent of the total number of plays.

Theorem 2 directly leads to upper bounds on the cumulative regret of the policy up to time n and on its expected value.

Since u_k depends on the parameter β, we will now write it $u_{k,\beta}$. The following lemma gives more explicit bounds on $u_{k,\beta}$.

Lemma 1 *Let* $w_k = \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}$. *We have* $u_{k,\beta} \leq 5w_k \log(w_k K\beta^{-1})$ *and* $u_{k,\beta} \leq w_k \log(4K\beta^{-1}) + 2w_k \log\left\{6w_k \log(w_k K\beta^{-1})\right\}$.

The first bound is the simplest but the least accurate. In the second one, the leading term is the first one (when β goes to 0).

Proof 3 *See Section A.3.*

Theorem 3 *With probability at least $1 - \beta$, for any time n, the cumulative regret of the β-UCB policy satisfies*

$$\sum_{k=1}^K T_k(n)\Delta_k \leq \sum_{k \in K} [u_{k,\beta} \wedge n]\Delta_k \tag{5}$$

Besides for any positive integer n, the expected cumulative regret of the $1/n$-UCB up to time n satisfies

$$\begin{array}{rcl} \sum_{k=1}^K \mathbb{E}[T_k(n)]\Delta_k & \leq & \sum_{k \in K} \left[(1 + u_{k,1/n}) \wedge n\right]\Delta_k \\ & \leq & C_1 \sum_{k \in K} \left\{ \left[\left(\frac{\sigma_k^2}{\Delta_k} + 1\right) \log(Kn) \right] \wedge \left[n\Delta_k\right] \right\} \\ & \leq & C_2 \log(2n) \sum_{k \neq k^*} \left(1 + \frac{\sigma_k^2}{\Delta_k}\right) \end{array} \tag{6}$$

for some universal constants C_1 and C_2.

Proof 4 *The first assertion is a direct consequence of Theorem 2. For the second assertion, the first inequality comes from $T_k(n) \leq n$ and $\mathbb{E}T_k(n) \leq \mathbb{P}[T_k(n) > u_k]n + \mathbb{P}[T_k(n) \leq u_k]u_k$. The second inequality uses Lemma 1. The third inequality follows by considering two cases: either $K > n$ (i.e. not enough time to explore all the arms), then the property is trivial, or $K \leq n$ which implies $\log(Kn) \leq 2\log(n)$ for any $n \geq 1$.*

3 Bounds for the expected regret

3.1 Adaptive β-UCB

As far as results in expectation are concerned (see second part of Theorem 3), we need to take β dependent on the time horizon (i.e. the total number of arms to be drawn) to obtain a regret bound of order $C \log n$.

Schematically an algorithm which needs to know its time horizon to have a $(\log n)$-cumulative regret bound up to time n can be modified into an adaptive algorithm having the same cumulative regret bound (up to a multiplicative constant close to 1). This is done by the doubling trick (see [3, p.33] and references within) in which we cut the time space into intervals of length 2^{2^k}. For each of this epoch, we launch the algorithm independently of what happens in the other epochs. The policy knows its time horizon and leads to a cumulative regret for epoch k of order 2^k. Summing these regrets up to some time horizon $T = 2^{2^K}$, we obtain a cumulative regret of order $\sum_{i=1}^{K} 2^i = 2^{K+1} - 1$, hence of order $\log T$.

For the β-UCB policy, we need neither to restart at each epoch the policy nor to cut the time space in epochs. Indeed, it suffices to take $\beta = 1/t$ at time t. To decrease β when the number of arms already drawn increases is natural: when the time t increases, the exploitation of an almost optimal arm becomes more and more detrimental to the quality of the policy, so we want to be a bit more sure that the optimal one is not in other arms. Consequently, we need to have better upper confidence bound, which means that β should be taken smaller. For this adaptive policy, one can show using the time cutting argument presented above that the results given in (6) still holds.

3.2 UCB-tuned policy

Define the confidence sequences of arm k

$$c_{t,s}^{(k)} \triangleq \sqrt{\frac{2V_{k,s}\log(4t^p)}{s}} + \frac{16\log(4t^p)}{3s}.$$

The following figure describes the UCB-tuned policy:

UCB-tuned policy: At time t, play an arm maximizing

$$\left(\overline{X}_{k,T_k(t-1)} + c_{t,T_k(t-1)}^{(k)}\right) \wedge 1.$$

A slight variation (with different confidence sequences) was proposed in Section 4 of [1]. In their experimental study these authors have found that this algorithms outperforms most previous algorithms under a wide range of conditions. However, no theoretical analysis of this algorithm has been attempted so far. (Theorem 4 of [1], which is a closely related result that applies to normally distributed payoffs only, is not a complete proof since it relies on some conjectures.) The next theorem shows that with $p > 2$, the regret of the above algorithm scales with the logarithm of the number of steps. A crucial feature of this result is that instead of scaling with $1/\Delta_j$, the regret scales with σ_j^2/Δ_j. This shows that the performance of UCB-tuned is less sensitive to whether the assumed payoff range is a good match to the true range of payoffs.

Theorem 4 *Let $p > 2$. For any time n, the expected regret of the UCB-tuned policy is bounded by*

$$\overline{R}_n \leq 16\big[\log(4) + p\log(n)\big] \sum_{k \neq k^*} \left(1 + \frac{\sigma_k^2}{2\Delta_k}\right) + 2\Big(1 + \frac{1}{p-2}\Big) \sum_{k \neq k^*} \Delta_k.$$

A Proofs of the results

A.1 Proof of Theorem 1

The following inequality is a direct consequence of Bernstein's inequality (see e.g. [2, p.124]).

Lemma 2 *Let $W_1, \dots, W_t$ be i.i.d. random variables taking their values in $[0;1]$. Let $E = \mathbb{E}W_1$ and $V = \mathbb{E}(W_1 - E)^2$ be the expectation and variance of these random variables. For any $\epsilon > 0$, with probability at least $1-\epsilon$, the empirical mean $\bar{E} = (\sum_{i=1}^t X_i)/t$ satisfies*

$$\bar{E} < E + \sqrt{\frac{2V\log(\epsilon^{-1})}{t}} + \frac{2\log(\epsilon^{-1})}{3t}. \tag{7}$$

To prove Theorem 1, we apply Lemma 2 for $\epsilon = \beta/3$ and three different i.i.d. random variables: $W_i = X_i$, $W_i = 1 - X_i$ and $W_i = 1 - (X_i - \mathbb{E}X_1)^2$. Let σ denote the standard deviation of X_1: $\sigma^2 \triangleq \mathbb{E}(X_i - \mathbb{E}X_1)^2$. Introduce $\mathcal{V} \triangleq \mathbb{V}\mathrm{ar}\left[(X_1 - \mathbb{E}X_1)^2\right]$. We obtain that with probability at least $1-\beta$, we simultaneously have

$$|\mu_t - \mu| \leq \left(\sigma\sqrt{\frac{2\log(3\beta^{-1})}{t}} + \frac{2\log(3\beta^{-1})}{3t}\right) \wedge 1 \tag{8}$$

and

$$\sigma^2 \leq \sigma_t^2 + (\mu - \mu_t)^2 + \sqrt{\frac{2\mathcal{V}\log(3\beta^{-1})}{t}} + \frac{2\log(3\beta^{-1})}{3t}. \tag{9}$$

Let δ be the r.h.s. of (8) and $L \triangleq \log(3\beta^{-1})/t$ Noting that $\mathcal{V} \leq \sigma^2$, we have

$$\sigma^2 \leq \sigma_t^2 + \delta^2 + \delta \leq \sigma_t^2 + 2\delta,$$

hence successively

$$\sigma^2 - 2\sigma\sqrt{2L} - 4L/3 - \sigma_t^2 \leq 0,$$

and

$$\sigma \leq \sqrt{2L} + \sqrt{\sigma_t^2 + 10L/3} \leq \sigma_t + (\sqrt{2} + \sqrt{10/3})\sqrt{L}.$$

Plugging this inequality in (8), we obtain

$$\mu \leq \mu_t + \sigma_t\sqrt{2L} + \left[2 + \sqrt{20/3} + 2/3\right]L \leq \mu_t + \sigma_t\sqrt{2L} + 16L/3$$

The reverse inequality is obtained in a similar way.

A.2 Proof of Theorem 2

Let $l_t \triangleq \log(\beta_t^{-1})$ (remember that $\beta_t = \beta/(4kt(t+1))$). Consider the event $\mathcal{E}$ on which

$$\forall t \in \mathbb{N}^+ \quad \forall k \in \{1, \dots, K\} \quad \begin{cases} \left|\overline{X}_{k,t} - \mu_k\right| & < & \sqrt{\frac{2\sigma_k^2 l_t}{t}} + \frac{2l_t}{3t} \\ \left|V_{k,t} - \sigma_k^2\right| & < & \sqrt{\frac{8\sigma_k^2 l_t}{t}} + \frac{4l_t}{3t} \end{cases} \tag{10}$$

Let us show that this event holds with probability at least $1-\beta$.

Proof 5 *We apply Lemma 2 with $\epsilon = \beta_t$ and different i.i.d. random variables: $W_i = X_{k,i}$, $W_i = 1 - X_{k,i}$, $W_i = (X_{k,i} - \mu_k)^2$ and $W_i = 1 - (X_{k,i} - \mu_k)^2$. We use that the variance of the last two random variables is bounded by $\mathbb{E}[(X_{k,1} - \mu_k)^4] \leq \sigma_k^2$ and that the empirical expectation of $(X_{k,i} - \mu_k)^2$ is*

$$\tfrac{1}{t}\textstyle\sum_{i=1}^t (X_{k,i} - \mu_k)^2 = V_{k,t} + (\overline{X}_{k,t} - \mu_k)^2.$$

We obtain that for any $t \in \mathbb{N}^+$ and $k \in \{1, \dots, K\}$, with probability at least $1 - \beta_t$

$$\begin{cases} \left|\overline{X}_{k,t} - \mu_k\right| < \sqrt{\frac{2\sigma_k^2 l_t}{t}} + \frac{2l_t}{3t} \\ \left|V_{k,t} + (\overline{X}_{k,t} - \mu_k)^2 - \sigma_k^2\right| < \sqrt{\frac{2\sigma_k^2 l_t}{t}} + \frac{2l_t}{3t} \end{cases}$$

Since $|\overline{X}_{k,t} - \mu_k| \leq 1$*, this last inequality leads to*

$$\left|V_{k,t} - \sigma_k^2\right| < \left|\overline{X}_{k,t} - \mu_k\right| + \sqrt{\tfrac{2\sigma_k^2 l_t}{t}} + \tfrac{2l_t}{3t},$$

which gives the second inequality of (10)*. Using an union bound, all these inequalities hold simultaneously with probability at least*

$$1 - 4\textstyle\sum_{k=1}^{K}\sum_{t\geq 1}\beta_t = 1 - \beta.$$

Now let us prove that on the event $\mathcal{E}$, we have $\mu_k \leq B_{k,t}$ and

$$B_{k,t} \leq \mu_k + \sigma_k\sqrt{\frac{8l_t}{t}} + \frac{8l_t}{t} \tag{11}$$

Proof 6 *For sake of simplicity, let us temporarily drop the* k *indices: e.g.* $B_{k,t}$*,* μ_k*,* σ_k^2 *and* $V_{k,t}$ *respectively become* B_t*,* μ*,* σ^2 *and* V_t*. Introduce* $L_t = \frac{l_t}{t}$*. By* (10)*,*

$$\sigma^2 - V_t < \sqrt{8\sigma^2 L_t} + \tfrac{4L_t}{3}.$$

Let $q(\sigma) = \sigma^2 - \sigma\sqrt{2L_t} + (-V_t - \frac{4L_t}{3})$*. Since* $q(\sigma)$ *is negative only between its two roots, the largest root gives a bound on the values* σ *can take when* $q(\sigma) < 0$ *(the "square root trick"):*

$$\sigma < \sqrt{2L_t} + \sqrt{V_t + \tfrac{10L_t}{3}} \leq \sqrt{V_t} + (1 + \sqrt{5/3})\sqrt{2L_t}.$$

Plugging this inequality in the first inequality of (10)*, we obtain*

$$\left|\mu_t - \mu\right| < \sqrt{2V_t L_t} + \tfrac{16L_t}{3},$$

and in particular $\mu \leq B_t$*. For the second part of the assertion, we use*

$$\left\{\begin{array}{l} \mu_t \leq \mu + \sigma\sqrt{2L_t} + \frac{2L_t}{3} \\ V_t \leq \sigma^2 + \sigma\sqrt{8L_t} + \frac{4L_t}{3} \leq \left(\sigma + \sqrt{2L_t}\right)^2 \end{array}\right.$$

and obtain

$$B_t \leq \mu + 2\sigma\sqrt{2L_t} + 8L_t,$$

which is the announced result.

Let $\check{K}$ be the set of non-optimal arms: $\check{K} = \left\{k \in K : \Delta_k > 0\right\}$. For any integers u_k where $k \in \check{K}$, we have

$$\begin{array}{l} \mathbb{P}[\exists k \in \check{K} \quad T_k(t) > u_k] \\ \quad = \mathbb{P}[\exists k \in \check{K} \quad T_k(t) > u_k; \mathcal{E}] + \mathbb{P}[\exists k \in \check{K} \quad T_k(t) > u_k; \mathcal{E}^c] \\ \quad \leq \mathbb{P}\big[\exists k \in \check{K} \quad \exists s < t \quad T_k(s) = u_k \text{ and } I_{s+1} = k; \mathcal{E}\big] + \mathbb{P}(\mathcal{E}^c) \\ \quad \leq \mathbb{P}\big[\exists k \in \check{K} \quad \exists s < t \quad T_k(s) = u_k \text{ and } B_{k,T_k(s)} \geq B_{k^*,T_k^*(s)}; \mathcal{E}\big] + \beta \\ \quad \leq \mathbb{P}\big[\exists k \in \check{K} \quad \exists s < t \quad B_{k,u_k} \geq \mu_{k^*} \text{ or } B_{k^*,T_k^*(s)} < \mu_{k^*}; \mathcal{E}\big] + \beta \\ \quad \leq \mathbb{P}\big[\exists k \in \check{K} \quad B_{k,u_k} \geq \mu_{k^*}; \mathcal{E}\big] + \mathbb{P}\big[\exists r < t \quad B_{k^*,r} < \mu_{k^*}; \mathcal{E}\big] + \beta \\ \quad \leq \mathbb{P}\Big(\exists k \in \check{K} \quad \mu_k + \sqrt{\frac{8\sigma_k^2 l_{k,u_k}}{u_k}} + \frac{8l_{k,u_k}}{u_k} \geq \mu_{k^*}\Big) + 0 + \beta \end{array} \tag{12}$$

The probability in this last r.h.s. is equal to zero provided that the u_k's are large enough. Precisely, we want u_k such that $t_k = \sqrt{l_{k,u_k}/u_k}$ satisfies

$$8t_k^2 + 2\sqrt{2\sigma_k^2}t_k - \Delta_k < 0,$$

equivalently dropping the k indices: $t < (\sqrt{2\sigma^2 + 8\Delta} - \sqrt{2\sigma^2})/8$. We get

$$\tfrac{u}{l_u} > \tfrac{64}{(\sqrt{2\sigma^2+8\Delta}-\sqrt{2\sigma^2})^2} = \tfrac{1}{\Delta^2}\left(\sqrt{2\sigma^2 + 8\Delta} + \sqrt{2\sigma^2}\right)^2.$$

This inequality is at least satisfied when

$$\tfrac{u}{l_u} > \tfrac{8\sigma^2}{\Delta^2} + \tfrac{16}{\Delta},$$

which ends the proof.

A.3 Proof of Lemma 1

Proof 7 *By the definition of u_k, we have $\frac{u_k-1}{\log[4Ku_k(u_k-1)\beta^{-1}]} \leq w_k$. This implies*

$$u_k \leq 1 + w_k \log[4Ku_k^2\beta^{-1}]. \tag{13}$$

Basic computations leads to $u_k \leq w_k^2 K\beta^{-1}$. This very rough upper bound can be used to have a tight upper bound of u_k. Indeed, after some simple computations using that $w_k \geq 16$, the first two recursive uses of (13) *gives*

$$u_k \leq 5w_k \log(w_k K\beta^{-1})$$

and

$$u_k \leq w_k \log(4K\beta^{-1}) + 2w_k \log\left\{6w_k \log(w_k K\beta^{-1})\right\},$$

which are the announced results.

B Proof of Theorem 4 (UCB-tuned's regret)

The choice of this confidence sequence is such that, for any fixed arm index k, any pairs (s,t) satisfying $1 \leq s \leq t$, the following holds:

$$\mathbb{P}(\mu_k \leq \overline{X}_{k,s} + c_{t,s}^{(k)}) \geq 1 - t^{-p}, \quad \mathbb{P}(\overline{X}_{k,s} + c_{t,s}^{(k)} \leq \mu_k + d_{t,s}^{(k)}) \geq 1 - t^{-p}. \tag{14}$$

Here

$$d_{t,s}^{(k)} \triangleq \sqrt{\frac{8\sigma_k^2 \log(4t^p)}{s}} + \frac{8\log(4t^p)}{s}.$$

Indeed, following the proof of Theorem 2, we derive that with probability $1-\beta$, we have for any fixed k,

$$\begin{array}{rcll} \left|\overline{X}_{k,t} - \mu_k\right| & < & \sqrt{\frac{2\sigma_k^2 \log(4\beta^{-1})}{t}} + \frac{2\log(4\beta^{-1})}{3t} & \text{and} \\ \left|V_{k,t} - \sigma_k^2\right| & < & \sqrt{\frac{8\sigma_k^2 \log(4\beta^{-1})}{t}} + \frac{4\log(4\beta^{-1})}{3t}. & \end{array} \tag{15}$$

¿From this we deduce, similarly to what is done in the proof of Theorem 2, that for any $s \geq 1$, with probability $1-\beta$, we have the two inequalities

$$\mu_k \leq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\log(4\beta^{-1})}{s}} + \frac{16\log(4\beta^{-1})}{3s} \leq \mu_k + \sqrt{\frac{8\sigma_k^2\log(4\beta^{-1})}{s}} + \frac{8\log(4\beta^{-1})}{s}$$

and (14) follows when choosing $\beta = t^{-p}$.

Now, pick a suboptimal arm, k (i.e., $\mu_k < \mu^*$). Then defining $u_{k,n} \geq 1$ an integer-valued sequence that will be selected later, we have:

$$\begin{array}{rcl} T_k(n) & = & \displaystyle\sum_{t=1}^n \mathbb{I}_{\{I_t=k\}} \\ & \leq & u_{k,n} - 1 + \displaystyle\sum_{t=1}^n \mathbb{I}_{\{I_t=k, T_k(t)\geq u_{k,n}\}} \\ & \leq & u_{k,n} - 1 + \displaystyle\sum_{t=1}^n \mathbb{I}_{\left\{\overline{X}^*_{T^*(t)} + c^*_{t,T^*(t)} \leq \overline{X}_{k,T_k(t)} + c^{(k)}_{t,T_k(t)}, T_k(t)\geq u_{k,n}\right\}} \\ & \leq & u_{k,n} - 1 + \displaystyle\sum_{t=1}^n \mathbb{I}_{\left\{\overline{X}^*_{T^*(t)} + c^*_{t,T^*(t)} \leq \mu^*\right\}} \\ & & \quad + \displaystyle\sum_{t=1}^n \mathbb{I}_{\left\{\overline{X}_{k,T_k(t)} + c^{(k)}_{t,T_k(t)} - d^{(k)}_{t,T_k(t)} \geq \mu_k\right\}} + \sum_{t=1}^n \mathbb{I}_{\left\{\mu^* < \mu_k + d^{(k)}_{t,T_k(t)}, T_k(t)\geq u_{k,n}\right\}} \qquad (16) \\ & \leq & u_{k,n} - 1 + \displaystyle\sum_{t=1}^n \sum_{s=1}^t \mathbb{I}_{\left\{\overline{X}^*_s + c^*_{t,s} \leq \mu^*\right\}} \\ & & \quad + \displaystyle\sum_{t=1}^n \sum_{s=1}^t \mathbb{I}_{\left\{\overline{X}_{k,s} + c^{(k)}_{t,s} - d^{(k)}_{t,s} \geq \mu_k\right\}} + \sum_{t=1}^n \sum_{s=u_{k,n}}^t \mathbb{I}_{\left\{d^{(k)}_{t,s} > \Delta_k\right\}} \qquad (17) \end{array}$$

where (16) follows as follows: Assume that $\overline{X}^*_{T^*(t)} + c^*_{t,T^*(t)} > \mu^*$ and $\overline{X}_{k,T_k(t)} + c^{(k)}_{t,T_k(t)} - d^{(k)}_{t,T_k(t)} < \mu_k$, and $\overline{X}^*_{T^*(t)} + c^*_{t,T^*(t)} \leq \overline{X}_{k,T_k(t)} + c^{(k)}_{t,T_k(t)}$. Then $\mu^* < \overline{X}^*_{T^*(t)} + c^*_{t,T^*(t)} \leq \overline{X}_{k,T_k(t)} + c^{(k)}_{t,T_k(t)} < \mu_k + d^{(k)}_{t,T_k(t)}$ and so under these conditions we must have $\mu^* < \mu_k + d^{(k)}_{t,T_k(t)}$.

The two first sums in the expression (17) are upper-bounded, in expectation, by

$$\sum_{t\geq 1}\sum_{s=1}^{t} \mathbb{P}(\overline{X}^*_s + c^*_{t,s} \leq \mu^*) + \mathbb{P}(\overline{X}_{k,s} + c^{(k)}_{t,s} - d^{(k)}_{t,s} \geq \mu_k)$$

which, from (14), is upper-bounded by $2\sum_{t\geq 1}\sum_{s=1}^{t} t^{-p} = 2\sum_{t\geq 1} t^{-p+1} \leq 2(1+1/(p-2))$ (bounding the sum by an integral) for $p > 2$.

Now, the last sum in (17) equals zero for some appropriate value of $u_{k,n}$. Indeed, thanks to the increasing monotonicity of $d^{(k)}_{\cdot,s}$ and the decreasing monotonicity of $d^{(k)}_{t,\cdot}$, the event

$$d^{(k)}_{t,s} > \Delta_k$$

for any $1 \leq u_{k,n} \leq s \leq t \leq n$ implies the event

$$d^{(k)}_{n,u_{k,n}} > \Delta_k.$$

But this last event never holds for a large enough value of $u_{k,n}$. Indeed, using the same argument as in the proof of Theorem 2, i.e. the fact that

$$\sqrt{\frac{8\sigma^2 l}{u}} + \frac{8l}{u} < \Delta$$

whenever $u/l > 8\sigma^2/\Delta^2 + 16/\Delta$, we deduce that for

$$\frac{u_{k,n}}{\log(4n^p)} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}$$

we have $d^{(k)}_{n,u_{k,n}} < \Delta_k$ thus $d^{(k)}_{t,s} < \Delta_k$ for all $1 \leq u_{k,n} \leq s \leq t \leq n$, and the third term of the sum in (17) is zero.

Thus, defining $u_{k,n} \triangleq 1 + \log(4n^p)\big(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}\big)$ and summing the three terms of (17) yields the logarithmic expected number of times a suboptimal arm is chosen:

$$\mathbb{E}[T_k(n)] \leq \log(4n^p)\left(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{16}{\Delta_k}\right) + 2 + \frac{2}{p-2},$$

and the logarithmic bound on the expected regret:

$$\overline{R}_n \leq \big[\log(4) + p\log(n)\big] \sum_{k\neq k^*}\left(\frac{8\sigma_k^2}{\Delta_k} + 16\right) + (2 + \frac{2}{p-2})\sum_{k\neq k^*}\Delta_k.$$

References

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

[2] L. Györfi L. Devroye and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[3] G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences.* PhD thesis, Université Paris 11, 2005.