Tuning bandit algorithms in stochastic environments

Jean-Yves Audibert, Rémi Munos, Csaba Szepesvari

Tuning bandit algorithms in stochastic environments

Jean-Yves Audibert[1] and Rémi Munos[2] and Csaba Szepesvári[3]

[1] CERTIS - Ecole des Ponts
19, rue Alfred Nobel - Cité Descartes
77455 Marne-la-Vallée - France
audibert@certis.enpc.fr
[2] INRIA Futurs Lille, SequeL project,
40 avenue Halley, 59650 Villeneuve d'Ascq, France
remi.munos@inria.fr
[3] University of Alberta, Edmonton T6G 2E8, Canada
szepesva@cs.ualberta.ca

Abstract. Algorithms based on upper-confidence bounds for balancing exploration and exploitation are gaining popularity since they are easy to implement, efficient and effective. In this paper we consider a variant of the basic algorithm for the stochastic, multi-armed bandit problem that takes into account the empirical variance of the different arms. In earlier experimental works, such algorithms were found to outperform the competing algorithms. The purpose of this paper is to provide a theoretical explanation of these findings and provide theoretical guidelines for the tuning of the parameters of these algorithms. For this we analyze the expected regret and for the first time the concentration of the regret. The analysis of the expected regret shows that variance estimates can be especially advantageous when the payoffs of suboptimal arms have low variance. The risk analysis, rather unexpectedly, reveals that except for some very special bandit problems, the regret, for upper confidence bounds based algorithms with standard bias sequences, concentrates only at a polynomial rate. Hence, although these algorithms achieve logarithmic expected regret rates, they seem less attractive when the risk of suffering much worse than logarithmic regret is also taken into account.

1 Introduction and notations

In this paper we consider *stochastic multi-armed bandit problems*. The original motivation of bandit problems comes from the desire to optimize efficiency in clinical trials when the decision maker can choose between treatments but initially does not know which of the treatments is the most effective one [9]. Multi-armed bandit problems became popular with the seminal paper of Robbins [8], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is defined by K distributions, $\nu_1, \ldots, \nu_K$, one for each "arm" of the bandit. Imagine a gambler playing with

these K slot machines. The gambler can pull the arm of any of the machines. Successive plays of arm k yield a sequence of independent and identically distributed (i.i.d.) real-valued random variables $X_{k,1}, X_{k,2}, \ldots$, coming from the distribution ν_k. The random variable $X_{k,t}$ is the payoff (or reward) of the k-th arm when this arm is pulled the t-th time. Independence also holds for rewards across the different arms. The gambler facing the bandit problem wants to pull the arms so as to maximize his cumulative payoff.

The problem is made challenging by assuming that the payoff distributions are initially unknown. Thus the gambler must use exploratory actions in order to learn the utility of the individual arms, making his decisions based on the available past information. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, efficient online algorithms must find the right balance between *exploration and exploitation.*

Since the gambler cannot use the distributions of the arms (which are not available to him) he must follow a *policy*, which is a mapping from the space of possible histories, $\cup_{t\in\mathbb{N}^+}\{1,\ldots,K\}^t \times \mathbb{R}^t$, into the set $\{1,\ldots,K\}$, which indexes the arms. Let $\mu_k = \mathbb{E}[X_{k,1}]$ denote the expected reward of arm k.[4] By definition, an *optimal arm* is an arm having the largest expected reward. We will use k^* to denote the index of such an arm. Let the optimal expected reward be $\mu^* = \max_{1\le k\le K} \mu_k$.

Further, let $T_k(t)$ denote the number of times arm k is chosen by the policy during the first t plays and let I_t denote the arm played at time t. The *(cumulative) regret at time* n is defined by

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

Oftentimes, the goal is to minimize the *expected (cumulative) regret of the policy*, $\mathbb{E}[\hat{R}_n]$. Clearly, this is equivalent to maximizing the total expected reward achieved up to time n. It turns out that the expected regret satisfies

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

where $\Delta_k \triangleq \mu^* - \mu_k$ is the expected loss of playing arm k. Hence, an algorithm that aims at minimizing the expected regret should minimize the expected sampling times of suboptimal arms.

Early papers studied stochastic bandit problems under Bayesian assumptions (e.g., [5]). Lai and Robbins [6] studied bandit problems with parametric uncertainties. They introduced an algorithm that follows what is now called the "optimism in the face of uncertainty". Their algorithm computes *upper confidence bounds* for all the arms by maximizing the expected payoff when the parameters are varied within appropriate confidence sets derived for the parameters. Then the algorithm chooses the arm with the highest such bound. They

[4] $\mathbb{N}$ denotes the set of natural numbers, including zero and $\mathbb{N}^+$ denotes the set of strictly positive integers.

show that the expected regret increases logarithmically only with the number of trials and prove that the regret is asymptotically the smallest possible up to a sublogarithmic factor for the considered family of distributions. Agrawal [1] has shown how to construct such optimal policies starting from the sample-mean of the arms. More recently, Auer et al. [3] considered the case when the rewards come from a bounded support, say $[0, b]$, but otherwise the reward distributions are unconstrained. They have studied several policies, most notably UCB1 which constructs the Upper Confidence Bounds (UCB) for arm k at time t by adding the *bias factor*

$$\sqrt{\frac{2b^2 \log t}{T_k(t-1)}}$$

to its sample-mean. They have proven that the expected regret of this algorithm satisfies

$$\mathbb{E}[\hat{R}_n] \leq 8 \left(\sum_{k:\mu_k<\mu^*} \frac{b^2}{\Delta_k} \right) \log(n) \; + O(1). \tag{1}$$

In the same paper they propose UCB1-NORMAL, that is designed to work with normally distributed rewards only. This algorithm estimates the variance of the arms and uses these estimates to refine the bias factor. They show that for this algorithm when the rewards are indeed normally distributed with means μ_k and variances σ_k^2,

$$\mathbb{E}[\hat{R}_n] \leq 8 \sum_{k:\mu_k<\mu^*} \left(\frac{32\sigma_k^2}{\Delta_k} + \Delta_k \right) \log(n) + O(1). \tag{2}$$

Note that one major difference of this result and the previous one is that the regret-bound for UCB1 scales with b^2, while the regret bound for UCB1-NORMAL scales with the variances of the arms. First, let us note that it can be proven that the scaling behavior of the regret-bound with b is not a proof artifact: The expected regret indeed scales with $\Omega(b^2)$. Since b is typically just an *a priori* guess on the size of the interval containing the rewards, which might be overly conservative, it is desirable to lessen the dependence on it.

Auer et al. introduced another algorithm, UCB1-Tuned, in the experimental section of their paper. This algorithm, similarly to UCB1-NORMAL uses the empirical estimates of the variance in the bias sequence. Although no theoretical guarantees were derived for UCB1-Tuned, this algorithm has been shown to outperform the other algorithms considered in the paper in essentially all the experiments. The superiority of this algorithm has been reconfirmed recently in the latest Pascal Challenge [4]. Intuitively, algorithms using variance estimates should work better than UCB1 when the variance of some suboptimal arms is much smaller than b^2, since these arms will be less often drawn: suboptimal arms are more easily spotted by algorithms using variance estimates.

In this paper we study the regret of *UCB-V*, which is a generic UCB algorithm that use variance estimates in the bias sequence. In particular, the bias sequences of UCB-V take the form

$$\sqrt{\frac{2V_{k,T_k(t-1)}\mathcal{E}_{T_k(t-1),t}}{T_k(t-1)}} + c\frac{3b\mathcal{E}_{T_k(t-1),t}}{T_k(t-1)},$$

where $V_{k,s}$ is the empirical variance estimate for arm k based on s samples, $\mathcal{E}$ (viewed as a function of (s,t)) is a so-called *exploration function* for which a typical choice is $\mathcal{E}_{s,t} = \zeta \log(t)$ (thus in this case, $\mathcal{E}$ independent of s). Here $\zeta, c > 0$ are tuning parameters that can be used to control the behavior of the algorithm.

One major result of the paper (Corollary 1) is a bound on the expected regret that scales in an improved fashion with b. In particular, we show that for a particular settings of the parameters of the algorithm,

$$\mathbb{E}[\hat{R}_n] \leq 10 \sum_{k:\mu_k<\mu^*} \left(\frac{\sigma_k^2}{\Delta_k} + 2b \right) \log(n).$$

The main difference to the bound (1) is that b^2 is replaced by σ_k^2, though b still appears in the bound. This is indeed the major difference to the bound (2).[5] In order to prove this result we will prove a novel tail bound on the sample average of i.i.d. random variables with bounded support that, unlike previous similar bounds, involves the empirical variance and which may be of independent interest (Theorem 1). Otherwise, the proof of the regret bound involves the analysis of the sampling times of suboptimal arms (Theorem 2), which contains significant advances compared with the one in [3]. This way we obtain results on the expected regret for a wide class of exploration functions (Theorem 3). For the "standard" logarithmic sequence we will give lower limits on the tuning parameters: If the tuning parameters are below these limits the loss goes up considerably (Theorems 4,5).

The second major contribution of the paper is the analysis of the risk that the studied upper confidence based policies have a regret much higher than its expected value. To our best knowledge no such analysis existed for this class of algorithms so far. In order to analyze this risk, we define the *(cumulative) pseudo-regret* at time n via

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

Note that the expectation of the pseudo-regret and the regret are the same: $\mathbb{E}[R_n] = \mathbb{E}[\hat{R}_n]$. The difference of the regret and the pseudo-regret comes from the randomness of the rewards. Sections 4 and 5 develop high probability bounds for the pseudo-regret . The same kind of formulae can be obtained for the cumulative regret (see Remark 2 p.13).

Interestingly, our analysis revealed some tradeoff that we did not expect: As it turns out, if one aims for logarithmic expected regret (or, more generally, for subpolynomial regret) then the regret does not necessarily concentrate exponentially fast around its mean (Theorem 7). In fact, this is the case when with positive probability the optimal arm yields a reward smaller than the expected

[5] Although this is unfortunate, it is possible to show that the dependence on b is unavoidable.

reward of some suboptimal arm. Take for example two arms satisfying this condition and with $\mu_1 > \mu_2$: the first arm is the optimal arm and $\Delta_2 = \mu_1 - \mu_2 > 0$. Then the distribution of the pseudo-regret at time n will have two modes, one at $\Omega(\log n)$ and the other at $\Omega(\Delta_2 n)$. The probability mass associated with this second mass will decay polynomially with n where the rate of decay depends on Δ_2. Above the second mode the distribution decays exponentially. By increasing the exploration rate the situation can be improved. Our risk tail bound (Theorem 6) makes this dependence explicit. Of course, increasing exploration rate increases the expected regret. This illustrates the tradeoff between the expected regret and the risk of achieving much worse than the expected regret. One lesson is thus that if in an application risk is important then it might be better to increase the exploration rate.

In Section 5, we study a variant of the algorithm obtained by $\mathcal{E}_{s,t} = \mathcal{E}_s$. In particular, we show that with an appropriate choice of $\mathcal{E}_s = \mathcal{E}_s(\beta)$, for any $0 < \beta < 1$, for an infinite number of plays, the algorithm achieves *finite* cumulative regret with probability $1 - \beta$ (Theorem 8). Hence, we name this variant PAC-UCB ("Probably approximately correct UCB"). Besides, for a finite time-horizon n, choosing $\beta = 1/n$ then yields a logarithmic bound on the regret that fails with probability $O(1/n)$ only. This should be compared with the bound $O(1/\log(n)^a)$, $a > 0$ obtained for the standard choice $\mathcal{E}_{s,t} = \zeta \log t$ in Corollary 2. Hence, we conjecture that knowing the time horizon might represent a significant advantage.

Due to limited space, some of the proofs are absent from this paper. All the proofs can be found in the extended version [2].

2 The UCB-V algorithm

For any $k \in \{1, \dots, K\}$ and $t \in \mathbb{N}$, let $\overline{X}_{k,t}$ and $V_{k,t}$ be the empirical estimates of the mean payoff and variance of arm k:

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

where by convention $\overline{X}_{k,0} \triangleq 0$ and $V_{k,0} \triangleq 0$. We recall that an *optimal arm* is an arm having the best expected reward $k^* \in \operatorname{argmax}_{k \in \{1,\dots,K\}} \mu_k$. We denote quantities related to the optimal arm by putting $*$ in the upper index.

In the following, we assume that the rewards are bounded. Without loss of generality, we may assume that all the rewards are almost surely in $[0, b]$, with $b > 0$. We summarize our assumptions on the reward sequence here:

Assumptions: Let $K > 2$, $\nu_1, \dots, \nu_K$ distributions over reals with support $[0, b]$. For $1 \le k \le K$, let $\{X_{k,t}\} \sim \nu_k$ be an i.i.d. sequence of random variables specifying the rewards for arm k.[6] Assume that the rewards of different arms are independent of each other, i.e., for any k, k', $1 \le k < k' \le K$, $t \in \mathbb{N}^+$,

[6] The i.i.d. assumption can be relaxed, see e.g., [7].

the collection of random variables, $(X_{k,1}, \dots, X_{k,t})$ and $(X_{k',1}, \dots, X_{k',t})$, are independent of each other.

2.1 The algorithm

Let $c \geq 0$. Let $\mathcal{E} = (\mathcal{E}_{s,t})_{s \geq 0, t \geq 0}$ be nonnegative real numbers such that for any $s \geq 0$, the function $t \mapsto \mathcal{E}_{s,t}$ is nondecreasing. We shall call $\mathcal{E}$ (viewed as a function of (s,t)) the exploration function. For any arm k and any nonnegative integers s, t, introduce

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\mathcal{E}_{s,t}}{s}} + c\frac{3b\mathcal{E}_{s,t}}{s} \tag{3}$$

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

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

Let us roughly describe the behavior of the algorithm. At the beginning (i.e., for small t), every arm that has not been drawn is associated with an infinite bound which will become finite as soon as the arm is drawn. The more an arm k is drawn, the closer the bound (3) gets close to its first term, and thus, from the law of large numbers, to the expected reward μ_k. So the procedure will hopefully tend to draw more often arms having greatest expected rewards.

Nevertheless, since the obtained rewards are stochastic it might happen that during the first draws the (unknown) optimal arm always gives low rewards. Fortunately, if the optimal arm has not been drawn too often (i.e., small $T_{k^*}(t-1)$), for appropriate choices of $\mathcal{E}$ (when $\mathcal{E}_{s,t}$ increases without bounds in t for any fixed s), after a while the last term of (3) will start to dominate the two other terms and will also dominate the bound associated with the arms drawn very often. Thus the optimal arm will be drawn even if the empirical mean of the obtained rewards, $\overline{X}_{k^*,T_{k^*}(t-1)}$, is small. More generally, such choices of $\mathcal{E}$ lead to the exploration of arms with inferior empirical mean. This is why $\mathcal{E}$ is referred to as the exploration function. Naturally, a high-valued exploration function also leads to draw often suboptimal arms. Therefore the choice of $\mathcal{E}$ is crucial in order to explore possibly suboptimal arms while keeping exploiting (what looks like to be) the optimal arm.

The actual form of $B_{k,s,t}$ comes from the following novel tail bound on the sample average of i.i.d. random variables with bounded support that, unlike previous similar bounds (Bennett's and Bernstein's inequalities), involves the empirical variance.

Theorem 1. *Let $X_1, \dots, X_t$ be i.i.d. random variables taking their values in $[0,b]$. Let $\mu = \mathbb{E}[X_1]$ be their common expected value. Consider the empirical expectation $\overline{X}_t$ and variance V_t defined respectively by*

$$\overline{X}_t = \frac{\sum_{i=1}^t X_i}{t} \quad \text{and} \quad V_t = \frac{\sum_{i=1}^t (X_i - \overline{X}_t)^2}{t}.$$

Then for any $t \in \mathbb{N}$ and $x > 0$, with probability at least $1 - 3e^{-x}$,

$$|\overline{X}_t - \mu| \leq \sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}. \tag{4}$$

Furthermore, introducing

$$\beta(x,t) = 3 \inf_{1<\alpha\leq 3} \Big(\frac{\log t}{\log \alpha} \wedge t \Big) e^{-x/\alpha}, \tag{5}$$

we have for any $t \in \mathbb{N}$ and $x > 0$, with probability at least $1 - \beta(x,t)$

$$|\overline{X}_s - \mu| \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s} \tag{6}$$

hold simultaneously for $s \in \{1, 2, \ldots, t\}$.

Remark 1. The uniformity in time is the only difference between the two assertions of the previous theorem. When we use (6), the values of x and t will be such that $\beta(x,t)$ is of order of $3e^{-x}$, hence there will be no real price to pay for writing a version of (4) that is uniform in time. In particular, this means that if $1 \leq S \leq t$ is a random variable then (6) still holds with probability at least $1 - \beta(x,t)$ and when s is replaced with S.

Note that (4) is useless for $t \leq 3$ since its r.h.s. is larger than b. For any arm k, time t and integer $1 \leq s \leq t$ we may apply Theorem 1 to the rewards $X_{k,1}, \ldots, X_{k,s}$, and obtain that with probability at least $1 - 3\sum_{s=4}^{\infty} e^{-(c\wedge 1)\mathcal{E}_{s,t}}$, we have $\mu_k \leq B_{k,s,t}$. Hence, by our previous remark at time t with high probability (for a high-valued exploration function $\mathcal{E}$) the expected reward of arm k is upper bounded by $B_{k,T_k(t-1),t}$. The user of the generic UCB-V policy has two parameters to tune: the exploration function $\mathcal{E}$ and the positive real number c.

A cumbersome technical analysis (not reproduced here) shows that there are essentially two interesting types of exploration functions:

- the ones in which $\mathcal{E}_{s,t}$ depends only on t (see Sections 3 and 4).
- the ones in which $\mathcal{E}_{s,t}$ depends only on s (see Section 5).

2.2 Bounds for the sampling times of suboptimal arms

The natural way of bounding the regret of UCB policies is to bound the number of times suboptimal arms are drawn (or the inferior sampling times). The bounds presented here significantly improve the ones used in [3]. This improvement is necessary to get tight bounds for the interesting case where the exploration function is logarithmic. The idea of the bounds is that the inferior sampling time of an arm can be bounded in terms of the number of times the UCB for the arm considered is over a some threshold value (τ in the statement below) and the number of times the UCB for an optimal arm is below the same threshold. Note that even though the above statements hold for any arm, they will be only useful for suboptimal arms. In particular, for a suboptimal arm the threshold can be chosen to lie between the payoff of an optimal arm and the payoff of the arm considered.

Theorem 2. *Consider UCB-V. Then, after K plays, each arm has been pulled once. Further, the following holds: Let arm k and time $n \in \mathbb{N}^+$ be fixed. For any $\tau \in \mathbb{R}$ and any integer $u > 1$, we have*

$$T_k(n) \leq u + \sum_{t=u+K-1}^{n} \Big(\mathbb{1}_{\{\exists s : u \leq s \leq t-1 \text{ s.t. } B_{k,s,t} > \tau\}} + \mathbb{1}_{\{\exists s^* : 1 \leq s^* \leq t-1 \text{ s.t. } \tau \geq B_{k^*,s^*,t}\}} \Big), \tag{7}$$

hence

$$\begin{aligned} \mathbb{E}\left[T_k(n)\right] \leq u &+ \sum_{t=u+K-1}^{n} \sum_{s=u}^{t-1} \mathbb{P}\big(B_{k,s,t} > \tau\big) \\ &+ \sum_{t=u+K-1}^{n} \mathbb{P}\big(\exists s : 1 \leq s \leq t-1 \text{ s.t. } B_{k^*,s,t} \leq \tau\big). \end{aligned} \tag{8}$$

Besides we have

$$\begin{aligned} &\mathbb{P}\big(T_k(n) > u\big) \\ &\quad \leq \sum_{t=3}^{n} \mathbb{P}\big(B_{k,u,t} > \tau\big) + \mathbb{P}\big(\exists s : 1 \leq s \leq n-u \text{ s.t. } B_{k^*,s,u+s} \leq \tau\big). \end{aligned} \tag{9}$$

Proof. The first assertion is trivial since at the beginning all arms has an infinite UCB, which becomes finite as soon as the arm has been played once. To obtain (7), we note that

$$T_k(n) - u \leq \sum_{t=u+K-1}^{n} \mathbb{1}_{\{I_t = k; T_k(t) > u\}} = \sum_{t=u+K-1}^{n} Z_{k,t,u},$$

where

$$\begin{aligned} Z_{k,t,u} &= \mathbb{1}_{\{I_t = k;\, u \leq T_k(t-1);\, 1 \leq T_{k^*}(t-1); B_{k,T_k(t-1),t} \geq B_{k^*,T_{k^*}(t-1),t}\}} \\ &\leq \mathbb{1}_{\{\exists s : u \leq s \leq t-1 \text{ s.t. } B_{k,s,t} > \tau\}} + \mathbb{1}_{\{\exists s^* : 1 \leq s^* \leq t-1 \text{ s.t. } \tau \geq B_{k^*,s^*,t}\}} \end{aligned}$$

Taking the expectation on both sides of (7) and using the probability union bound, we obtain (8). Finally, (9) comes from a more direct argument that uses the fact that the exploration function $\xi_{s,t}$ is a nondecreasing function with respect to t. Consider an event such that the following statements hold:

$$\begin{cases} \forall t : 3 \leq t \leq n \text{ s.t. } B_{k,u,t} \leq \tau, \\ \forall s : 1 \leq s \leq n-u \text{ s.t. } B_{k^*,s,u+s} > \tau. \end{cases}.$$

Then for any $1 \leq s \leq n-u$ and $u+s \leq t \leq n$,

$$B_{k^*,s,t} \geq B_{k^*,s,u+s} > \tau \geq B_{k,u,t}.$$

This implies that arm k will not be pulled a $(u+1)$-th time. Therefore we have proved by contradiction that

$$\begin{aligned} \big\{T_k(n) > u\big\} \subset \Big(&\big\{\exists t : 3 \leq t \leq n \text{ s.t. } B_{k,u,t} > \tau\big\} \\ &\cup \big\{\exists s : 1 \leq s \leq n-u \text{ s.t. } B_{k^*,s,u+s} \leq \tau\big\} \Big), \end{aligned} \tag{10}$$

which by taking probabilities of both sides gives the announced result.

3 Expected regret of UCB-V

In this section, we consider that the exploration function does not depend on s (yet, $\mathcal{E} = (\mathcal{E}_t)_{t\geq 0}$ is still nondecreasing with t). We will see that as far as the expected regret is concerned, a natural choice of $\mathcal{E}_t$ is the logarithmic function and that c should not be taken too small if one does not want to suffer polynomial regret instead of logarithmic one. We derive bounds on the expected regret and conclude by specifying natural constraints on c and $\mathcal{E}_t$.

Theorem 3. *We have*

$$\mathbb{P}(B_{k,s,t} > \mu^*) \leq 2e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}, \tag{11}$$

and

$$\begin{aligned}\mathbb{E}[R_n] \leq \sum_{k:\Delta_k>0} \Bigg\{ & 1 + 8(c\vee 1)\mathcal{E}_n\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right) \\ & + ne^{-\mathcal{E}_n}\left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) + \sum_{t=16\mathcal{E}_n}^{n} \beta\big((c\wedge 1)\mathcal{E}_t, t\big)\Bigg\}\Delta_k,\end{aligned} \tag{12}$$

where we recall that $\beta\big((c\wedge 1)\mathcal{E}_t, t\big)$ is essentially of order $e^{-(c\wedge 1)\mathcal{E}_t}$ (see (5) and Remark 1).

Proof. Let $\mathcal{E}'_n = (c\vee 1)\mathcal{E}_n$. We use (8) with u being the smallest integer larger than $8\big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\big)\mathcal{E}'_n$ and $\tau = \mu^*$. For any $s \geq u$ and $t \geq 2$, we have

$$\begin{aligned}&\mathbb{P}(B_{k,s,t} > \mu^*) = \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2V_{k,s}\mathcal{E}_t}{s}} + 3bc\frac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2[\sigma_k^2+b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\tfrac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\big) + \mathbb{P}\big(V_{k,s} \geq \sigma_k^2 + b\Delta_k/2\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} - \mu_k > \Delta_k/2\big) + \mathbb{P}\Big(\tfrac{\sum_{j=1}^s (X_{k,j}-\mu_k)^2}{s} - \sigma_k^2 \geq b\Delta_k/2\Big) \\ &\leq 2e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)},\end{aligned} \tag{13}$$

proving (11). Here in the last step we used Bernstein's inequality twice and in the second inequality we used that the choice of u guarantees that for any $u \leq s < t$ and $t \geq 2$,

$$\begin{aligned}\sqrt{\tfrac{2[\sigma_k^2+b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\tfrac{\mathcal{E}_t}{s} &\leq \sqrt{\tfrac{[2\sigma_k^2+b\Delta_k]\mathcal{E}'_n}{u}} + 3b\tfrac{\mathcal{E}'_n}{u} \leq \sqrt{\tfrac{[2\sigma_k^2+b\Delta_k]\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]}} + \tfrac{3b\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]} \\ &= \tfrac{\Delta_k}{2}\left[\sqrt{\tfrac{2\sigma_k^2+b\Delta_k}{2\sigma_k^2+4b\Delta_k}} + \tfrac{3b\Delta_k}{4\sigma_k^2+8b\Delta_k}\right] \leq \tfrac{\Delta_k}{2},\end{aligned} \tag{14}$$

since the last inequality is equivalent to $(x-1)^2 \geq 0$ with $x = \sqrt{\frac{2\sigma_k^2+b\Delta_k}{2\sigma_k^2+4b\Delta_k}}$. Summing up the probabilities in Equation (13) we obtain

$$\sum_{s=u}^{t-1} \mathbb{P}(B_{k,s,t} > \mu^*) \leq 2\sum_{s=u}^{\infty} e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} = 2\frac{e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}{1-e^{-\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}$$
$$\leq \left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} \leq \left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) e^{-\mathcal{E}_n'}, \qquad (15)$$

where we have used that $1 - e^{-x} \geq 2x/3$ for $0 \leq x \leq 3/4$. By using (6) of Theorem 1 to bound the other probability in (8), we obtain that

$$\mathbb{E}\left[T_k(n)\right] \leq 1 + 8\mathcal{E}_n' \left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right) + ne^{-\mathcal{E}_n'} \left(\frac{24\sigma_k^2}{\Delta_k^2} + \frac{4b}{\Delta_k}\right) + \sum_{t=u+1}^{n} \beta((c \wedge 1)\mathcal{E}_t, t),$$

which by $u \geq 16\mathcal{E}_n$ gives the announced result.

In order to balance the terms in (12) the exploration function should be chosen to be proportional to $\log t$. For this choice, the following corollary gives an explicit bound on the expected regret:

Corollary 1. *If $c = 1$ and $\mathcal{E}_t = \zeta \log t$ for $\zeta > 1$, then there exists a constant c_ζ depending only on ζ such that for $n \geq 2$*

$$\mathbb{E}[R_n] \leq c_\zeta \sum_{k:\Delta_k>0} \left(\frac{\sigma_k^2}{\Delta_k} + 2b\right) \log n. \qquad (16)$$

For instance, for $\zeta = 1.2$, the result holds for $c_\zeta = 10$.

Proof (Sketch of the proof). The first part, (16), follows directly from Theorem 3. Let us thus turn to the numerical result. For $n \geq K$, we have $R_n \leq b(n-1)$ (since in the first K rounds, the optimal arm is chosen at least once). As a consequence, the numerical bound is nontrivial only for $20 \log n < n - 1$, so we only need to check the result for $n > 91$. For $n > 91$, we bound the constant term using $1 \leq \frac{\log n}{\log 91} \leq a_1 \frac{2b}{\Delta_k}(\log n)$, with $a_1 = 1/(2\log 91) \approx 0.11$. The second term between the brackets in (12) is bounded by $a_2\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right)\log n$, with $a_2 = 8 \times 1.2 = 9.6$. For the third term, we use that for $n > 91$, we have $24n^{-0.2} < a_3 \log n$, with $a_3 = \frac{24}{91^{0.2} \times \log 91} \approx 0.21$. By tedious computations, the fourth term can be bounded by $a_4 \frac{2b}{\Delta_k}(\log n)$, with $a_4 \approx 0.07$. This gives the desired result since $a_1 + a_2 + a_3 + a_4 \leq 10$.

As promised, Corollary 1 gives a logarithmic bound on the expected regret that has a linear dependence on the range of the reward, contrary to bounds for algorithms that do not take into account the empirical variance of the rewards (see e.g. the bound (1) that holds for UCB1).

The previous corollary is well completed by the following result, which essentially says that we should not use $\mathcal{E}_t = \zeta \log t$ with $\zeta < 1$.

Theorem 4. *Consider $\mathcal{E}_t = \zeta \log t$ and let n denote the total number of draws. Whatever c is, if $\zeta < 1$, then there exist some reward distributions (depending on n) such that*

- *the expected number of draws of suboptimal arms using the UCB-V algorithm is polynomial in the total number of draws*
- *the UCB-V algorithm suffers a polynomial loss.*

So far we have seen that for $c = 1$ and $\zeta > 1$ we obtain a logarithmic regret, and that the constant ζ should not be taken below 1 (whatever c is) without risking to suffer polynomial regret. Now we consider the last term in $B_{k,s,t}$, which is linear in the ratio $\mathcal{E}_t/s$, and show that this term is also necessary to obtain a logarithmic regret, since we have:

Theorem 5. *Consider $\mathcal{E}_t = \zeta \log t$. Whatever ζ is, if $c\zeta < 1/6$, there exist probability distributions of the rewards such that the UCB-V algorithm suffers a polynomial loss.*

To conclude the above analysis, natural values for the constants appearing in the bound are the following ones

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\log t}{s}} + \frac{b\log t}{2s}.$$

This choice corresponds to the critical exploration function $\mathcal{E}_t = \log t$ and to $c = 1/6$, that is, the minimal associated value of c in view of the previous theorem. In practice, it would be unwise (or risk seeking) to use smaller constants in front of the last two terms.

4 Concentration of the regret

In real life, people are not only interested in the expected rewards that they can obtain by some policy. They also want to estimate probabilities of obtaining much less rewards than expected, hence they are interested in the concentration of the regret. This section starts with the study of the concentration of the pseudo-regret, since, as we will see in Remark 2 p.13, the concentration properties of the regret follow from the concentration properties of the pseudo-regret.

We still assume that the exploration function does not depend on s and that $\mathcal{E} = (\mathcal{E}_t)_{t\geq 0}$ is nondecreasing. Introduce

$$\tilde{\beta}_n(t) \triangleq 3 \min_{\substack{\alpha\geq 1 \quad M\in\mathbb{N} \\ s_0=0<s_1<\cdots<s_M=n \\ \text{s.t. } s_{j+1}\leq\alpha(s_j+1)}} \sum_{j=0}^{M-1} e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+t+1}}{\alpha}}. \tag{17}$$

We have seen in the previous section that in order to obtain logarithmic expected regret, it is natural to take a logarithmic exploration function. In this case, and also when the exploration function goes to infinity faster than the

logarithmic function, the complicated sum in (17), up to second order logarithmic terms, is of the order $e^{-(c\wedge 1)\mathcal{E}_t}$. This can be seen by considering (disregarding rounding issues) the geometric grid $s_j = \alpha^j$ with α close to 1. The next theorem provides a bound for the tails of the pseudo-regret.

Theorem 6. *Let*

$$v_k \triangleq 8(c \vee 1)\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{4b}{3\Delta_k}\right), \qquad r_0 \triangleq \sum_{k:\Delta_k>0} \Delta_k\big(1 + v_k\mathcal{E}_n\big).$$

Then, for any $x \geq 1$, we have

$$\mathbb{P}\big(R_n > r_0 x\big) \leq \sum_{k:\Delta_k>0} \left\{2ne^{-(c\vee 1)\mathcal{E}_n x} + \tilde{\beta}_n(\lfloor v_k\mathcal{E}_n x\rfloor)\right\}, \tag{18}$$

where we recall that $\tilde{\beta}_n(t)$ is essentially of order $e^{-(c\wedge 1)\mathcal{E}_t}$ (see the above discussion).[7]

Proof (sketch of the proof). First note that

$$\begin{aligned}\mathbb{P}\big(R_n > r_0 x\big) &= \mathbb{P}\Big\{\sum_{k:\Delta_k>0} \Delta_k T_k(n) > \sum_{k:\Delta_k>0} \Delta_k(1 + v_k\mathcal{E}_n)x\Big\} \\ &\leq \sum_{k:\Delta_k>0} \mathbb{P}\Big\{T_k(n) > (1 + v_k\mathcal{E}_n)x\Big\}.\end{aligned}$$

Let $\mathcal{E}'_n = (c\vee 1)\mathcal{E}_n$. We use (9) with $\tau = \mu^*$ and $u = \lfloor(1 + v_k\mathcal{E}_n)x\rfloor \geq v_k\mathcal{E}_n x$. From (11) of Theorem 3 , we have $\mathbb{P}(B_{k,u,t} > \mu^*) \leq 2e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} \leq 2e^{-\mathcal{E}'_n x}$. To bound the other probability in (9), we use $\alpha \geq 1$ and the grid $s_0, \ldots, s_M$ realizing the minimum of (17) when $t = u$. Let $I_j = \{s_j + 1, \ldots, s_{j+1}\}$. Then

$$\begin{aligned}\mathbb{P}\big(\exists s : 1 \leq s \leq n - u \text{ s.t. } B_{k^*,s,u+s} \leq \mu^*\big) &\leq \sum_{j=0}^{M-1} \mathbb{P}\big(\exists s \in I_j \text{ s.t. } B_{k^*,s,s_j+u+1} \leq \mu^*\big) \\ &\leq \sum_{j=0}^{M-1} \mathbb{P}\big(\exists s \in I_j \text{ s.t. } s(\overline{X}_{k^*,s} - \mu^*) + \sqrt{2sV_{k^*,s}\mathcal{E}_{s_j+u+1}} + 3bc\mathcal{E}_{s_j+u+1} \leq 0\big) \\ &\leq 3\sum_{j=0}^{M-1} e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+u+1}}{\alpha}} = \tilde{\beta}_n(u) \leq \tilde{\beta}_n(\lfloor v_k\mathcal{E}_n x\rfloor),\end{aligned}$$

where the second to last inequality comes from an appropriate union bound argument (see [2] for details).

When $\mathcal{E}_n \geq \log n$, the last term is the leading term. In particular, when $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$, Theorem 6 leads to the following corollary, which essentially says that for any $z > \gamma \log n$ with γ large enough,

$$\mathbb{P}\big(R_n > z\big) \leq C\, z^{-\zeta},$$

for some constant $C > 0$:

[7] Here $\lfloor x\rfloor$ denotes the largest integer smaller or equal to x.

Corollary 2. *When $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$, there exist $\kappa_1 > 0$ and $\kappa_2 > 0$ depending only on b, K, $(\sigma_k)_{k\in\{1,\dots,K\}}$, $(\Delta_k)_{k\in\{1,\dots,K\}}$ satisfying that for any $\varepsilon > 0$ there exists $\Gamma_\varepsilon > 0$ (tending to infinity when ε goes to 0) such that for any $n \geq 2$ and any $z > \kappa_1 \log n$*

$$\mathbb{P}\big(R_n > z\big) \leq \kappa_2 \frac{\Gamma_\varepsilon \log z}{z^{\zeta(1-\varepsilon)}}$$

Since the regret is expected to be of order $\log n$ the condition $z = \Omega(\log n)$ is not an essential restriction. Further, the regret concentration, although increasing with ζ, is pretty slow. For comparison, remember that a zero-mean martingale M_n with increments bounded by 1 would satisfy $\mathbb{P}(M_n > z) \leq \exp(-2z^2/n)$. The slow concentration for UCB-V happens because the first $\Omega(\log(t))$ choices of the optimal arm can be unlucky, in which case the optimal arm will not be selected any more during the first t steps. Hence, the distribution of the regret will be of a mixture form with a mode whose position scales linearly with time and which decays only at a polynomial rate, which is controlled by ζ.[8] This reasoning relies crucially on that the choices of the optimal arm can be unlucky. Hence, we have the following result:

Theorem 7. *Consider $\mathcal{E}_t = \zeta \log t$ with $c\zeta > 1$. Let $\tilde{k}$ denote the second optimal arm. If the essential infimum of the optimal arm is strictly larger than $\mu_{\tilde{k}}$, then the pseudo-regret has exponentially small tails. Inversely, if the essential infimum of the optimal arm is strictly smaller than μ_k, then the pseudo-regret has only polynomial tail.*

Remark 2. In Theorem 6 and Corollary 2, we have considered the pseudo-regret: $R_n = \sum_{k=1}^K T_k(n)\Delta_k$ instead of the regret $\hat{R}_n \triangleq \sum_{t=1}^n X_{k^*,t} - \sum_{t=1}^n X_{I_t,T_{I_t}(t)}$. Our main motivation for this was to provide as simple as possible formulae and assumptions. The following computations explains that when the optimal arm is unique, one can obtain similar contration bounds for the regret. Consider the interesting case when $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$. By modifying the analysis slightly in Corollary 2, one can get that there exists $\kappa_1 > 0$ such that for any $z > \kappa_1 \log n$, with probability at least $1 - z^{-1}$, the number of draws of suboptimal arms is bounded by $C\,z$ for some $C > 0$. This means that the algorithm draws an optimal arm at least $n - C\,z$ times. Now if the optimal arm is unique, this means that $n - Cz$ terms cancel out in the summations of the definition of the regret. For the Cz terms which remain, one can use standard Bernstein inequalities and union bounds to prove that with probability $1 - Cz^{-1}$, we have $\hat{R}_n \leq R_n + C'\sqrt{z}$. Since the bound on the pseudo-regret is of order z (Corollary 2), a similar bound holds for the regret.

5 PAC-UCB

In this section, we consider that the exploration function does not depend on t: $\mathcal{E}_{s,t} = \mathcal{E}_s$. We show that for an appropriate sequence $(\mathcal{E}_s)_{s\geq 0}$, this leads to

[8] Note that entirely analogous results hold for UCB1.

an UCB algorithm which has nice properties with high probability (Probably Approximately Correct), hence the name of it. Note that in this setting, the quantity $B_{k,s,t}$ does not depend on the time t so we will simply write it $B_{k,s}$. Besides, in order to simplify the discussion, we take $c = 1$.

Theorem 8. *Let $\beta \in (0,1)$. Consider a sequence $(\mathcal{E}_s)_{s\geq 0}$ satisfying $\mathcal{E}_s \geq 2$ and*

$$4K \sum_{s\geq 7} e^{-\mathcal{E}_s} \leq \beta. \tag{19}$$

Consider u_k the smallest integer such that

$$\frac{u_k}{\mathcal{E}_{u_k}} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}. \tag{20}$$

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

Let $q > 1$ be a fixed parameter. A typical choice for $\mathcal{E}_s$ is

$$\mathcal{E}_s = \log(Ks^q\beta^{-1}) \vee 2, \tag{21}$$

up to some additive constant ensuring that (19) holds. For this choice, Theorem 8 implies that for some positive constant κ, with probability at least $1-\beta$, for any suboptimal arm k (i.e., $\Delta_k > 0$), its number of play is bounded by

$$\mathcal{T}_{k,\beta} \triangleq \kappa\Big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{1}{\Delta_k}\Big) \log\Big[K\Big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{b}{\Delta_k}\Big)\beta^{-1}\Big],$$

which is independent of the total number of plays! This directly leads to the following upper bound on the regret of the policy at time n

$$\sum_{k=1}^K T_k(n)\Delta_k \leq \sum_{k:\Delta_k>0} \mathcal{T}_{k,\beta}\Delta_k. \tag{22}$$

One should notice that the previous bound holds with probability at least $1-\beta$ and on the complement set no small upper bound is possible: one can find a situation in which with probability of order β, the regret is of order n (even if (22) holds with probability greater than $1-\beta$). More formally, this means that the following bound cannot be essentially improved (unless additional assumptions are imposed):

$$\mathbb{E}[R_n] = \sum_{k=1}^K \mathbb{E}[T_k(n)]\Delta_k \leq (1-\beta) \sum_{k:\Delta_k>0} \mathcal{T}_{k,\beta}\Delta_k + \beta n$$

As a consequence, if one is interested in having a bound on the expected regret at some fixed time n, one should take β of order $1/n$ (up to a logarithmic factor):

Theorem 9. *Let $n \geq 7$ be fixed. Consider the sequence $\mathcal{E}_s = \log[Kn(s+1)]$. For this sequence, the PAC-UCB policy satisfies*

- *with probability at least $1 - \frac{4\log(n/7)}{n}$, for any $k : \Delta_k > 0$, the number of plays of arm k up to time n is bounded by $1 + \big(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}\big)\log(Kn^2)$.*
- *the expected regret at time n satisfies*

$$\mathbb{E}[R_n] \leq \sum_{k:\Delta_k>0} \Big(\frac{24\sigma_k^2}{\Delta_k} + 30b\Big)\log(n/3). \tag{23}$$

6 Open problem

When the horizon time n is known, one may want to choose the exploration function $\mathcal{E}$ depending on the value of n. For instance, in view of Theorems 3 and 6, one may want to take $c = 1$ and a constant exploration function $\mathcal{E} \equiv 3 \log n$. This choice ensures logarithmic expected regret and a nice concentration property:

$$\mathbb{P}\Big\{ R_n > 24 \sum_{k:\Delta_k>0} \Big(\frac{\sigma_k^2}{\Delta_k} + 2b \Big) \log n \Big\} \leq \frac{C}{n}. \tag{24}$$

This algorithm does not behave as the one which simply takes $\mathcal{E}_{s,t} = 3 \log t$. Indeed the algorithm with constant exploration function $\mathcal{E}_{s,t} = 3 \log n$ concentrates its exploration phase at the beginning of the plays, and then switches to exploitation mode. On the contrary, the algorithm which adapts to the time horizon explores and exploits during all the time interval $[0; n]$. However, in view of Theorem 7, it satisfies only

$$\mathbb{P}\Big\{ R_n > 24 \sum_{k:\Delta_k>0} \Big(\frac{\sigma_k^2}{\Delta_k} + 2b \Big) \log n \Big\} \leq \frac{C}{(\log n)^C}.$$

which is significantly worse than (24). The open question is: is there an algorithm that adapts to time horizon which has a logarithmic expected regret and a concentration property similar to (24)? We conjecture that the answer is no.

Acknowledgements. Csaba Szepesvári greatly acknowledges the support received through the Alberta Ingenuity Center for Machine Learning (AICML) and the Computer and Automation Research Institute of the Hungarian Academy of Sciences, and the PASCAL pump priming project "Sequential Forecasting". The authors thank Yizao Wang for the careful reading of an earlier version of this paper.

References

1. R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.
2. J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Variance estimates and exploration function in multi-armed bandit. Research report 07-31, Certis - Ecole des Ponts, http://cermics.enpc.fr/ audibert/RR0731.pdf, 2007.
3. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
4. P. Auer, N. Cesa-Bianchi, and J. Shawe-Taylor. Exploration versus exploitation challenge. In *2nd PASCAL Challenges Workshop*. Pascal Network, 2006.
5. J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley-Interscience series in systems and optimization. Wiley, Chichester, NY, 1989.
6. T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

7. T.L. Lai and S. Yakowitz. Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, 40:1199–1209, 1995.
8. H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.
9. W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.