Variance estimates and exploration function in multi-armed bandit

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

CERTIS Research Report 07-31
also Willow Technical report 01-07
April 2007

Ecole des ponts - Certis
6-8 avenue Blaise Pascal
77420 Champs-sur-Marne
France

Inria - Rocquencourt
Domaine Voluceau-Rocquencourt
78153 Le Chesnay Cedex
France

Ecole Normale Supérieure - DI
45, rue d'Ulm
75005 Paris
France

[1]Willow Project, Certis Lab, ParisTech-Ecole des Ponts, 77455 Marne la Vallee, France,http://www.enpc.fr/certis/
[2]Sequel, INRIA Futurs, Université Lille 3, France
[3]Department of Computing Science, University of Alberta, Canada

Variance estimates and exploration function in multi-armed bandit

Estimation de la variance et exploration pour le bandit à plusieurs bras

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

[1]Willow Project, Certis Lab, ParisTech-Ecole des Ponts, 77455 Marne la Vallee, France,http://www.enpc.fr/certis/

[2]Sequel, INRIA Futurs, Université Lille 3, France

[3]Department of Computing Science, University of Alberta, Canada

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 some very special bandit problems, for upper confidence bound based algorithms with standard bias sequences, the regret concentrates only at a polynomial rate. Hence, although these algorithms achieve logarithmic expected regret rates, they seem less attractive when the risk of achieving much worse than logarithmic cumulative regret is also taken into account.

Résumé

Les algorithmes réalisant le compromis exploration-exploitation à base de bornes supérieures des récompenses deviennent de plus en plus populaire en raison de leur succès pratiques récents Dans ce travail, nous considérons une variante de l'algorithme de base pour le problème du bandit à plusieurs bras. Cette variante, qui prend en compte les variances empiriques des récompenses obtenues sur les différents bras, a amélioré nettement les résultats obtenus précédemment. Le but de ce rapport est de fournir une explication rigoureuse de ces découvertes. Par ailleurs, nous clarifions les choix des paramètres de l'algorithme, et analysons la concentration du regret. Nous prouvons que de dernier est concentré seulement si la distribution des récompenses du bras optimal suit une hypothèse non triviale, ou quand l'algorithme est modifié de manière à explorer plus.

Contents

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 he does not know which of the treatments is the most effective one [11]. Multi-armed bandit problems became popular with the seminal paper of Robbins [10], 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 on-line 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.[1] By definition, *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 $\mu^* = \max_{1\leq k\leq K} \mu_k$.

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

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

achieved up to time n. It turns out that the expected regret satisfies

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

where $\Delta_k = \mu^* - \mu_k$ 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., [6]). Lai and Robbins [8] studied bandit problems with parametric uncertainties. They introduced an algorithm that follows what is now called the "optimism in the face of uncertainty principle". 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 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 has shown how to construct such optimal policies starting from the sample-means of the arms [1]. More recently, Auer et. al considered the case when the rewards come from a bounded support, say $[0, b]$, but otherwise the reward distributions are unconstrained [3]. 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 more desirable the 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 $\mathcal{E}_{s,t} = \zeta \log(t)$. 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).[2] 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

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

"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.16).

Interestingly, our analysis revealed some tradeoffs 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 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, hence 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 $0 < \beta < 1$, the algorithm achieves finite cumulative regret with probability $1 - \beta$ (Theorem 8). Hence, we name this variant PAC-UCB ("Probably approximately correct UCB"). Given 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.

Such high probability bounds show how risk behaves and thus have important

practical implications. Further, concentration of the regret also plays an important role in algorithms like UCT which treat sequential stochastic search problems as a sequences of nested bandit problems [7].

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 $b > 0$. For easy reference we summarize our assumptions on the reward sequence here:

Assumption A1 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.[3] Assume that the rewards of different arms are independent of each other, i.e., for any k, k', $1 \le k < k' \le K$, $t \in \mathbb{N}^+$, 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 $c \ge 0$. Let $\mathcal{E} = (\mathcal{E}_{s,t})_{s \ge 0, t \ge 0}$ be nonnegative real numbers such that for any $s \ge 0$, 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 $1/0 = +\infty$.

[3] The i.i.d. assumption can be relaxed, see e.g., [9].

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

Let us roughly describe the behaviour 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 means. 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 optimal 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*

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

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

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

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

Proof. See Section A.1. □

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* (4) *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. The bounds presented here significantly improve the ones used in [3]. The improvement is a necessary step to get tight bounds for the interesting case where the exploration function is logarithmic.

Theorem 2. *After K plays, each arm has been pulled once. Let arm k and time $n \in \mathbb{N}^+$ be fixed. For any $\tau \in \mathbb{R}$ and any integer $u > 1$, we have*

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

hence

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

Besides we have

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

Even if the above statements hold for any arm, they will be only useful for suboptimal arms.

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

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

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 that the exploration function $\xi_{s,t}$ is a nondecreasing function with respect to t. Consider an event such that the following statements hold:

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

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

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

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 (still, $\mathcal{E} = (\mathcal{E}_t)_{t\geq 0}$ is a nondecreasing of 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*

$$\begin{aligned}\mathbb{E}[R_n] \;\leq\; \sum_{k:\Delta_k>0} \Bigg\{ & 1 + 8(c\vee 1)\mathcal{E}_n \left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right) \\ & + n e^{-\mathcal{E}_n}\left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) + \sum_{t=16\mathcal{E}_n}^{n} \beta\big((c\wedge 1)\mathcal{E}_t, t\big)\Bigg\}\Delta_k,\end{aligned} \tag{11}$$

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 the smallest integer larger than $8\big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\big)\mathcal{E}'_n$ and $\tau = \mu^*$. The above choice of u guarantees that for any $u \leq s < t$ and $t \geq 2$,

$$\begin{aligned}&\sqrt{\frac{2[\sigma_k^2 + b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\frac{\mathcal{E}_t}{s} \leq \sqrt{\frac{[2\sigma_k^2 + b\Delta_k]\mathcal{E}'_n}{u}} + 3b\frac{\mathcal{E}'_n}{u} \\ &\quad\leq \sqrt{\tfrac{[2\sigma_k^2+b\Delta_k]\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]}} + \tfrac{3b\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]} = \tfrac{\Delta_k}{2}\left[\sqrt{\tfrac{2\sigma_k^2+b\Delta_k}{2\sigma_k^2+4b\Delta_k}} + \tfrac{3b\Delta_k}{4\sigma_k^2+8b\Delta_k}\right] \leq \tfrac{\Delta_k}{2},\end{aligned} \tag{12}$$

since the last inequality is equivalent to $(x-1)^2 \geq 0$ for $x = \sqrt{\frac{2\sigma_k^2+b\Delta_k}{2\sigma_k^2+4b\Delta_k}}$.

For any $s \geq u$ and $t \geq 2$, we have

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

where in the last step we used Bernstein's inequality (see (33)) twice. Summing

up these probabilities we obtain

$$\sum_{s=u}^{t-1}\mathbb{P}(B_{k,s,t}>\mu^*)\leq 2\sum_{s=u}^{\infty}e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}=2\frac{e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}{1-e^{-\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}$$
$$\leq\left(\tfrac{24\sigma_k^2}{\Delta_k^2}+\tfrac{4b}{\Delta_k}\right)e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}\leq\left(\tfrac{24\sigma_k^2}{\Delta_k^2}+\tfrac{4b}{\Delta_k}\right)e^{-\mathcal{E}'_n},\qquad(14)$$

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

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

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

In order to balance the terms in (11) 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 $n\geq 2$*

$$\mathbb{E}[R_n]\quad\leq\quad c_\zeta\sum_{k:\Delta_k>0}\left(\frac{\sigma_k^2}{\Delta_k}+2b\right)\log n.\qquad(15)$$

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

Proof. The first part follows directly from Theorem 3. The numerical assertion is tedious. It consists in bounding the four terms between brackets in (11). First it uses that

- bn is always a trivial upper bound on R_n,
- $b(n-1)$ is a trivial upper bound on R_n when $n\geq K$ (since in the first K rounds, you draw exactly once the optimal arm).

As a consequence, the numerical bound is non-trivial 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 (11) is bounded by $a_2\big(\frac{\sigma_k^2}{\Delta_k^2}+\frac{2b}{\Delta_k}\big)\log n$, 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 on algorithms that does not take into account the empirical variance of the reward distributions (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.*

Proof. We consider the following reward distributions:

- arm 1 concentrates its rewards on 0 and 1 with equal probability.
- the other arms always provide a reward equal to $\frac{1}{2} - \varepsilon_n$.

Arm 1 is therefore the optimal arm. After $\tilde{s}$ plays of the optimal arm, since we have necessarily $V_{k,\tilde{s}} \leq 1/4$, using $\tilde{b} \triangleq 3cb\zeta$, we can write for any $t \leq n$

$$\begin{aligned} B_{1,\tilde{s},t} &= \overline{X}_{1,\tilde{s}} + \sqrt{\tfrac{2V_{1,\tilde{s}}\zeta \log t}{\tilde{s}}} + \tilde{b}\tfrac{\log t}{\tilde{s}} \\ &\leq \tfrac{1}{2} + \left(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2}\right) + \sqrt{\tfrac{\zeta \log n}{2\tilde{s}}} + \tilde{b}\tfrac{\log n}{\tilde{s}}. \end{aligned} \tag{16}$$

On the other hand, for any $0 \leq s \leq t$, we have

$$B_{2,s,t} = \tfrac{1}{2} - \varepsilon_n + \tilde{b}\tfrac{\log t}{s} \geq \tfrac{1}{2} - \varepsilon_n. \tag{17}$$

So the algorithm will behave badly if with non-negligible probability, for some $s^* \ll n$, we have $B_{1,s^*,t} < 1/2 - \varepsilon_n$.

n is large enough). To help us choosing $\tilde{s}$ and ε_n, we need a lower bound on the deviation of $\overline{X}_{1,\tilde{s}} - 1/2$. This is obtained through Stirling's formula

$$n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n+1)} < n! < n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n)}, \tag{18}$$

since for ℓ such that $(\tilde{s}+\ell)/2 \in \mathbb{N}$, it leads to:

$$\begin{aligned}
&\mathbb{P}\big(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2} = -\tfrac{\ell}{2\tilde{s}}\big) \\
&= \big(\tfrac{1}{2}\big)^{\tilde{s}} \binom{\tilde{s}}{\frac{\tilde{s}+\ell}{2}} \\
&\geq \big(\tfrac{1}{2}\big)^{\tilde{s}} \frac{(\frac{\tilde{s}}{e})^{\tilde{s}}\sqrt{2\pi\tilde{s}}e^{\frac{1}{12\tilde{s}+1}}}{(\frac{\tilde{s}+\ell}{2e})^{\frac{\tilde{s}+\ell}{2}}(\frac{\tilde{s}-\ell}{2e})^{\frac{\tilde{s}-\ell}{2}}\sqrt{\pi(\tilde{s}+\ell)}\sqrt{\pi(\tilde{s}-\ell)}e^{\frac{1}{6(\tilde{s}+\ell)}}e^{\frac{1}{6(\tilde{s}-\ell)}}} \\
&= \frac{1}{(1+\frac{\ell}{\tilde{s}})^{\frac{\tilde{s}+\ell}{2}}(1-\frac{\ell}{\tilde{s}})^{\frac{\tilde{s}-\ell}{2}}}\sqrt{\frac{2\tilde{s}}{\pi(\tilde{s}^2-\ell^2)}}e^{\frac{1}{12\tilde{s}+1}-\frac{1}{6(\tilde{s}+\ell)}-\frac{1}{6(\tilde{s}-\ell)}} \\
&\geq \sqrt{\frac{2}{\pi\tilde{s}}}\Big(1-\frac{\ell^2}{\tilde{s}^2}\Big)^{-\frac{\tilde{s}}{2}}\Big(\frac{1-\frac{\ell}{\tilde{s}}}{1+\frac{\ell}{\tilde{s}}}\Big)^{\frac{\ell}{2}}e^{-\frac{1}{6(\tilde{s}+\ell)}-\frac{1}{6(\tilde{s}-\ell)}} \\
&\geq \sqrt{\frac{2}{\pi\tilde{s}}}e^{-\frac{\ell^2}{2\tilde{s}}-\frac{1}{6(\tilde{s}+\ell)}-\frac{1}{6(\tilde{s}-\ell)}}.
\end{aligned} \tag{19}$$

Let $\lfloor x \rfloor$ be the largest integer smaller or equal to x. Introduce κ a constant parameter. By summing $\lfloor\sqrt{\tilde{s}}\rfloor$ well chosen probabilities, i.e., the largest probabilities $\mathbb{P}\big(\overline{X}_{1,\tilde{s}} - \frac{1}{2} = -\frac{\ell}{2\tilde{s}}\big)$ for $\ell \geq \sqrt{2\kappa\tilde{s}\log\tilde{s}}$, we get that for some positive constant $C > 0$

$$\mathbb{P}\left(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2} \leq -\sqrt{\frac{\kappa\log\tilde{s}}{2\tilde{s}}}\right) \geq C\tilde{s}^{-\kappa}. \tag{20}$$

Let $\zeta' \in]\zeta; 1[$ such that $n^{\zeta'/\kappa}$ is an integer number. We consider $\tilde{s} = n^{\zeta'/\kappa}$ so that from (16), we obtain

$$\mathbb{P}\left(B_{1,\tilde{s},t} \leq \tfrac{1}{2} - (\sqrt{\zeta'}-\sqrt{\zeta})\sqrt{\frac{\log n}{2n^{\kappa'}}} + \tilde{b}\frac{\log n}{n^{\kappa'}}\right) \geq Cn^{-\kappa\kappa'}. \tag{21}$$

In view of (17), we take $\varepsilon_n = \frac{\sqrt{\zeta'}-\sqrt{\zeta}}{2}\sqrt{\frac{\log n}{2n^{\kappa'}}}$ such that with probability at least $Cn^{-\zeta'}$, we draw the optimal arm no more than $\tilde{s} = n^{\zeta'/\kappa}$ times. Up to multiplicative constants, this leads to an expected number of draws of suboptimal arms larger than $(n - n^{\zeta'/\kappa})n^{-\zeta'} \approx n^{1-\zeta'}$ and an expected regret larger than $(n - n^{\zeta'/\kappa})\varepsilon_n n^{-\zeta'} \approx n^{1-\zeta'} \approx n^{1-\zeta'-\zeta'/\kappa}$ up to a logarithmic factor. Taking κ sufficiently large, for $\zeta < 1$, there exists $\zeta' \in]\zeta; 1[$ such that $1 - \zeta' - \zeta'/\kappa > 0$, so that we have obtained that polynomial expected regret can occur as soon as $\zeta < 1$. □

So far we have seen that for $c = 1$ and $\zeta > 1$ we obtain a logarithmic regret, and that the constant ζ could 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.*

Proof. See Section A.2. □

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.16, 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

$$\tilde{\beta}(t) \triangleq 3 \min_{\substack{\alpha\geq 1 \quad M\in\mathbb{N} \\ s_0=0<s_1<\cdots<s_M=n \\ \text{s.t. } s_{j+1}\leq\alpha(s_j+1)}} \sum_{j=0}^{M-1} e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+t+1}}{\alpha}}. \tag{22}$$

We have seen in the previous section that 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 complicate sum of (22), up to second order logarithmic terms, is of the order of $e^{-(c\wedge 1)\mathcal{E}_t}$. This can be seen by considering (disregarding rounding issues) the geometric grid $s_j = \alpha^j$ with α close to 1. Let $\lfloor x \rfloor$ still denote the largest integer smaller or equal to x. The next theorem provides a bound for the tails of the pseudo-regret.

Theorem 6. *Let*

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

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

$$\mathbb{P}\big(R_n > r_0 x\big) \leq \sum_{k:\Delta_k>0} \Big\{ 2ne^{-(c\vee 1)\mathcal{E}_n x} + \tilde{\beta}(\lfloor v_k \mathcal{E}_n x \rfloor) \Big\}, \tag{23}$$

where we recall that $\tilde{\beta}(t)$ is essentially of order $e^{-(c\wedge 1)\mathcal{E}_t}$ (see (22)).

Proof. First note that

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

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 (13), 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 $s_0, \ldots, s_M$ of $\{1, \ldots, n\}$ realizing the minimum of (22) when $t = u$. Let $I_j = \{s_j + 1, \ldots, s_{j+1}\}$. Then

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

which gives the desired result. □

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,

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

for some constant $C > 0$:

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, $(\sigma_k)_{k\in\{1,\ldots,K\}}$, $(\Delta_k)_{k\in\{1,\ldots,K\}}$ 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 $z > \kappa_1 \log n$*

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

Proof. For $\kappa_3 > 0$ and $\kappa_4 > 0$ well chosen and depending only on b, K, $(\sigma_k)_{k\in\{1,\dots,K\}}$, $(\Delta_k)_{k\in\{1,\dots,K\}}$, Theorem 6 can be written as

$$\mathbb{P}(R_n > \kappa_3 \mathcal{E}_n x) \leq 2nKe^{-\mathcal{E}_n x} + K\tilde{\beta}(z'),$$

where $z' = \lfloor \kappa_4 \mathcal{E}_n x \rfloor$. Considering $x = z/(\kappa_3 \mathcal{E}_n)$, we obtain

$$\mathbb{P}(R_n > z) \leq 2nKe^{-z/\kappa_3} + K\tilde{\beta}(z').$$

For $\kappa_1 \triangleq 2\kappa_3$ and $z > \kappa_1 \log n$, the first term of the r.h.s is bounded with $2Ke^{-z/(2\kappa_3)}$, which can be bounded with $\kappa_2 \frac{\log z}{z^\zeta}$ for appropriate choice of κ_2 (depending only on b, K, $(\sigma_k)_{k\in\{1,\dots,K\}}$, $(\Delta_k)_{k\in\{1,\dots,K\}}$). To upper bound $\tilde{\beta}(z')$ (see defintion in (22)), we consider a geometric grid of step $\alpha = 1/(1-\varepsilon)$, and cut the sum in $\tilde{\beta}$ in two parts: for the j's for which $s_j \leq z'$, we use

$$e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+z'+1}}{\alpha}} \leq e^{-\frac{\mathcal{E}_{z'}}{\alpha}} = (z')^{-\zeta(1-\varepsilon)},$$

whereas for the j's for which $s_j \leq t$, $e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+z'+1}}{\alpha}} \leq e^{-\frac{\mathcal{E}_{s_j}}{\alpha}} \leq e^{-j\frac{\log\alpha}{\alpha}}$. The first sum on j's has at most $1 + (\log z')/\log[1/(1-\varepsilon)]$ terms, whereas the second sum on j's is of order of its first term since it is geometrically decreasing. This finishes the proof. □

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 increases with increasing 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 (yielding small regret) in which case the optimal regret 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 whose decays only at a polynomial rate, which is controlled by ζ.[4] 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.*

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

Proof. Let $\tilde{\mu}$ be the essential infimum of the optimal arm. Assume that $\tilde{\mu} > \mu_{\tilde{k}}$. Then there exists μ' such that $\mu_{\tilde{k}} < \mu' < \tilde{\mu}$. For any arm k, introduce $\delta_k = \mu' - \mu_k$. Let us use (9) with $\tau = \mu'$ and where u is the smallest integer larger than $8\big(\frac{\sigma_k^2}{\delta_k^2} + \frac{2b}{\delta_k}\big)\mathcal{E}'_n$. This value of τ makes the last probability in (9) vanish. The first term is controlled as in the proof of Theorem 6. Precisely, we obtain for $v'_k \triangleq 8(c \vee 1)\big(\frac{\sigma_k^2}{\delta_k^2} + \frac{2b}{\delta_k}\big)$, $r'_0 \triangleq \sum_{k:\Delta_k>0} \Delta_k\big(1 + v'_k\mathcal{E}_n\big)$ and any $x \geq 1$

$$\mathbb{P}\big(R_n > r'_0 x\big) \;\; \leq \;\; 2e^{\log(Kn) - (c\vee 1)\mathcal{E}_n x},$$

which proves that R_n has exponential tails in this case.

On the contrary, when $\tilde{\mu} < \mu_{\tilde{k}}$, we consider the following reward distributions:

- the optimal arm concentrates its rewards on $\tilde{\mu}$ and b such that its expected reward is strictly larger than $\mu_{\tilde{k}}$,
- all suboptimal arms are deterministic to the extent that they always provide a reward equal to $\mu_{\tilde{k}}$.

Let q be any positive integer. Consider the event:

$$\big\{X_{1,1} = X_{1,2} = \ldots = X_{1,q} = \tilde{\mu}\big\}.$$

Let $c_2 \triangleq c\zeta$ and $\eta \triangleq \mu_{\tilde{k}} - \tilde{\mu}$. On this event, we have for any $t \leq e^{\eta q/c_2}$

$$B_{1,q,t} = \tilde{\mu} + c_2\frac{\log t}{q} \leq \mu_{\tilde{k}}.$$

Besides for any $0 \leq s \leq t$, we have

$$B_{2,s,t} = \mu_{\tilde{k}} + c_2\frac{\log t}{s} > \mu_{\tilde{k}}.$$

This means that the optimal arm cannot be played more than q times during the first $e^{\eta q/c_2}$ plays. This gives a regret and a pseudo-regret of at least $\Delta_{\tilde{k}}\big(e^{\eta q/c_2} - q\big)$. So the pseudo-regret cannot have thinnest tails than polynomial ones. □

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 slightly modifying the analysis 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 Cz for some $C > 0$. This means that the algorithm draws an*

optimal arm at least $n - Cz$. 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 appropriate sequence $(\mathcal{E}_s)_{s\geq 0}$, this leads to 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*

$$4K \sum_{s\geq 7} e^{-\mathcal{E}_s} \leq \beta. \tag{24}$$

Consider u_k the smallest integer such that

$$\frac{u_k}{\mathcal{E}_{u_k}} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}. \tag{25}$$

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

Proof. See Section A.3. □

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

$$\mathcal{E}_s = \log(Ks^q\beta^{-1}) \vee 2, \tag{26}$$

up to some additive constant ensuring that (24) 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

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

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

$$\sum_{k=1}^K T_k(n)\Delta_k \leq \sum_{k:\Delta_k>0} \mathcal{T}_{k,\beta}\Delta_k. \tag{27}$$

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 (27) holds with probability greater than $1-\beta$). More formally, this means that the following bound cannot be essentially improved (unless putting additional assumptions):

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

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

Theorem 9. *Let $n \geq 7$. 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 $1 + \left(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}\right)\log(Kn^2)$.*
- *the expected regret at time n satisfies*

$$\mathbb{E}[R_n] \leq \sum_{k:\Delta_k>0} \left(\frac{24\sigma_k^2}{\Delta_k} + 30b\right)\log(n/3). \tag{28}$$

Proof. See Section A.4. □

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:

$$\mathbb{P}\left\{R_n > 24\sum_{k:\Delta_k>0}\left(\frac{\sigma_k^2}{\Delta_k} + 2b\right)\log n\right\} \quad \leq \quad \frac{C}{n}. \tag{29}$$

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

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

which is significantly worse than (29). 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 (29)? We conjecture that the answer is no.

A Proofs of the results

A.1 Proof of Theorem 1

The result follows from a version of Bennett's inequality which gives a high probability confidence interval for the mean of an i.i.d. sequence:

Lemma 1. *Let U be a real-valued random variable such that almost surely $U \leq b''$ for some $b'' \in \mathbb{R}$. Let $\mu = \mathbb{E}[U]$, $b' \triangleq b'' - \mu$, and $b''_+ = b'' \vee 0$. Let $U_1, \ldots, U_n$ be i.i.d. copies of U, $\overline{U}_t = 1/t \sum_{s=1}^{t} U_s$. The following statements are true for any $x > 0$:*

- *with probability at least $1 - e^{-x}$, simultaneously for $1 \leq t \leq n$,*

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{E}[U^2]\, x} + b''_+ x/3, \tag{30}$$

- *with probability at least $1 - e^{-x}$, simultaneously for $1 \leq t \leq n$,*

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{V}\mathrm{ar}(U)\, x} + b' x/3. \tag{31}$$

Proof of Lemma 1. Let $v = (\mathbb{V}\mathrm{ar}\, U)/(b')^2$. To prove this inequality, we use Result (1.6) of Freedman [5] to obtain that for any $a > 0$

$$\mathbb{P}\big(\exists t : 0 \leq t \leq n \text{ and } t(\overline{U}_t - \mu)/b' \geq a\big) \leq e^{a + (a+nv)\log[nv/(nv+a)]}.$$

In other words, introducing $h(u) = (1+u)\log(1+u) - u$, with probability at least $1 - e^{-nvh[a/(nv)]}$, simultaneously for $1 \leq t \leq n$,

$$t(\overline{U}_t - \mu) < ab'$$

Consider $a = \sqrt{2nvx} + x/3$. To prove (31), it remains to check that

$$nvh[a/(nv)] \geq x. \tag{32}$$

This can be done by introducing $\varphi(r) = (1 + r + r^2/6)\log(1 + r + r^2/6) - r - 2r^2/3$. For any $r \geq 0$, we have $\varphi'(r) = (1 + r/3)\log(1 + r + r^2/6) - r$ and $3\varphi''(r) = \log(1 + r + r^2/6) - (r + r^2/6)/(1 + r + r^2/6)$, which is nonnegative since $\log(1 + r') \geq r'/(1 + r')$ for any $r' \geq 0$. The proof of (31) is finished since $\varphi(\sqrt{2x/(nv)}) \geq 0$ implies (32).

To prove (30), we need to modify the martingale argument underlying Freedman's result. Precisely, let $g(r) \triangleq (e^r - 1 - r)/r^2$, we replace

$$\mathbb{E}\left[e^{\lambda[U - \mathbb{E}U - \lambda g(\lambda b')\mathbb{V}\mathrm{ar}\, U]}\right] \leq 1$$

by (see e.g., [2, Chap. 2: Inequality (8.2) and Remark 8.1])

$$\mathbb{E}\left[e^{\lambda[U-\mathbb{E}U-\lambda g(\lambda b'')\mathbb{E}U^2]}\right] \leq 1.$$

By following Freedman's arguments, we get

$$\mathbb{P}\big(\exists t : 0 \leq t \leq n \text{ and } t(\overline{U}_t - \mu) \geq a\big) \leq \min_{\lambda>0} e^{-\lambda a+\lambda^2 g(\lambda b'')n\mathbb{E}[U^2]}.$$

Now if $b'' \leq 0$, this minimum is upper bounded with

$$\min_{\lambda>0} e^{-\lambda a+\frac{1}{2}\lambda^2 n\mathbb{E}[U^2]} = e^{-\frac{a^2}{2n\mathbb{E}[U^2]}},$$

which leads to (30) when $b'' \leq 0$. When $b'' > 0$, the minimum is reached for $\lambda b'' = \log\big(\frac{b''a+n\mathbb{E}[U^2]}{n\mathbb{E}[U^2]}\big)$ and then the computations are similar to the one developed to obtain (31). □

Remark 3. *Lemma 1 differs from the standard version of Bernstein's inequality in a few ways. The standard form of Bernstein's inequality (using the notation of this lemma) is as follows: for any $w > 0$,*

$$\mathbb{P}\big(\overline{U}_n - \mu > w\big) \leq e^{-\frac{nw^2}{2\mathbb{V}\mathrm{ar}(U)+(2b'w)/3}}. \tag{33}$$

When this inequality is used to derive high-probability confidence interval, we get

$$n(\overline{U}_n - \mu) \leq \sqrt{2n\mathbb{V}\mathrm{ar}\,(U)\,x} + 2\tfrac{b'x}{3}.$$

Compared with (31) *we see that the second term here is larger by a multiplicative factor of* 2*. This factor is saved thanks to the use of Bennett's inequality. Another difference is that Lemma 1 allows the time indices to vary in an interval. This form follows from a martingale's argument due to Freedman [5].*

Given Lemma 1, the proof of Theorem 1 essentially reduces to an application of the "square-root trick". For the first part of the theorem, we will prove a result slightly stronger since it will be useful to obtain the second part of Theorem 1: for any $x > 0$ and $n \in \mathbb{N}$, with probability at least $1 - 3e^{-x}$, for any $0 \leq t \leq n$,

$$|\overline{X}_t - \mu| < \tfrac{\sqrt{2nV_t x}}{t} + \tfrac{3bnx}{t^2}. \tag{34}$$

First, notice that if we prove the theorem for random variables with $b = 1$ then the theorem follows for the general case by a simple scaling argument.

Let σ denote the standard deviation of X_1: $\sigma^2 \triangleq \mathbb{V}\mathrm{ar}\, X_1$, and introduce $\mathcal{V} \triangleq \mathbb{E}\big[(X_1 - \mathbb{E}X_1)^4\big]$. Lemma 1, (31) with the choices $U_i = X_i$, $U_i = -X_i$, and

Lemma 1, (30) with the choice $U_i = -(X_i - \mathbb{E}[X_1])^2$ yield that with probability at least $1 - 3e^{-x}$, for any $0 \leq t \leq n$, we simultaneously have

$$|\overline{X}_t - \mu| \leq \sigma \tfrac{\sqrt{2nx}}{t} + \tfrac{x}{3t} \tag{35}$$

and

$$\sigma^2 \leq V_t + (\mu - \overline{X}_t)^2 + \tfrac{\sqrt{2n\mathcal{V}x}}{t}. \tag{36}$$

Let $L \triangleq nx/t^2$. We claim that from (35) and (36), it follows that

$$\sigma \leq \sqrt{V_t} + 1.8\sqrt{L}. \tag{37}$$

Since the random variable X_1 takes its values in $[0,1]$, we necessarily have $\sigma \leq 1/2$. Hence, when $1.8\sqrt{L} \geq 1/2$ then (37) is trivially satisfied, so from now on we may assume that $1.8\sqrt{L} \leq 1/2$, i.e., $L \leq (3.6)^{-2}$. Noting that $\mathcal{V} \leq \sigma^2$, by plugging (35) into (36) we obtain for any $0 \leq t \leq n$

$$\begin{array}{rcl} \sigma^2 & \leq & V_t + 2L\sigma^2 + \frac{2L}{3}\sigma\sqrt{2L} + \frac{L^2}{9} + \sigma\sqrt{2L} \\ & \leq & V_t + \frac{\sqrt{L}\sigma}{3.6} + \frac{2}{3\times(3.6)^2}\sigma\sqrt{2L} + \frac{L}{9\times(3.6)^2} + \sigma\sqrt{2L} \\ & \leq & V_t + 1.77\sqrt{L}\sigma + \frac{L}{100}, \end{array}$$

or $\sigma^2 - 1.77\sqrt{L}\sigma - (V_t + \frac{L}{100}) \leq 0$. The l.h.s. when viewed as a second order polynomial in σ has a positive leading term, hence its larger root gives an upper bound on σ: $\sigma \leq \frac{1.77}{2}\sqrt{L} + \sqrt{V_t + 0.8L} \leq \sqrt{V_t} + 1.8\sqrt{L}$, which finished the proof of (37). Plugging (37) into (35), we obtain

$$|\overline{X}_t - \mu| \leq \sqrt{2V_tL} + \big[1.8\sqrt{2} + 1/3\big]L < \sqrt{2V_tL} + 3L,$$

which, given the definition of L, ends the proof of (34), and thus the proof of the first part of Theorem 1.

Let us now consider the second part of the theorem: Fix $t_1 \leq t_2$, $t_1, t_2 \in \mathbb{N}$, let $\alpha \geq t_2/t_1$. From (34), with probability at least $1 - 3e^{-x/\alpha}$, for $t \in \{t_1, \ldots, t_2\}$, we have

$$\begin{array}{rcl} t|\overline{X}_t - \mu| & < & \sqrt{2t_2V_tx/\alpha} + 3x/\alpha \\ & \leq & \sqrt{2tV_tx} + 3x. \end{array} \tag{38}$$

To finish the proof, we use the previous inequality for well chosen intervals $[t_1; t_2]$ forming a partition of $[3; n]$. This last interval starts from 4 since (38) is trivial for $t < 4$. Precisely, introduce

$$\bar{\beta}(x,n) \triangleq 3 \min_{\substack{M\in\mathbb{N} \\ s_0=3<s_1<\cdots<s_M=n \\ \text{s.t. } s_{j+1}\leq\alpha(s_j+1)}} \sum_{j=0}^{M-1} e^{-x/\alpha}.$$

and let $s_0, \dots, s_M$ be the grid realizing the above minimum. We have

$$\begin{aligned}
&\mathbb{P}\big(\exists t : 1 \le t \le n \text{ s.t. } |\overline{X}_t - \mu| > \sqrt{\tfrac{2V_t x}{t}} + \tfrac{3x}{t}\big) \\
&\le \textstyle\sum_{j=0}^{M-1} \mathbb{P}\big(\exists t : s_j < t \le s_{j+1} \text{ s.t.} \\
&\qquad\qquad t|\overline{X}_t - \mu| > \sqrt{2tV_t x} + 3x\big) \\
&\le 3 \textstyle\sum_{j=0}^{M-1} e^{-x/\alpha} \\
&= \bar{\beta}(x, n) \\
&\le \beta(x, n),
\end{aligned}$$

where the last inequality comes from the use of a geometric grid of step α and a complete grid $\{3, 4, \dots, n\}$. This ends the proof of Theorem 1.

A.2 Proof of Theorem 5

We want to prove that if $c\zeta < 1/6$ then there exists a bandit problem such that UCB-V suffers a polynomial loss.

Let ϵ be a number in the $(0, 1)$ interval to be chosen later. Consider the following two-armed bandit problem: Let $\{X_{1t}\}$ be an i.i.d. Bernoulli sequence with $\mathbb{P}(X_{1t} = 1) = \epsilon$. Let $\{X_{2t}\}$ be the deterministic sequence given by $X_{2t} = \epsilon/2$. Thus, $\mu^* = \mu_1 = \mathbb{E}[X_{11}] = \epsilon > \epsilon/2 = \mathbb{E}[X_{21}] = \mu_2$, i.e., the first arm is the optimal one. Note that $b = 1$.

Since $c\zeta < 1/6$, we have $\delta \triangleq 1/6 - c\zeta > 0$. Hence we can choose ϵ in $(0, 1)$ such that

$$\tfrac{\log(1/(1-\varepsilon))}{\varepsilon} < \tfrac{1-3\delta}{1-6\delta}. \tag{39}$$

Indeed, such a value exists since $\lim_{\varepsilon \to 0} \log(1/(1-\varepsilon))/\varepsilon = 1$ and $(1-3\delta)/(1-6\delta) > 1$. Let $\gamma = (1-3\delta)/\log(1/(1-\varepsilon))$. Note that $\gamma > 0$. The following claim holds then:

Claim: Fix $n \in \mathbb{N}$ and consider an event when during the first $T = \lceil \gamma \log n \rceil$ pulls the optimal arm always returns 0. On such an event the optimal arm is not pulled more than T times during the time interval $[1, n]$, i.e., $T_1(n) \le T$.

Proof. Note that on the considered event $V_{1t} = 0$, $\overline{X}_{1t} = 0$ and hence

$$B_{1,T_1(t-1),t} = 3c\zeta \log(t)/T_1(t-1).$$

Further,

$$B_{2,T_2(t-1),t} = \varepsilon/2 + 3c\zeta \log(t)/T_2(t) \ge \varepsilon/2.$$

Let t_1 be the time t when arm one is pulled the T-th time. If $t_1 \ge n$ then the claim holds. Hence, assume that $t_1 < n$. In the next time step, $t = t_1 + 1$, we have

$T_1(t-1) = T$, hence

$$\begin{aligned} B_{1,T_1(t-1),t} &= 3c\tfrac{\zeta \log(t)}{T} \\ &\leq 3c\tfrac{\zeta \log(n)}{T} \\ &\leq 3c\tfrac{\zeta}{\gamma} \\ &= (1-6\delta)\tfrac{\log(1/(1-\varepsilon))}{2(1-3\delta)} \\ &< \tfrac{\varepsilon}{2}, \end{aligned}$$

where the last step follows by (39). Since $\varepsilon/2 \leq B_{2,T_2(t-1),t}$ it follows that the algorithm chooses arm 2 at time step $t_1 + 1$ and $T_1(t) = T$. Since the same argument can be repeated for $t_1 + 2, t_1 + 3, \ldots, n$, the claim follows. □

Now observe that the probability of the event that the optimal arm returns 0 during its first T pulls is

$$(1-\varepsilon)^T \geq (1-\varepsilon)^{\gamma \log n} = n^{\gamma \log(1-\varepsilon)} = n^{-(1-3\delta)}.$$

Further, when this event holds the regret is at least $(n-T)\varepsilon/2$. Thus, the expected regret is at least

$$\tfrac{\varepsilon}{2} n^{1-(1-3\delta)}(1 - \gamma(\log n)/n) = \tfrac{\varepsilon}{2} n^{3\delta}(1 - \gamma(\log n)/n),$$

thus finishing the proof.

A.3 Proof of Theorem 8

Without loss of generality (by a scaling argument), we may assume that $b = 1$. Consider the event $\mathcal{A}$ on which

$$\forall s \geq 7 \quad \forall k \in \{1, \ldots, K\} \quad \begin{cases} \left|\overline{X}_{k,s} - \mu_k\right| < \sigma_k \sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{\mathcal{E}_s}{3s} \\ \sigma_k \leq \sqrt{V_{k,s}} + 1.8\sqrt{\frac{\mathcal{E}_s}{s}} \\ \sqrt{V_{k,s}} \leq \sigma_k + \sqrt{\frac{\mathcal{E}_s}{2s}} \end{cases} \tag{40}$$

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

Proof. To prove the first two inequalities, the arguments are similar to the ones used in the proof of Theorem 1. The main difference here is that we want the third inequality to simultaneously hold. We apply Lemma 1 with $x = \mathcal{E}_s$, $n = s$ and different i.i.d. random variables: $W_i = X_{k,i}$, $W_i = -X_{k,i}$, $W_i = (X_{k,i} - \mu_k)^2$ and $W_i = -(X_{k,i} - \mu_k)^2$. We use that the second moment of the last two random

variables satisfies $\mathbb{E}[(X_{k,1}-\mu_k)^4] \leq \sigma_k^2$ and that the empirical expectation of $(X_{k,i}-\mu_k)^2$ is

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

We obtain that for any $s \geq 7$ and $k \in \{1,\dots,K\}$, with probability at least $1-4e^{-\mathcal{E}_s}$

$$\left\{\begin{array}{l} \left|\overline{X}_{k,s}-\mu_k\right| < \sigma_k\sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{\mathcal{E}_s}{3s} \\ \sigma_k^2 \leq V_{k,s} + (\overline{X}_{k,s}-\mu_k)^2 + \sqrt{\frac{2\sigma_k^2\mathcal{E}_s}{s}} \\ V_{k,s} + (\overline{X}_{k,s}-\mu_k)^2 \leq \sigma_k^2 + \sigma_k\sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{\mathcal{E}_s}{3s} \leq \left(\sigma_k + \sqrt{\frac{\mathcal{E}_s}{2s}}\right)^2 \end{array}\right.$$

As we have seen in Section A.1, the above first two inequalities give the first two inequalities of (40). Finally, taking the square root in the above third inequality gives the last inequality of (40).

Using an union bound, all these inequalities hold simultaneously with probability at least

$$1 - 4\sum_{k=1}^{K}\sum_{s\geq 7} e^{-\mathcal{E}_s} \geq 1-\beta.$$

■

□

Remember that $B_{k,s} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\mathcal{E}_s}{s}} + \frac{3\mathcal{E}_s}{s}$. Now let us prove that on the event $\mathcal{A}$, for any $s \geq 1$ and $k \in \{1,\dots,K\}$, we have $\mu_k \leq B_{k,s}$ and

$$B_{k,s} \leq \mu_k + 2\sigma_k\sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{13\mathcal{E}_s}{3s} \tag{41}$$

Proof. The inequality $\mu_k \leq B_{k,s}$ is obtained by plugging the second inequality of (40) in the first one of (40) and by noting that since $\mathcal{E}_s \geq 2$, the inequality is trivial for $s \leq 6$. Introduce $L_s = \frac{\mathcal{E}_s}{s}$. To prove (41), we used the first and third inequalities of (40) to obtain

$$\begin{array}{rcl} B_{k,s} & \leq & \mu_k + \sigma_k\sqrt{2L_s} + \frac{L_s}{3} + \sqrt{2L_s}\left(\sigma_k + \sqrt{L_s/2}\right) + 3L_s \\ & = & \mu_k + 2\sigma_k\sqrt{2L_s} + \frac{13L_s}{3}. \end{array}$$

Once more, the inequality is trivial for $s \leq 6$.

■

□

Now let us prove that the choice of u_k in Theorem 8 guarantees that

$$\mu_k + 2\sigma_k\sqrt{\tfrac{2\mathcal{E}_{u_k}}{u_k}} + \tfrac{13\mathcal{E}_{u_k}}{3u_k} < \mu^*. \tag{42}$$

Proof. For the sake of lightening the notation, let us drop for a moment the k indices, so that (42) is equivalent to

$$2\sigma\sqrt{\tfrac{2\mathcal{E}_u}{u}} + \tfrac{13\mathcal{E}_u}{3u} < \Delta. \tag{43}$$

Let $r = u/\mathcal{E}_u$. We have

$$\begin{array}{rcl} (43) & \Leftrightarrow & r - \frac{13}{3\Delta} > \frac{2\sigma}{\Delta}\sqrt{2r} \\ & \Leftrightarrow & r > \frac{13}{3\Delta} \quad \text{and} \quad \big(r - \frac{13}{3\Delta}\big)^2 > \frac{8\sigma^2}{\Delta^2} r \\ & \Leftrightarrow & r > \frac{13}{3\Delta} \quad \text{and} \quad r^2 - \big(\frac{8\sigma^2}{\Delta^2} + \frac{26}{3\Delta}\big) r + \frac{169}{9\Delta^2} > 0 \end{array}$$

This trivially holds for $r > \frac{8\sigma^2}{\Delta^2} + \frac{26}{3\Delta}$.

□

By adapting the argument leading to (10), we obtain

$$\begin{aligned} &\big\{\exists k : T_k(\infty) > u_k\big\} \\ &\quad \subset \Big(\big\{\exists k \text{ s.t. } B_{k,u_k} > \tau\big\} \cup \big\{\exists s \geq 1 \text{ s.t. } B_{k^*,s} \leq \tau\big\}\Big). \end{aligned}$$

Taking $\tau = \mu^*$ and using (42),we get

$$\begin{aligned} &\big\{\exists k : T_k(\infty) > u_k\big\} \\ &\quad \subset \Big(\big\{\exists k \text{ s.t. } B_{k,u_k} > \mu_k + 2\sigma_k\sqrt{\tfrac{2\mathcal{E}_{u_k}}{u_k}} + \tfrac{13\mathcal{E}_{u_k}}{3u_k}\big\} \\ &\qquad\qquad \cup\big\{\exists s \geq 1 \text{ s.t. } B_{k^*,s} \leq \mu^*\big\}\Big) \\ &\quad \subset \mathcal{A}. \end{aligned}$$

So we have proved that

$$\mathbb{P}\big(\exists k : T_k(\infty) > u_k\big) \leq \mathbb{P}(\mathcal{A}) \leq \beta,$$

which is the desired result.

A.4 Proof of Theorem 9

Consider the following sequence $\mathcal{E}'_s = \log[Kn(s+1)]$ for $s \leq n$ and $\mathcal{E}'_s = \infty$ otherwise. For this sequence, the assumptions of Theorem 8 are satisfied for $\beta = \frac{4\log(n/7)}{n}$ since $\sum_{7\leq s\leq n} 1/(s+1) \leq \log(n/7)$. Besides, to consider the sequence

$(\mathcal{E}'_s)_{s\geq 0}$ instead of $(\mathcal{E}_s)_{s\geq 0}$ does not modify the algorithm up to time n. Therefore with probability at least $1-\beta$, we have

$$\frac{T_k(n)-1}{\mathcal{E}_{T_k(n)-1}} \leq \frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k},$$

hence

$$T_k(n) \leq 1 + \left(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}\right)\log[KnT_k(n)], \tag{44}$$

which gives the first assertion.

For the second assertion, first note that since $R_n \leq n$, (28) is useful only when $30(K-1)\log(n/3) < n$. So the bound is trivial when $n \leq 100$ or when $K \geq n/50$. For $n > 100$ and $K < n/50$, (44) gives

$$T_k(n) \leq 1 + \left(\frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}\right)\log[n^3/50] \leq \left(\frac{24\sigma_k^2}{\Delta_k^2} + \frac{26b}{\Delta_k}\right)\log(n/3),$$

hence

$$\mathbb{E}[T_k(n)] \leq 4\log(n/7) + \left(\frac{24\sigma_k^2}{\Delta_k^2} + \frac{26b}{\Delta_k}\right)\log(n/3) \leq \left(\frac{24\sigma_k^2}{\Delta_k^2} + \frac{30b}{\Delta_k}\right)\log(n/3).$$

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. *PAC-Bayesian statistical learning theory*. PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. `http://cermics.enpc.fr/~audibert/ThesePack.zip`.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multi-armed 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] D.A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

[6] J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley-Interscience series in systems and optimization. Wiley, Chichester, NY, 1989.

[7] L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML-2006)*, pages 282–293, 2006.

[8] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[9] T.L. Lai and S. Yakowitz. Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, 40:1199–1209, 1995.

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

[11] 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.