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
April 2007, revised January 2008

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, 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, France, http://www.enpc.fr/certis/
[2]Sequel, INRIA Futurs, Université Lille 3, France
[3]Department of Computing Science, University of Alberta, Canada

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.

Exploration-exploitation trade-off using variance estimates in the multiarmed bandit setting

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] Department of Computing Science
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. This paper considers 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 paper provides a first analysis of the expected regret of such algorithms and of the concentration of the regret of upper confidence bounds algorithm. As expected, these analyses of the regret suggest that the algorithm that use the variance estimates can have a major advantage over its alternatives that do not use such estimates when the variances of the payoffs of the suboptimal arms are low. This work, however, reveals that the regret concentrates only at a polynomial rate. This holds for all the upper confidence bound based algorithms and for all bandit problems except those rare ones where with probability one the payoffs coming from the optimal arm are always larger than the expected payoff for the second best arm.
Hence, although upper confidence bound bandit algorithms achieve logarithmic expected regret rates, a risk-averse decision maker may prefer some alternative algorithm. The paper also illustrates some of the results with computer simulations.

1 Introduction and notations

In this paper we consider algorithms for *stochastic multi-armed bandit problems*. Bandit problems illustrate the fundamental difficulty of decision making in the face of uncertainty: A decision maker must choose between following what seems

to be the best choice ("exploit") or to explore some alternative hoping to discover a choice that is even better than the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when patients arrive sequentially and the effectiveness of the treatments are initially unknown (Thompson, 1933). Multi-armed bandit problems became popular with the seminal paper of Robbins (Robbins, 1952), 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 specified by K distributions, $\nu_1, \ldots, \nu_K$. The decision maker initially does not know these distributions, but can sample from them one by one. The samples are considered as rewards. The goal of the decision maker is to maximize the sum of the rewards, or, what is equivalent, to minimize his *regret*, i.e., the loss as compared to the total payoff that can be obtained given full knowledge of the problem.

The name 'bandit' comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which as a result produces a random payoff. If arm k is pulled the random payoff is drawn from ν_k. Thus payoff is assumed to be independent of all previous payoffs. Independence also holds across the arms. We will denote the payoff received when the k-th arm is pulled the t-th time by $X_{k,t}$.

Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, an efficient bandit algorithm must carefully balance *exploration and exploitation*.

A gambler learning about the distributions of the arms' payoffs can use all past information to decide about its next action. Designing a strategy for the gambler means that we pick a mapping ("policy") that maps the space of possible histories, $\cup_{t \in \mathbb{N}^+} \{1, \ldots, K\}^t \times \mathbb{R}^t$, into the set $\{1, \ldots, K\}$ (indexing the arms).

Let us formalize now the goal of the design problem. For this let $\mu_k = \mathbb{E}[X_{k,1}]$ denote the expected reward of arm k. 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 (we do not assume that the optimal arm is unique). 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 $I_t \in 1, \ldots, K$ denote the index of the arm arm played at time t. The *(cumulative) regret at time n* is defined by

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

Hence, the goal of the decision maker can be formalized as the problem of minimizing 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. Wald's equation implies 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., (Gittins, 1989)). Lai and Robbins (Lai & Robbins, 1985) studied bandit problems with parametric uncertainties in a minimax framework. They introduced an algorithm that follows what is now called the "optimism in the face of uncertainty principle". The algorithm works by computing *upper confidence bounds* for all the arms and then choosing the arm with the highest such bound. The upper confidence bound of an algorithm is obtained by maximizing the expected payoff when the parameters are varied within an appropriate confidence set. They proved that the expected regret of their algorithm increases at most at a logarithmic rate with the number of trials and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions). Agrawal has shown how to construct upper confidence bound algorithms that use the sample-means of the arms (Agrawal, 1995). 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 (Auer et al., 2002). They have studied several policies, most notably UCB1 which constructs the Upper Confidence Bound (UCB) for arm k at time t by adding the *bias factor*

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

to its sample-mean. In this paper the authors proved that the expected regret of this algorithm satisfies

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

In the same paper they propose UCB1-NORMAL, that is restricted to the case when the payoffs are normally distributed with unknown mean and variance. This algorithm estimates the variance of the arms and uses these estimates to refine the bias factor. Under the normality assumption they show that

$$\mathbb{E}[\hat{R}_n] \leq 8 \sum_{k:\mu_k<\mu^*} \left(\frac{32\sigma_k^2}{\Delta_k} + \Delta_k \right) \log(n) + O(1), \tag{3}$$

where μ_k denotes the mean payoff for arm k (as before), while σ_k^2 denotes the arm's variance.

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 $\Omega(b^2)$ (see Proposition 1, Section A.1). Since in many cases b is just a conservative, *a priori* guess on the size of the interval containing the rewards, it is more than desirable to lessen the

dependence of the algorithm on it. We see that UCB1-NORMAL achieves this perfectly. However, the price is high: We have to assume that the payoffs are normally distributed.

In the experimental section of their paper Auer et al. introduced another algorithm, called UCB1-Tuned. This algorithm, similarly to UCB1-NORMAL uses the empirical estimates of the variance in the bias sequence. However, unlike UCB1-NORMAL, this algorithm is designed to work with any bounded payoff distribution. The experiments of Auer et al. indicate that the idea of using empirical variance estimates works: UCB1-Tuned 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 (Auer et al., 2006). Intuitively, algorithms using variance estimates should work better than ones that do not use such estimates (like UCB1) when the variance of some suboptimal arm is much smaller than b^2. If this is the case a "variance-aware" algorithm can spot the suboptimal arms after a few trials, thereby reducing the regret suffered.

In this paper we study the regret of *UCB-V*, which is a generic UCB-type 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, $\mathcal{E} = \mathcal{E}_{\cdot,\cdot}$ (viewed as a function of (s,t)) is a so-called *exploration function*. A typical choice for this function is $\mathcal{E}_{s,t} = \zeta \log(t)$, where $\zeta, c > 0$ are tuning parameters that can be used to control the behavior of the algorithm.

The first major contribution of the paper (Theorem 4) is a bound on the expected regret of UCB-V that scales in an improved fashion with b. In particular, we show that for a particular settings of the parameters of the algorithm,

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

The main difference to the bound (2) is that b^2 is replaced by σ_k^2. However, notice that b still appears in the bound, which is a major difference to the bound (3). Although, this is unfortunate, it is possible to show that the dependence on b is unavoidable.

In order to prove the above result we will prove a novel tail bound on the sample average of i.i.d. random variables with bounded support. Unlike previous similar bounds, this bound uses the empirical variance and thus it might be of independent interest (Theorem 1).

Just like as it was done in the work of Auer et al., our regret bound also relies on the analysis of the sampling times of suboptimal arms (Theorem 2). However, compared to the analysis of (Auer et al., 2002), the new result is significantly improved. Thanks to this result. we obtain results on the expected regret for

a wide class of exploration functions (Theorem 3), leading to the main result already cited (Theorem 4). In addition, 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 5,6).

The second major contribution of the paper is the probabilistic analysis of the risk that the regret of the studied algorithm is much higher than its expected value. To our best knowledge, for this class of algorithms no such analysis existed previously. The concentration of regret results obtained are potentially important in the analysis of algorithms that nest sequences of bandits, such as the UCT algorithm proposed in (Kocsis & Szepesvári, 2006), which recently was proven to be very efficient in computer go (e.g., (Gelly & Silver, 2007)).

In order to analyze the risk, we will study the *(cumulative) pseudo-regret* defined by

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

Note that the expectation of the pseudo-regret and the regret are the same: $\mathbb{E}[R_n] = \mathbb{E}[\hat{R}_n]$, but the the randomness of the rewards influences the pseudo-regret only indirectly (i.e., through $\{T_k(n)\}$). In order to analyze the risk, in Sections 5 and 6 we develop high probability bounds for the pseudo-regret. Similar results can be obtained for the cumulative regret (see Remark 2 p.19).

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 10). In fact, this is always 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. Let the first arm be the optimal one: $\mu_1 > \mu_2$, $\Delta_2 = \mu_1 - \mu_2 > 0$. 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 9) makes the dependence of the risk on the algorithm's parameters 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. The theoretical findings of this part of the paper are illustrated in a series of experiments, described in Section 5.1.

In the final part of the paper (Section 6) we consider a variant of the problem when the time-horizon is given *a priori*. As it turns out in this case a good choice of the exploration function is to make it independent of the global time index t: $\mathcal{E}_{s,t} = \mathcal{E}_s$. 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 11). Hence, we name this variant of the algorithm 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 to hold at most with probability $O(1/n)$. 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 1. Hence, we conjecture that knowing the time horizon might represent a significant advantage.

2 The UCB-V algorithm

Let $\mathbb{N}$ denote the set of natural numbers including zero and let $\mathbb{N}^+$ denote the set of positive integers. For any $k \in \{1, \dots, K\}$ and $t \in \mathbb{N}$, let $\overline{X}_{k,t}$ and $V_{k,t}$ be the respective 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 any arm that has 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.[4] 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. Decision maker does not the distributions of the arms, but knows b.

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 fixed value of $s \ge 0$ 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 nonnegative integers s, t, introduce

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\mathcal{E}_{s,t}}{s}} + c\frac{3b\mathcal{E}_{s,t}}{s} \tag{4}$$

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

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

[4] The i.i.d. assumption can be relaxed, see e.g., (Lai & Yakowitz, 1995).

Let us summarize the main ideas underlying the algorithm. As long as an arm is never chosen its bound is infinite. Hence, initially the algorithm tries all the arms at least once (one by one). Then, the more an arm k has been tested, the closer the bound (4) gets to the sample-mean, and hence, by the law of large numbers, to the expected reward μ_k. So the procedure will hopefully tend to draw more often arms having the largest 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. This might make the sample-mean of this arm smaller than that of the other arms and hence an algorithm that only uses sample-means might get stuck with not choosing the optimal arm any more. The UCB-V policy uses the exploration function, $\mathcal{E}$, to prevent this situation. Indeed, assuming that for any fixed s, $\mathcal{E}_{s,t}$ increases without bounds in t we see that after a while the last term of (4) will start to dominate the two other terms and will also dominate the bound associated with the arms drawn very often. This will allow the algorithm to draw the optimal arm again, giving it a chance to develop a better estimate of the mean. We thus see that an appropriate choice of $\mathcal{E}$ encourages exploration; hence it's name. Naturally, an exploration function that tends to dominate the sample-means will not give enough room for the observed payoffs to influence the choices of the actions – the algorithm might draw suboptimal arms too often. Therefore $\mathcal{E}$ must be carefully chosen so as to balance exploration and exploitation. The major idea of upper-confidence bounds algorithms is that $\mathcal{E}$ should be selected such that $B_{k,s,t}$ is a high probability upper bound on the payoff of arm k. Then if no confidence bound fails then a suboptimal arm can only be chosen if its confidence bound is larger than its payoff difference to the optimal arm. Since confidence intervals shrink with increasing sample sizes the number of times the previous situation can happen is limited. Further, by designing $\mathcal{E}$ such that the error probabilities decay fast enough, we can make sure that the total error committed due to the failure of the confidence intervals is not too large either.

In our algorithm, the actual form of the exploration function comes from the following novel tail bound on the sample average of i.i.d. random variables with bounded support. The novelty of this bound is that, unlike previous similar bounds (e.g., Bennett's and Bernstein's inequalities), it 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 mean $\overline{X}_t$ 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 $1 - 3e^{-x}$,

$$|\overline{X}_t - \mu| \leq \sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}. \tag{5}$$

Furthermore, introducing

$$\beta(x,t) = 3 \inf_{1<\alpha\leq 3} \Big(\frac{\log t}{\log \alpha} \wedge t \Big) e^{-x/\alpha}, \tag{6}$$

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

$$|\overline{X}_s - \mu| \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s} \tag{7}$$

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

Proof. See Section A.2.

Remark 1. The uniformity in time is the only difference between the two assertions of the previous theorem. When we use (7), 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 (5) that is uniform in time. In particular, this means that if $1 \leq S \leq t$ is a random variable then (5) still holds with probability at least $1 - \beta(x,t)$ and when s is replaced with S.

Note that (5) is useless for $t \leq 3$ since its right-hand side (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 if $\mathcal{E}$ takes "sufficiently high values" then with high probability 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 types of exploration functions leading to interesting properties of the resulting algorithms in terms of expected regret, PAC bounds on the regret and adaptivity with respect to the total number of plays:

- the ones in which $\mathcal{E}_{s,t}$ depends only on t (see Sections 3 and 5).
- the ones in which $\mathcal{E}_{s,t}$ depends only on s (see Section 6).

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. In this section we derive such bounds, generalizing and improving upon the previous analysis of (Auer et al., 2002). The improvement is a necessary step to get *tight* bounds for exploration functions scaling logarithmically with t, which represent the most interesting class of exploration functions.

Since all the statements here make use of Assumption A1, we will refrain from citing it. Further, all the results in these sections are for algorithm UCB-V.

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

$$\begin{aligned} T_k(n) \le u + \textstyle\sum_{t=u+K-1}^{n} \big(& \mathbb{1}_{\{\exists s : u \le s \le t-1 \text{ s.t. } B_{k,s,t} > \tau\}} \\ & + \mathbb{1}_{\{\exists s^* : 1 \le s^* \le t-1 \text{ s.t. } \tau \ge B_{k^*,s^*,t}\}} \big), \end{aligned} \tag{8}$$

hence

$$\begin{aligned} \mathbb{E}\left[T_k(n)\right] \le u + \textstyle\sum_{t=u+K-1}^{n} \sum_{s=u}^{t-1} \mathbb{P}\big(B_{k,s,t} > \tau\big) \\ + \textstyle\sum_{t=u+K-1}^{n} \mathbb{P}\big(\exists s : 1 \le s \le t-1 \text{ s.t. } B_{k^*,s,t} \le \tau\big). \end{aligned} \tag{9}$$

Besides we have

$$\begin{aligned} & \mathbb{P}\big(T_k(n) > u\big) \\ & \quad \le \textstyle\sum_{t=u+1}^{n} \mathbb{P}\big(B_{k,u,t} > \tau\big) + \mathbb{P}\big(\exists s : 1 \le s \le n-u \text{ s.t. } B_{k^*,s,u+s} \le \tau\big). \end{aligned} \tag{10}$$

Note that even though if the above statements hold for any arm, naturally they provide useful bounds only for suboptimal arms.

Proof. The first assertion is trivial since at the beginning each arm has an infinite UCB value, which becomes finite as soon as the arm has been played once.

To obtain (8), we note that

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

where

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

Taking the expectation of both sides of (8) and using a union bound, we obtain (9).

Finally, (10) comes from a more direct argument that uses that the exploration function $\xi_{s,t}$ is a nondecreasing function with respect to t, which is developed next: Consider an event such that the following statements hold:

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

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

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

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

$$\begin{aligned} \big\{T_k(n) > u\big\} \subset \Big(& \big\{\exists t : u+1 \le t \le n \text{ s.t. } B_{k,u,t} > \tau\big\} \\ & \cup \big\{\exists s : 1 \le s \le n-u \text{ s.t. } B_{k^*,s,u+s} \le \tau\big\} \Big). \end{aligned} \tag{11}$$

By taking probabilities of both sides gives the announced result.

3 Expected regret of UCB-V

In this section, we assume that the exploration function does not depend on s (still, $\mathcal{E} = (\mathcal{E}_t)_{t\geq 0}$ is a nondecreasing function of t). We will see that as far as the expected regret is concerned, a natural choice for $\mathcal{E}_t$ 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 $\mathcal{E}_t$.

3.1 Upper bound on the expected regret

Theorem 3. *We have*

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

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 (6) and Remark 1).

Proof. Let $\mathcal{E}'_n = (c\vee 1)\mathcal{E}_n$. We use Equation (9) where we choose u to be the smallest integer larger than $8\big(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\big)\mathcal{E}'_n$ and we choose $\tau = \mu^*$.

This choice of u guarantees that for any $u \leq s < t$ and $t \geq 2$,

$$\begin{aligned} &\sqrt{\frac{2[\sigma_k^2 + b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\frac{\mathcal{E}_t}{s} \leq \sqrt{\frac{[2\sigma_k^2 + b\Delta_k]\mathcal{E}'_n}{u}} + 3b\frac{\mathcal{E}'_n}{u} \\ &\quad \leq \sqrt{\tfrac{[2\sigma_k^2+b\Delta_k]\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]}} + \tfrac{3b\Delta_k^2}{8[\sigma_k^2+2b\Delta_k]} = \tfrac{\Delta_k}{2}\left[\sqrt{\tfrac{2\sigma_k^2+b\Delta_k}{2\sigma_k^2+4b\Delta_k}} + \tfrac{3b\Delta_k}{4\sigma_k^2+8b\Delta_k}\right] \leq \tfrac{\Delta_k}{2}, \end{aligned} \tag{13}$$

where the last inequality holds as it 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

$$\begin{aligned} &\mathbb{P}(B_{k,s,t} > \mu^*) \leq \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2V_{k,s}\mathcal{E}_t}{s}} + 3bc\frac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} + \sqrt{\tfrac{2[\sigma_k^2+b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\tfrac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\big) + \mathbb{P}\big(V_{k,s} \geq \sigma_k^2 + b\Delta_k/2\big) \\ &\leq \mathbb{P}\big(\overline{X}_{k,s} - \mu_k > \Delta_k/2\big) + \mathbb{P}\Big(\tfrac{\sum_{j=1}^{s}(X_{k,j}-\mu_k)^2}{s} - \sigma_k^2 \geq b\Delta_k/2\Big) \\ &\leq 2e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}, \end{aligned} \tag{14}$$

where in the last step we used Bernstein's inequality (see (42)) twice. Summing up these probabilities we obtain

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

where we have used that $1 - e^{-x} \geq 2x/3$ for $0 \leq x \leq 3/4$. By using (7) of Theorem 1 to bound the other probability in (9), 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 gives the announced result since by assumption $u \geq 16\mathcal{E}_n$.

In order to balance the terms in (12) the exploration function should be chosen to be proportional to $\log t$, yielding the following upper estimate of the payoff of arm k that was chosen s times up to time t:

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2\zeta V_{k,s} \log t}{s}} + c\frac{3b \log t}{s}. \tag{16}$$

For this choice, the following theorem, the main result of this section, gives an explicit bound on the expected regret:

Theorem 4. *Let $c = 1$ and $\mathcal{E}_t = \zeta \log t$ for $\zeta > 1$. Then there exists a constant c_ζ depending only on ζ such that for $n \geq 2$*

$$\mathbb{E}[R_n] \leq c_\zeta \sum_{k:\Delta_k > 0} \left(\frac{\sigma_k^2}{\Delta_k} + 2b\right) \log n. \tag{17}$$

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

Proof. The first part follows directly from Theorem 3. The numerical assertion is tedious. It consists in bounding the four terms between brackets in (12). 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 (12) 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, Theorem 4 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 do not take into account the empirical variance of the reward distributions (see e.g. the bound (2) that holds for UCB1).

3.2 Lower limits on the bias sequence

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

Theorem 5. *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 $\frac{1}{2} - \varepsilon_n$.

Define $\tilde{b} \triangleq 3cb\zeta$.

Notice that arm 1 is the optimal arm. After $\tilde{s}$ plays of this arm, since we necessarily have $V_{k,\tilde{s}} \leq 1/4$, for any $t \leq n$ we have

$$\begin{aligned} B_{1,\tilde{s},t} &= \overline{X}_{1,\tilde{s}} + \sqrt{\tfrac{2V_{1,\tilde{s}}\zeta \log t}{\tilde{s}}} + \tilde{b}\tfrac{\log t}{\tilde{s}} \\ &\leq \tfrac{1}{2} + \left(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2}\right) + \sqrt{\tfrac{\zeta \log n}{2\tilde{s}}} + \tilde{b}\tfrac{\log n}{\tilde{s}}. \end{aligned} \tag{18}$$

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

$$B_{2,s,t} = \tfrac{1}{2} - \varepsilon_n + \tilde{b}\tfrac{\log t}{s} \geq \tfrac{1}{2} - \varepsilon_n. \tag{19}$$

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

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

$$n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n+1)} < n! < n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n)}. \tag{20}$$

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

$$\begin{aligned} &\mathbb{P}\left(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2} = -\tfrac{\ell}{2\tilde{s}}\right) \\ &\quad = \left(\tfrac{1}{2}\right)^{\tilde{s}} \binom{\tilde{s}}{\frac{\tilde{s}+\ell}{2}} \\ &\quad \geq \left(\tfrac{1}{2}\right)^{\tilde{s}} \frac{\left(\frac{\tilde{s}}{e}\right)^{\tilde{s}}\sqrt{2\pi\tilde{s}}\, e^{\frac{1}{12\tilde{s}+1}}}{\left(\frac{\tilde{s}+\ell}{2e}\right)^{\frac{\tilde{s}+\ell}{2}}\left(\frac{\tilde{s}-\ell}{2e}\right)^{\frac{\tilde{s}-\ell}{2}}\sqrt{\pi(\tilde{s}+\ell)}\sqrt{\pi(\tilde{s}-\ell)}\, e^{\frac{1}{6(\tilde{s}+\ell)}}\, e^{\frac{1}{6(\tilde{s}-\ell)}}} \\ &\quad = \frac{1}{\left(1+\frac{\ell}{\tilde{s}}\right)^{\frac{\tilde{s}+\ell}{2}}\left(1-\frac{\ell}{\tilde{s}}\right)^{\frac{\tilde{s}-\ell}{2}}}\sqrt{\frac{2\tilde{s}}{\pi(\tilde{s}^2-\ell^2)}}\, e^{\frac{1}{12\tilde{s}+1} - \frac{1}{6(\tilde{s}+\ell)} - \frac{1}{6(\tilde{s}-\ell)}} \\ &\quad \geq \sqrt{\frac{2}{\pi\tilde{s}}}\left(1 - \frac{\ell^2}{\tilde{s}^2}\right)^{-\frac{\tilde{s}}{2}}\left(\frac{1-\frac{\ell}{\tilde{s}}}{1+\frac{\ell}{\tilde{s}}}\right)^{\frac{\ell}{2}} e^{-\frac{1}{6(\tilde{s}+\ell)} - \frac{1}{6(\tilde{s}-\ell)}} \\ &\quad \geq \sqrt{\frac{2}{\pi\tilde{s}}}\, e^{-\frac{\ell^2}{2\tilde{s}} - \frac{1}{6(\tilde{s}+\ell)} - \frac{1}{6(\tilde{s}-\ell)}}. \end{aligned} \tag{21}$$

Let $\lfloor x \rfloor$ be the largest integer smaller or equal to x. Introduce a parameter, κ. 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 $C > 0$,

$$\mathbb{P}\left(\overline{X}_{1,\tilde{s}} - \tfrac{1}{2} \leq -\sqrt{\tfrac{\kappa \log \tilde{s}}{2\tilde{s}}}\right) \geq C\tilde{s}^{-\kappa}. \tag{22}$$

Let $\zeta' \in]\zeta; 1[$ such that $n^{\zeta'/\kappa}$ is an integer number. We consider $\tilde{s} = n^{\zeta'/\kappa}$ so that from (18), we obtain

$$\mathbb{P}\left(B_{1,\tilde{s},t} \leq \tfrac{1}{2} - (\sqrt{\zeta'} - \sqrt{\zeta})\sqrt{\tfrac{\log n}{2n^{\zeta'/\kappa}}} + \tilde{b}\tfrac{\log n}{n^{\zeta'/\kappa}}\right) \geq Cn^{-\zeta'}. \tag{23}$$

In view of (19), we take $\varepsilon_n = \frac{\sqrt{\zeta'}-\sqrt{\zeta}}{2}\sqrt{\frac{\log n}{2n^{\zeta'/\kappa}}}$ 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 $\zeta > 1$ the algorithm achieves logarithmic regret, and that the constant ζ could not be taken below 1 (independently of the value of c) without risking to suffer a polynomial regret. Now, let us consider the last term, which is linear in the ratio $\mathcal{E}_t/s$, in $B_{k,s,t}$. The next result shows that this term is also necessary to obtain a logarithmic regret:

Theorem 6. *Consider $\mathcal{E}_t = \zeta \log t$. Independently of the value of ζ, 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.3.

To conclude the above analysis, the natural choice for the bias sequence is

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

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 Risk bounds

Often decision makers are not satisfied with a good expected return. A stronger requirement is that the algorithms should give guaranteed returns with high

probability. With such a guarantee the decision maker is guaranteed to avoid unnecessarily high risks involving potentially huge losses. Hence the interest in studying the distributional properties of the regret. In the next section we provide tail bounds for the regret of UCB1 (we also provide a refined analysis of its expected regret), followed by a result in the subsequent section that concerns the tail behavior of UCB-V. These results are illustrated by computer experiments in Section 5.1.

4.1 Risk bounds for UCB1

In this section, we analyze the behavior of UCB1 in terms of the expected regret and the probability of high regret when the bias factor depends on a exploration coefficient $\alpha > 1$. The upper bounds take the form:

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + b\sqrt{\frac{\alpha \log t}{s}}. \tag{24}$$

We remind that in the original version of UCB1, the exploration coefficient was set to $\alpha = 2$. We show in the next result that the expected regret is $\mathbb{E}[R_n] = O(\alpha \log(n))$, which exhibits a linear dependency w.r.t. the coefficient α (the greater α the greater the exploration of all arms). Next, we provide an upper bound on the probability of high (pseudo-) regret of the form $\mathbb{P}(R_n > z) = O(z^{1-2\alpha})$ (the greater α the thinner the tail on the pseudo-regret).

The user may thus choose a range of possible algorithms between an algorithm (when setting α to a value close to 1) which yields low regret on the average but which may be risky (high probability of obtaining less rewards than expected), or an algorithm (when α is larger) which has a higher regret on the average, but which is more secure, in the sense that the actual regret is more concentrated around its expectation. Thus, the algorithm exhibits a tradeoff between expected reward and risk.

Theorem 7. *Let $\alpha > 1$. The expected pseudo-regret for UCB1 defined by (24) satisfies:*

$$\mathbb{E}[R_n] \leq \sum_{k:\Delta_k>0} \Big[\frac{4b^2}{\Delta_k}\alpha \log(n) + \Delta_k\big(\frac{3}{2} + \frac{1}{2(\alpha-1)}\big)\Big]. \tag{25}$$

Proof. For any sub-optimal arm k, let u denote the smallest integer larger than $\big(\frac{2b}{\Delta_k}\big)^2 \alpha \log(n)$. Thus, for any $u \leq s < t \leq n$, we have $b\sqrt{\frac{\alpha \log(t)}{s}} \leq \Delta_k/2$. From (9) with $\tau = \mu^*$, it comes

$$\begin{aligned}\mathbb{E}\left[T_k(n)\right] \leq u + \textstyle\sum_{t=u+1}^{n}\sum_{s=u}^{t-1} \mathbb{P}\big(B_{k,s,t} > \mu^*\big) \\ + \textstyle\sum_{t=u+1}^{n}\sum_{s=1}^{t-1} \mathbb{P}\big(B_{k^*,s,t} \leq \mu^*\big).\end{aligned} \tag{26}$$

Now, for $s \geq u$, $\mathbb{P}\big(B_{k,s,t} > \mu^*\big) = \mathbb{P}\big(\overline{X}_{k,s} + b\sqrt{\frac{\alpha \log(t)}{s}} > \mu_k + \Delta_k\big) \leq \mathbb{P}\big(\overline{X}_{k,s} > \mu_k + \Delta_k/2\big) \leq e^{-s\Delta_k^2/(2b^2)} \leq e^{-u\Delta_k^2/(2b^2)} \leq n^{-2\alpha}$, where we used the Chernoff-Hoeffding bound. We deduce that: $\sum_{t=u+1}^{n}\sum_{s=u}^{t-1} \mathbb{P}\big(B_{k,s,t} > \mu^*\big) \leq n^{2(1-\alpha)}/2$.

The first sum of probabilities in (26) is thus bounded by $n^{2(1-\alpha)}/2 \leq 1/2$ whenever $n \geq 1$.

For the second sum, we have $\mathbb{P}\big(B_{k^*,s,t} \leq \mu^*\big) \leq t^{-2\alpha}$ from the Chernoff-Hoeffding bound. Thus $\sum_{t=u+1}^{n}\sum_{s=1}^{t-1}\mathbb{P}\big(B_{k^*,s,t} \leq \mu^*\big) \leq \sum_{t=u+1}^{n} t^{1-2\alpha} \leq \int_u^\infty t^{1-2\alpha}dt = \frac{u^{-2(\alpha-1)}}{2(\alpha-1)}$ for $\alpha > 1$.

Thus (26) implies that $\mathbb{E}\left[T_k(n)\right] \leq \left(\frac{2b}{\Delta_k}\right)^2 \alpha\log(n) + \frac{3}{2} + \frac{1}{2(\alpha-1)}$ holds for all $n \geq 1$. The bound on the expected regret follows.

Theorem 8. *Let $\Delta_{\min} \triangleq \min_{k:\Delta_k>0}\Delta_k$. The pseudo-regret for UCB1 defined by (24) satisfies, for any $z \geq 4K\frac{b^2}{\Delta_{\min}}\alpha\log(n)$:*

$$\mathbb{P}\big(R_n > z\big) \leq \sum_{k:\Delta_k>0}\left\{e^{-z\frac{\Delta_k}{4Kb^2}} + z^{1-2\alpha}\frac{(K\Delta_k)^{2\alpha-1}}{2\alpha-1}\right\}. \tag{27}$$

Since the second term is dominant, we thus have $\mathbb{P}\big(R_n > z\big) = O(z^{1-2\alpha})$.

Proof. We have:

$$\begin{aligned}\mathbb{P}\big(R_n > z\big) &= \mathbb{P}\big(\sum_{k:\Delta_k>0}\Delta_k T_k(n) > z\big) \leq \mathbb{P}\big(\max_{k:\Delta_k>0}\Delta_k T_k(n) > z/K\big)\\ &\leq \sum_{k:\Delta_k>0}\mathbb{P}\big(T_k(n) > z/(K\Delta_k)\big)\end{aligned} \tag{28}$$

Notice that from the condition on z, we have $z/(K\Delta_k) \geq \left(\frac{2b}{\Delta_k}\right)^2\alpha\log(n)$. Define u_k to be the smallest integer larger than $z/(K\Delta_k)$. We start by providing an upper bound on $\mathbb{P}\big(T_k(n) > u_k\big)$.

Applying (10) with u_k and $\tau = \mu^*$, it comes that

$$\mathbb{P}\big(T_k(n) > u_k\big) \leq \sum_{t=u_k+1}^{n}\mathbb{P}\big(B_{k,u_k,t} > \mu^*\big) + \sum_{s=1}^{n-u_k}\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big). \tag{29}$$

From the condition on z, we have $b\sqrt{\frac{\alpha\log(n)}{u_k}} \leq \Delta_k/2$, thus $\mathbb{P}\big(B_{k,u_k,t} > \mu^*\big) \leq \mathbb{P}\big(\overline{X}_{k,u_k} > \mu_k + \Delta_k/2\big) \leq e^{-u_k\Delta_k^2/(2b^2)}$. Since $\alpha > 1$ we have $n \leq e^{u_k\left(\frac{\Delta_k}{2b}\right)^2}$. Thus the first sum of probabilities in (29) is bounded by $e^{-u_k(\Delta_k/(2b))^2}$.

Now, $\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big) \leq (u_k+s)^{-2\alpha}$ from the Chernoff-Hoeffding bound. Thus the second sum of probabilities in (29) is bounded by $\sum_{s=1}^{n-u_k}\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big) \leq \sum_{s=1}^{n-u_k}(u_k+s)^{-2\alpha} \leq \int_{u_k}^\infty t^{-2\alpha}dt = \frac{u_k^{1-2\alpha}}{2\alpha-1}$. Thus,

$$\mathbb{P}\big(T_k(n) > u_k\big) \leq e^{-u_k\left(\frac{\Delta_k}{2b}\right)^2} + \frac{u_k^{1-2\alpha}}{2\alpha-1}.$$

We deduce that $\mathbb{P}\big(T_k(n) > \frac{z}{K\Delta_k}\big) \leq e^{-z\frac{\Delta_k}{4Kb^2}} + z^{1-2\alpha}\frac{(K\Delta_k)^{2\alpha-1}}{2\alpha-1}$, and (27) follows from (28).

Theorem 8 shows that the regret has at least polynomial tails. In fact, this result cannot be improved to the extent that there exist distributions of the rewards for which for some constant $C > 0$, for any z large enough, $\mathbb{P}\big(R_n > z\big) \geq 1/(Cz^C)$. This can be proved by a simple adaptation of the arguments used in the proof of Theorem 10.

Theorems 7 and 8 show that the more we explore (i.e. larger α is), the smaller the tails of the regret is. However, the price of this extra exploration is a larger expected regret. In the next section, a similar tradeoff between expected rewards and risk is obtained for the UCB-V algorithm.

5 Risk bounds for UCB-V

In this section we concentrate on the analysis of the concentration properties of the pseudo-regret for UCB-V. As we will see in Remark 2 p.19, the concentration properties of the regret follow from the concentration properties of the pseudo-regret, hence there is no compromise in studying 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

$$\tilde{\beta}_n(t) \triangleq 3 \min_{\substack{\alpha\geq 1 \quad M\in\mathbb{N} \\ s_0=0<s_1<\cdots<s_M=n \\ \text{s.t.} \quad s_{j+1}\leq \alpha(s_j+1)}} \sum_{j=0}^{M-1} e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+t+1}}{\alpha}}. \tag{30}$$

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 complicated sum in (30), 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 is 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 9. *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

$$\mathbb{P}\big(R_n > r_0 x\big) \leq \sum_{k:\Delta_k>0} \left\{2ne^{-(c\vee 1)\mathcal{E}_n x} + \tilde{\beta}_n(\lfloor v_k\mathcal{E}_n x \rfloor)\right\}, \tag{31}$$

where we recall that $\tilde{\beta}_n(t)$ is essentially of order $e^{-(c\wedge 1)\mathcal{E}_t}$ (see (30)).

Proof. First note that

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

Let $\mathcal{E}'_n = (c \vee 1)\mathcal{E}_n$. We use (10) with $\tau = \mu^*$ and $u = \lfloor (1+v_k\mathcal{E}_n)x \rfloor \ge v_k\mathcal{E}_n x$. From (14), we have $\mathbb{P}(B_{k,u,t} > \mu^*) \le 2e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} \le 2e^{-\mathcal{E}'_n x}$. To bound the other probability in (10), we use $\alpha \ge 1$ and the grid $s_0, \dots, s_M$ of $\{1, \dots, n\}$ realizing the minimum of (30) when $t = u$. Let $I_j = \{s_j + 1, \dots, s_{j+1}\}$. Then

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

which gives the desired result.

When $\mathcal{E}_n \ge \log n$, the last term is the leading term. In particular, when $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$, Theorem 9 leads to the following corollary, which essentially says that for any $z > \gamma \log n$ with γ large enough,

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

for some constant $C > 0$:

Corollary 1. *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 $n \ge 2$ and any $z > \kappa_1 \log n$*

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

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 9 can be written as

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

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

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

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 $\tilde{\beta}_n(z')$ (see defintion in (30)), we consider a geometric grid of step $\alpha = 1/(1-\varepsilon)$, and cut the sum in $\tilde{\beta}_n$ in two parts: for the j's for which $s_j \le z'$, we use

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

whereas for the j's for which $s_j \le t$, $e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+z'+1}}{\alpha}} \le e^{-\frac{\mathcal{E}_{s_j}}{\alpha}} \le e^{-j\frac{\log\alpha}{\alpha}}$. 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 $\mathbb{P}(M_n > z) \le \exp(-2z^2/n)$. The slow concentration for UCB-V happens because the first $\Omega(\log(t))$ choices of the optimal arm can be unlucky (yielding small rewards) 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 whose decays only at a polynomial rate, which is controlled by ζ.[5] This reasoning relies crucially on that the choices of the optimal arm can be unlucky. Hence, we have the following result:

Theorem 10. *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.*

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 (10) 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 (10) vanish. The first term is controlled as in the proof of Theorem 9. 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 $x \ge 1$

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

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 $\mu_{\tilde{k}}$.

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

Let q be any positive integer. Consider the event:

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

Let $c_2 \triangleq 3bc\zeta$ and $\eta \triangleq \mu_{\tilde{k}} - \tilde{\mu}$. On this event, we have for any $t \leq e^{\eta q/c_2}$

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

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

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

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 $\Delta_{\tilde{k}}\big(e^{\eta q/c_2} - q\big)$. Now consider q large enough in order to have $e^{\eta q/c_2} - q \geq \frac{1}{2}e^{\eta q/c_2}$. Let $w > 0$ such that $e^{w^{-1}e^{\eta q/c_2}}$ is an integer larger than $e^{\eta q/c_2}$. Consider $n = e^{w^{-1}e^{\eta q/c_2}}$. We have

$$\mathbb{P}\big(R_n \geq \tfrac{\Delta_{\tilde{k}}}{2} w \log n\big) \geq \mathbb{P}(\Gamma) = \mathbb{P}(X_{1,1} = \tilde{\mu})^q = 1/(w \log n)^C$$

for some constant $C > 0$ depending only on c_2, η and $\mathbb{P}(X_{1,1} = \tilde{\mu})$. So the pseudo-regret cannot have thinner tails than polynomial ones.

Now in the general case, when $\tilde{\mu} < \mu_{\tilde{k}}$ but without specific reward distributions. One can also prove the regret has no thinner tails than polynomial ones. The proof is essentially the same but more cumbersome. For instance, instead of considering an event on which $X_{1,1} = X_{1,2} = \ldots = X_{1,q} = \tilde{\mu}$, we consider an event on which $X_{1,1}, X_{1,2}, \ldots, X_{1,q}$ are below μ'' with $\tilde{\mu} < \mu'' < \mu_{\tilde{k}}$, and on which, for the second optimal arm, the empirical means stay close to the associated expected mean $\mu_{\tilde{k}}$.

Remark 2. In Theorem 9 and Corollary 1, 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 formulae and assumptions which are as simple as possible. The following computations show that when the optimal arm is unique, one can obtain similar concentration 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 1, one can derive 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 $n - Cz$ times. Now, if the optimal arm is unique, then $n - Cz$ terms cancel out in sum defining the regret. For the Cz terms that remain, one can use standard Bernstein inequalities and union bounds to prove that with probability $1 - Cz^{-1}$, $\hat{R}_n \leq R_n + C'\sqrt{z}$ holds. Since the bound on the pseudo-regret is of order z (Corollary 1), a similar bound holds for the regret.

5.1 Illustration of the risk bounds

The purpose of this section is to illustrate the tail bounds obtained. For this we ran some computer experiments with bandits with two arms: the payoff of the optimal arm follows a Bernoulli distribution with expectation 0.5, while the payoff of the suboptimal arm is deterministic and assumes a value p which is slightly less than 0.5. This arrangement makes the job of the bandit algorithms very hard: All algorithms learn the value of the suboptimal arm quickly (although UCB1 will be very optimistic about this arm despite that all the payoffs received are the same). Since the difference of 0.5 and p is kept very small, it takes a lot of trials to identify the optimal arm and in particular to achieve this, the algorithm has to try the optimal arm many times. If the experiments start in an unlucky way, the algorithms will have the tendency to choose the suboptimal arm, further delaying the time of recognizing the true identity of the optimal arm. In all cases, $10,000$ independent runs were used to estimate the quantities of interest and the algorithms were run for $T = 2^{20} \approx 1,000,000$ time steps.

We have run experiments with both UCB1 and UCB-V. In the case of UCB1 the exploration coefficient, α (cf. Equation (24)), was chosen to take the value of 2, which can be considered as a typical choice. In the case of UCB-V we used $\zeta = 1$, $c = 1$, as a not too conservative choice (cf. Equation (16)). In both cases we set $b = 1$. For the considered bandit problems the difference between UCB1 and UCB-V is the result of that in the case of UCB-V the upper confidence value of the suboptimal arm will converge significantly faster to the true value than the same value computed by UCB1 since the estimated variances will always take the value of zero (the payoff is deterministic).

Fix $\alpha \geq 0$. Define the *value at risk* for the risk level α as the upper α-percentile of the regret:

$$R_n(\alpha) = \inf\{r : \mathbb{P}(R_n \geq r) \leq \alpha\}.$$

Hence, $R_n(\alpha)$ is a lower bound on the loss that might happen with α probability. Notice that the tail bounds of the previous section predict that the value at risk can be excessively large for difficult bandit problems. In particular, the more aggressive an algorithm is in optimizing the expected regret, the larger the value at risk is.

Figures 1 and 2 compare the estimated value at risk as a function of time for UCB1 and UCB-V for an easier ($p = 0.48$) and a more difficult problem ($p = 0.495$). Note that UCB-V, having tighter confidence intervals, can be considered as a more aggressive algorithm. For the figures the risk parameters were chosen to be $\alpha = 0.5$, 0.16 and 0.5 (the latter value corresponding to the median). These figures also show the mean regret. The figures also show the value at risk estimated by drawing 10,000 samples of a Gaussian distribution fitted to the regret distribution at each time step, for both algorithms. (These curves are labeled by pasting "(n)" after the α value.) If the regret is normally distributed, we can expect a good match across the value-at-risk measurements.

As expected, in the case of the "easy" problem UCB-V outperforms UCB1 by a large margin (which partially confirms the results on the scaling of the expected regret with the variance of the suboptimal arms). For UCB1 the distribution of

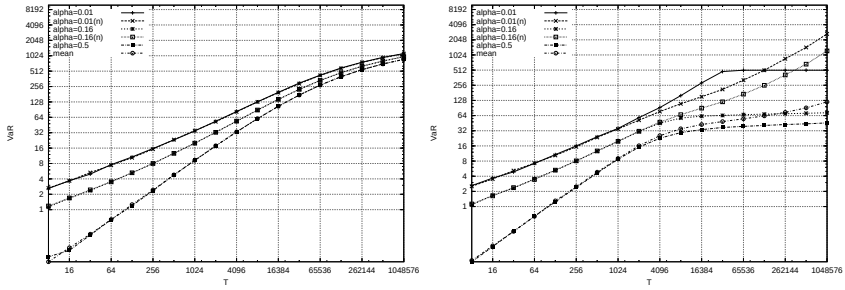

Fig. 1. Value at risk as a function of time when the expected payoff of the suboptimal arm is $p = 0.48$. The left-hand side (l.h.s.) figure depicts results for UCB1, while the right-hand side (r.h.s.) figure depicts results for UCB-V. Note the logarithmic scale of the time axis. For more details see the text.

regret is well approximated by a Gaussian at all time steps. In the case of UCB-V, we see that the Gaussian approximation overestimates the tail. Actually, in this case the regret distribution is bimodal (figures for the difficult problem will be shown later), but the r.h.s. mode has a very small mass (ca. 0.3% at the end of the experiment). Note that by the end of the experiment the expected regret of UCB-V is ca. 120, while the expected regret of UCB1 is ca. 870. This task is already quite challenging for both algorithms: They both have a hard time identifying the optimal arm. Looking at the distributions (not shown) of how many times the optimal arm is played, it turns out that UCB1 fails to shifts the vast majority of the probability mass to the optimal arm by the end of the experiment. At the same time, UCB-V shifts this happens at around $T = 8,192$. Note that in their initial (transient) phase both algorithms try both actions equally often (hence in the initial phase the expected regret grows linearly). The main difference is that UCB-V shrinks the confidence interval of the suboptimal arm much faster and hence eventually suffers a much smaller regret.

On the more challenging problem, the performance of UCB-V deteriorates considerably. Although the respective expected regrets of the algorithms are comparable (1213 and 1195, respectively, for UCB-V and UCB1), the value at risk for α close to zero for UCB-V is significantly larger than that for UCB1.

In order to illustrate what "goes wrong" with UCB-V we plotted the time evaluation of the proportion of time when the suboptimal arm is chosen as a function of time for 20 independent runs. That is, by introducing $T_{\text{bad}}(t) = \sum_{s=1}^{t} \mathbb{I}_{\{I_s \text{ is the bad arm}\}}$, the figure shows the time evolution of $T_{\text{bad}}(t)/t$ for 20 different runs. The result is shown in Figure 3. We see that quite a few runs tend to prefer the suboptimal arm as time goes by, although ultimately the curves for all runs converge towards 0.

In order to get a better picture of the distribution of picking the wrong arm, we plotted this distribution as a function of time (Figure 4). Note that at around time $T = 2,048$ ($\log_2(T) = 11$) the probability mass becomes bimodal. At this

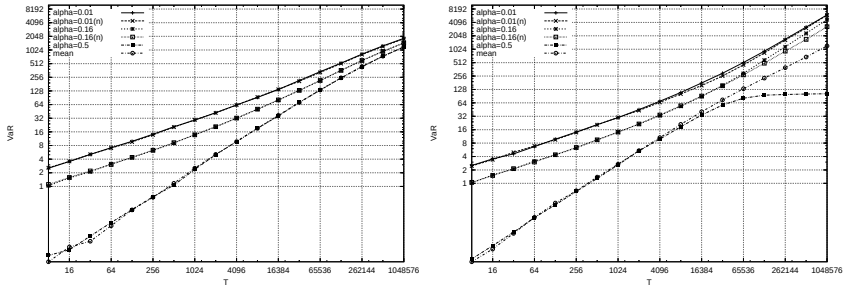

Fig. 2. Value at risk as a function of time when the expected payoff of the suboptimal arm is $p = 0.495$. The l.h.s. figure depicts results for UCB1, while the r.h.s. figure depicts results for UCB-V. For more details see the text.

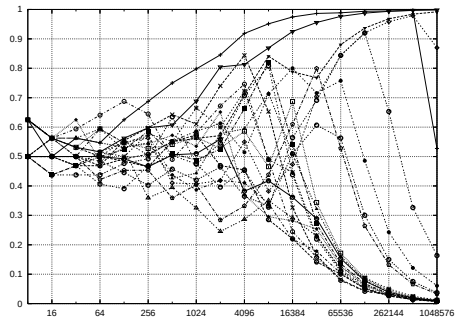

Fig. 3. $T_{\mathrm{bad}}(t)/t$, the proportion of time of using the suboptimal arm in the first t time-steps as a function of time for 20 independent runs. The bandit problem has parameter $p = 0.495$ and the algorithm is UCB-V.

time, a larger mass is shifted towards the (desired) mode with value 0, while a smaller, but still substantial mass is drifting towards 1. The mass of this second mode is continuously decreasing, although at a slow rate. The slow rate of this drift causes the large regret of UCB-V. A similar figure for UCB1 (not shown here) reveals that for UCB1 the distribution stays unimodal (up to the precision of estimation) and the mode starts to drift (slowly) towards 0 as late as at time $T = 2^{17}$.

In order to asses the rate of leakage of the probability mass from the right-side mode, we plotted the estimated probability of selecting the suboptimal arm more than α-fraction of the time (i.e., $\mathbb{P}(T_{\mathrm{bad}}(t) \geq \alpha t)$), as a function of time and for various values of α, see Figure 5. The figure reinforces that in the initial phase $T_{\mathrm{bad}}(t)$ is concentrated around $0.5t$. At the time when the two modes appear most of the mass starts to drift towards zero, though at the same time some mass is drifting towards t as indicated by the divergence of $\mathbb{P}(T_{\mathrm{bad}}(t) \geq \alpha t)$ away from zero. That all curves are converging to each other reveals that the

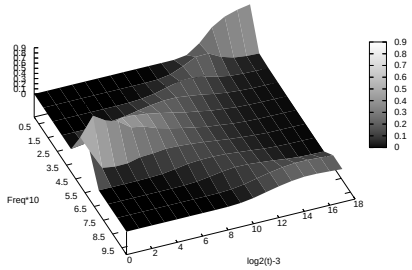

Fig. 4. The distribution of $T_{\mathrm{bad}}(t)/t$, the frequency of using the suboptimal arm, plotted against time. The bandit problem has parameter $p = 0.495$ and the algorithm is UCB-V.

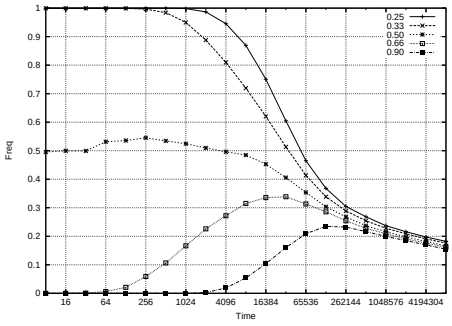

Fig. 5. The probability of choosing the suboptimal arm more than α-fraction of time plotted against time and various values of α. Note that the experiment was continued up to $T = 2^{24}$ steps to show the beginning of the asymptotic phase.

distribution becomes rather concentrated around the two modes, located at 0 and t. As the rate of convergence of the curves toward zero was hard to judge from the first $T = 2^{20}$ steps (the transient phase hardly ends by this time), we continued the experiment up to $T = 2^{24}$ time steps (the figure shows the results up to this time). Plotting the same figure on a log-log scale, it looks as if in the asymptotic phase these curves followed a polynomial curve.

To show that the regret also follows a bimodal distribution we plotted the histogram of the regret at times $T_1 = 16,384$ and $T_2 = 524,288$, shown on the left- and r.h.s. subfigures of Figure 6, respectively. The first time point, T_1, was selected so that the arm-choice distribution and hence also the regret distribution is still unimodal. However, already at this time the regret distribution looks heavy tailed on the right. By time T_2 the regret distribution is already bimodal, with a substantial mass belonging to the right-side mode (based on the previous figure, this mass is estimated to be about 25% of the total mass). Note that

the left-side mode is close to zero, while the right-side mode is close to $\Delta T_2 = 0.005 \times T_2 \approx 2,600$, confirming that runs contributing to either of the modes tend to stay with the mode from the very beginning of the experiments. Hence, the distribution of the regret appears to be of a mixture Gaussians.

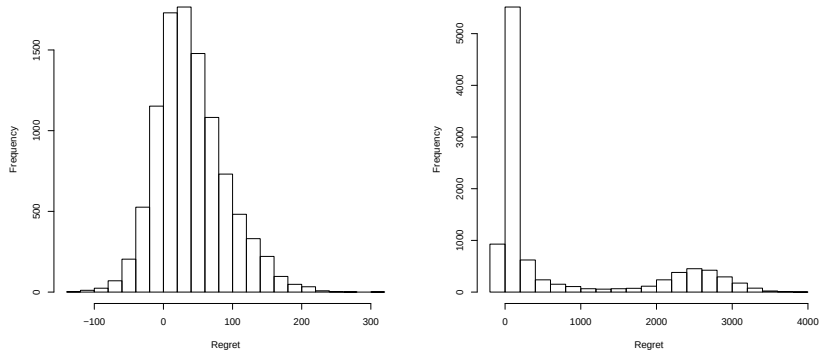

Fig. 6. Distribution of the regret for UCB-V at time $T_1 = 16,384$ (l.h.s. figure) and $T_2 = 524,288$ (r.h.s. figure). The bandit problem has parameter $p = 0.495$.

6 PAC-UCB

In this section, we consider the case when 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 an UCB algorithm which play any suboptimal arm only a few times, with high probability. Hence, the algorithm is "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 $B_{k,s}$. Besides, in order to simplify the discussion, we take $c = 1$.

Theorem 11. *Let $\beta \in (0,1)$. Consider a sequence $(\mathcal{E}_s)_{s\geq 0}$ satisfying $\mathcal{E}_s \geq 2$ and*

$$4K \sum_{s\geq 7} e^{-\mathcal{E}_s} \leq \beta. \tag{32}$$

Consider u_k the smallest integer such that

$$\frac{u_k}{\mathcal{E}_{u_k}} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k}. \tag{33}$$

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

Proof. See Section A.4.

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

$$\mathcal{E}_s = \log(Ks^q\beta^{-1}) \vee 2, \tag{34}$$

up to some additive constant ensuring that (32) holds. For this choice, Theorem 11 implies that for some positive constant κ, with probability at least $1-\beta$, for any suboptimal arm k (i.e., $\Delta_k > 0$), the number of plays is bounded by

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

Notice that this 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

$$\textstyle\sum_{k=1}^K T_k(n)\Delta_k \le \sum_{k:\Delta_k>0} \mathcal{T}_{k,\beta}\Delta_k. \tag{35}$$

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

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

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 12. *Let $n \ge 7$. 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 suboptimal arm k, the number of plays 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*

$$\mathbb{E}[R_n] \le \textstyle\sum_{k:\Delta_k>0} \big(\frac{24\sigma_k^2}{\Delta_k} + 30b\big)\log(n/3). \tag{36}$$

Proof. See Section A.5.

7 Open problem

When the time horizon 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 9, 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:

$$\mathbb{P}\Big\{R_n > 24\textstyle\sum_{k:\Delta_k>0}\Big(\frac{\sigma_k^2}{\Delta_k} + 2b\Big)\log n\Big\} \le \frac{C}{n}. \tag{37}$$

The behavior of this algorithm should be contrasted to the one with $\mathcal{E}_{s,t} = 3\log t$. 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 the exploitation mode. On the contrary, the algorithm which adapts to the time horizon explores and exploits at any time during the interval $[0; n]$. However, in view of Theorem 10, it satisfies only

$$\mathbb{P}\Big\{R_n > 24\sum_{k:\Delta_k>0}\Big(\frac{\sigma_k^2}{\Delta_k} + 2b\Big)\log n\Big\} \leq \frac{1}{C(\log n)^C}.$$

which is significantly worse than (37). 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 (37)? We conjecture that the answer is no.

A Proofs of the results

A.1 Lower bound for UCB1

Proposition 1. *There exists arm rewards in $[0, b]$ such that UCB1 (defined by the bias factor (1)) has an expected regret $\mathbb{E}[R_n] = \Omega(b^2\log(n))$.*

Proof. Consider the 2-armed deterministic bandit problem such that arm 1 yields the reward Δ, and arm 2 the reward 0. In order to obtain a lower bound on the regret, we derive a lower bound on $T_2(n)$.

First consider the "balance equation"

$$\Delta + b\sqrt{\frac{2\log(n+1)}{n-p(n)}} = b\sqrt{\frac{2\log(n+1)}{p(n)}}, \tag{38}$$

where $p(n)$ is considered as a function of $n \geq 1$. Note that solving (38) yields

$$p(n) = \frac{n}{2}\left[1 - \sqrt{1 - 4\Big(\frac{\sqrt{1 + n\Delta^2/(2b^2\log(n+1))} - 1}{n\Delta^2/(2b^2\log(n+1))}\Big)^2}\right].$$

Besides, we have the property that: $p(n) \geq \frac{2b^2}{\Delta^2}\log(n+1) - \frac{4\sqrt{2}b^3}{\Delta^3}\frac{(\log(n+1))^{3/2}}{\sqrt{n}}$, whose first term is dominant when n is large. Thus $p(n) = \Omega(\frac{b^2}{\Delta^2}\log(n+1))$

The intuition is that UCB1 works by keeping the upper bound $B_{1,T_1(n),n+1}$ of the first arm close to that of the second arm $B_{2,T_2(n),n+1}$ since the algorithm chooses at each time step the arm that has the highest bound, which decreases as a consequence its value. Thus we expect that $T_2(n)$ will be close to $p(n)$. For that purpose, let us prove the following result.

Lemma 1. *At any time step $n+1$, if UCB1 chooses arm 1 then we have $T_2(n) \geq p(n)$, otherwise we have $T_2(n) \leq p(n)$. We deduce that for all $n \geq 3$, $T_2(n) \geq p(n-1)$.*

Proof. The fist part of the lemma comes from the fact that if $T_2(n) < p(n)$, then $T_1(n) > n - p(n)$, thus

$$B_{2,T_2(n),n+1} = b\sqrt{\frac{2\log(n+1)}{T_2(n)}} > b\sqrt{\frac{2\log(n+1)}{p(n)}} = \Delta + b\sqrt{\frac{2\log(n+1)}{n-p(n)}}$$
$$> \Delta + b\sqrt{\frac{2\log(n+1)}{T_1(n)}} = B_{1,T_1(n),n+1},$$

which implies that arm 2 is chosen. A similar reasoning holds in the other case.

Now the second part of the lemma is proven by contradition. Assume there exists $n \geq 3$ such that $T_2(n) < p(n-1)$, and let n denote the first such time. Thus $T_2(n-1) \geq p(n-2)$ (note that this is also true if $n = 3$ since $T_2(2) = 1$ and $p(1) \leq 1/2$). Thus $T_2(n-1) \leq T_2(n) < p(n-1)$ which, from the first part of the proposition, implies that at time n, arm 2 is chosen. We deduce that

$$p(n-1) > T_2(n) = T_2(n-1) + 1 \geq p(n-2) + 1.$$

This is impossible since the function $x \to p(x)$ has a slope bounded by $1/2$ in the domain $[1, \infty)$, thus $p(n-1) \leq p(n-2) + 1/2$.

From the previous lemma, we deduce that $T_2(n) = \Omega(\frac{b^2}{\Delta^2}\log(n))$ and thus the regret $R_n = T_2(n)\Delta = \Omega(\frac{b^2}{\Delta}\log(n))$, which proves the proposition.

A.2 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 2. *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 $1 \leq t \leq n$,*

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{E}[U^2]\, x} + b''_+ x/3, \tag{39}$$

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

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{V}\mathrm{ar}(U)\, x} + b' x/3. \tag{40}$$

Proof (Proof of Lemma 2). Let $v = (\mathbb{V}\mathrm{ar}\, U)/(b')^2$. To prove this inequality, we use Result (1.6) of Freedman (Freedman, 1975) 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 (40), it remains to check that

$$nvh[a/(nv)] \geq x. \tag{41}$$

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 (40) is finished since $\varphi(\sqrt{2x/(nv)}) \geq 0$ implies (41).

To prove (39), 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., (Audibert, 2004, 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 (39) when $b'' \leq 0$. When $b'' > 0$, the minimum is reached for $\lambda b'' = \log\left(\frac{b''a + n\mathbb{E}[U^2]}{n\mathbb{E}[U^2]}\right)$ and then the computations are similar to the one developed to obtain (40).

Remark 3. Lemma 2 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 $w > 0$,

$$\mathbb{P}\big(\overline{U}_n - \mu > w\big) \leq e^{-\frac{nw^2}{2\mathbb{V}\mathrm{ar}(U) + (2b'w)/3}}. \tag{42}$$

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

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

Compared with (40) 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 2 allows the time indices to vary in an interval. This form follows from a martingale's argument due to Freedman (Freedman, 1975).

Given Lemma 2, 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 $0 \leq t \leq n$,

$$|\overline{X}_t - \mu| < \tfrac{\sqrt{2nV_t x}}{t} + \tfrac{3bnx}{t^2}. \tag{43}$$

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 2, (40) with the choices $U_i = X_i$, $U_i = -X_i$, and Lemma 2, (39) 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

$$|\overline{X}_t - \mu| \leq \sigma \tfrac{\sqrt{2nx}}{t} + \tfrac{x}{3t} \tag{44}$$

and

$$\sigma^2 \leq V_t + (\mu - \overline{X}_t)^2 + \tfrac{\sqrt{2n\mathcal{V}\, x}}{t}. \tag{45}$$

Let $L \triangleq nx/t^2$. We claim that from (44) and (45), it follows that

$$\sigma \leq \sqrt{V_t} + 1.8\sqrt{L}. \tag{46}$$

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 (46) 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 (44) into (45) we obtain for any $0 \leq t \leq n$

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

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 (46). Plugging (46) into (44), we obtain

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

which, given the definition of L, ends the proof of (43), 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 (43), with probability at least $1 - 3e^{-x/\alpha}$, for $t \in \{t_1, \ldots, t_2\}$, we have

$$\begin{aligned}
t|\overline{X}_t - \mu| &< \sqrt{2t_2 V_t x/\alpha} + 3x/\alpha \\
&\leq \sqrt{2tV_t x} + 3x.
\end{aligned} \tag{47}$$

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 (47) is trivial for $t < 4$. Precisely, introduce

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

and let $s_0, \ldots, s_M$ 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 $\{3, 4, \ldots, n\}$. This ends the proof of Theorem 1.

A.3 Proof of Theorem 6

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

$$\frac{\log(1/(1-\varepsilon))}{\varepsilon} < \frac{1-3\delta}{1-6\delta}. \tag{48}$$

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,

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

Let t_1 be the time t when arm one is pulled the T-th time. If $t_1 \geq n$ 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 (48). 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.4 Proof of Theorem 11

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

$$\forall s \geq 7 \quad \forall k \in \{1, \ldots, K\} \quad \begin{cases} \left|\overline{X}_{k,s} - \mu_k\right| < \sigma_k \sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{\mathcal{E}_s}{3s} \\ \sigma_k \leq \sqrt{V_{k,s}} + 1.8\sqrt{\frac{\mathcal{E}_s}{s}} \\ \sqrt{V_{k,s}} \leq \sigma_k + \sqrt{\frac{\mathcal{E}_s}{2s}} \end{cases} \tag{49}$$

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 2 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

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

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

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

As we have seen in Section A.2, the above first two inequalities give the first two inequalities of (49). Finally, taking the square root in the above third inequality gives the last inequality of (49).

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

$$B_{k,s} \leq \mu_k + 2\sigma_k \sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{13\mathcal{E}_s}{3s} \tag{50}$$

Proof. The inequality $\mu_k \leq B_{k,s}$ is obtained by plugging the second inequality of (49) in the first one of (49) 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 (50), we used the first and third inequalities of (49) to obtain

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

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

■

Now let us prove that the choice of u_k in Theorem 11 guarantees that

$$\mu_k + 2\sigma_k \sqrt{\frac{2\mathcal{E}_{u_k}}{u_k}} + \frac{13\mathcal{E}_{u_k}}{3u_k} < \mu^*. \tag{51}$$

Proof. For the sake of lightening the notation, let us drop for a moment the k indices, so that (51) is equivalent to

$$2\sigma \sqrt{\frac{2\mathcal{E}_u}{u}} + \frac{13\mathcal{E}_u}{3u} < \Delta. \tag{52}$$

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

$$\begin{aligned} (52) &\Leftrightarrow r - \tfrac{13}{3\Delta} > \tfrac{2\sigma}{\Delta}\sqrt{2r} \\ &\Leftrightarrow r > \tfrac{13}{3\Delta} \quad \text{and} \quad \left(r - \tfrac{13}{3\Delta}\right)^2 > \tfrac{8\sigma^2}{\Delta^2} r \\ &\Leftrightarrow r > \tfrac{13}{3\Delta} \quad \text{and} \quad r^2 - \left(\tfrac{8\sigma^2}{\Delta^2} + \tfrac{26}{3\Delta}\right) r + \tfrac{169}{9\Delta^2} > 0 \end{aligned}$$

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

By adapting the argument leading to (11), 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 (51),we get

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

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.5 Proof of Theorem 12

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 11 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

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

hence

$$T_k(n) \leq 1 + \big(\tfrac{8\sigma_k^2}{\Delta_k^2} + \tfrac{26b}{3\Delta_k}\big)\log[KnT_k(n)], \tag{53}$$

which gives the first assertion.

For the second assertion, first note that since $R_n \leq n$, (36) 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 $K < n/50$, (53) gives

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

hence

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

Bibliography

Agrawal, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, *27*, 1054–1078.

Audibert, J.-Y. (2004). *PAC-bayesian statistical learning theory.* Doctoral dissertation, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. http://cermics.enpc.fr/ audibert/ThesePack.zip.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Auer, P., Cesa-Bianchi, N., & Shawe-Taylor, J. (2006). Exploration versus exploitation challenge. *2nd PASCAL Challenges Workshop.*

Freedman, D. (1975). On tail probabilities for martingales. *The Annals of Probability*, *3*, 100–118.

Gelly, S., & Silver, D. (2007). Combining online and offline knowledge in UCT. *International Conference on Machine Learning (ICML).*

Gittins, J. C. (1989). *Multi-armed bandit allocation indices.* Wiley-Interscience series in systems and optimization. Chichester, NY: Wiley.

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. *Proceedings of the 17th European Conference on Machine Learning (ECML-2006)* (pp. 282–293).

Lai, T., & Yakowitz, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, *40*, 1199–1209.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*, 4–22.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, *58*, 527–535.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.