Exploration-exploitation trade-off using variance estimates in multi-armed bandits

Jean Yves Audibert*

Université Paris-Est, Ecole des Ponts ParisTech, CERTIS
6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France
&
Willow - ENS / INRIA
45 rue d'Ulm, 75005 Paris, France

Rémi Munos

INRIA Lille - Nord Europe, SequeL project,
40 avenue Halley, 59650 Villeneuve d'Ascq, France

Csaba Szepesvári*,[1]

Department of Computing Science
University of Alberta
Edmonton T6G 2E8, Canada

Abstract

Algorithms based on upper confidence bounds for balancing exploration and exploitation are gaining popularity since they are easy to implement, efficient and effective. 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. We provide the first analysis of the expected regret for such algorithms. As expected, our results show that the algorithm that uses the variance estimates has a major advantage over its alternatives that do not use such estimates provided that the variances of the payoffs of the suboptimal arms are low. We also prove 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 special ones where with probability one the payoff obtained by pulling the optimal arm is larger than the expected payoff for the second best arm. Hence, although upper confidence bound bandit algorithms achieve logarithmic expected regret rates,

*Corresponding author.
Email addresses: audibert@certis.enpc.fr (Jean Yves Audibert), remi.munos@inria.fr (Rémi Munos), szepesva@cs.ualberta.ca (Csaba Szepesvári)
[1]Csaba Szepesvári is on leave from MTA SZTAKI, Budapest, Hungary.

they might not be suitable for a risk-averse decision maker. We illustrate some of the results by computer simulations.

Key words: exploration-exploitation tradeoff, multi-armed bandits, Bernstein inequality, high probability bound, risk analysis

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 test ("explore") some alternative, hoping to discover a choice that beats the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially [13]. These bandit problems became popular with the seminal paper of Robbins [12], 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 real-valued distributions, $\nu_1, \ldots, \nu_K$. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the *regret* which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, i.e., when the arm giving the highest expected reward is pulled all the time.

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 produces a random payoff as a result: When arm k is pulled the random payoff is drawn from ν_k. The payoffs are 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, to play well the gambler must carefully balance *exploration and exploitation*.

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

Let us state the goal of this design problem formally. 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. The expected payoff of such an arm is the optimal expected reward: $\mu^* = \max_{1 \le k \le K} \mu_k$. Let $T_k(t)$ denote the number of times arm k is chosen by the policy during the first t plays and let $I_t \in 1, \ldots, K$ be the index of the arm played at time t. The *(cumulative) regret of the gambler's strategy up to time* n is defined by

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

where k^* is the index of an optimal arm (when multiple optimal arms exist we pick one such arm arbitrarily). The goal is to design a policy whose *expected (cumulative) regret*, $\mathbb{E}[\hat{R}_n]$, is as small as possible. (Clearly, this is equivalent to maximizing the total expected reward achieved up to time n.) Wald's equation implies that the expected regret satisfies

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

where $\Delta_k = \mu^* - \mu_k$ is the expected loss of playing arm k. Hence, a policy 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 [7]). Lai and Robbins [10] 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". Their 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 [1] has shown how to construct upper confidence bound algorithms that use the sample-means of the arms. More recently, Auer et al. [3] considered the non-parametric case when all the knowledge the decision maker has is that the rewards have bounded range, say they belong to $[0, b]$. 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. They proved that the expected regret of this algorithm satisfies

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

In the same paper they proposed UCB1-NORMAL, a policy specialized to the case when the payoffs are normally distributed with unknown mean and variance. This algorithm estimates the arms' variances to refine the bias factor. Under the normality assumption they show that

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

where σ_k^2 denotes the variance of the k^{th} arm.

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 UCB1's regret-bound with b is not a proof artifact: The expected regret indeed scales with[2] $\Omega(b^2)$ (see Proposition 1, Section A.2). Since in many cases b is 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. [3] 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. [3] indicate that the idea of using empirical variance estimates works: UCB1-Tuned outperformed the other algorithms in essentially all the experiments. The superiority of this algorithm has been reconfirmed recently in the latest Pascal Challenge [4]. Intuitively, algorithms using variance estimates should work better than 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 then a "variance-aware" algorithm can spot the suboptimal arms much faster, thereby reducing the regret suffered.

One purpose of this paper is to study such "variance-aware" algorithms. For this we study the regret of *UCB-V*, which is a generic UCB-type algorithm that use variance estimates in its 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 the so-called *exploration function*. A typical choice for this function is $\mathcal{E}_{s,t} = \zeta \log(t)$. With this choice the algorithm's behavior is controlled by the parameters $\zeta, c > 0$.

[2]Through the paper, we will use the Landau notation: $\Omega(g)$ is a term asymptotically bounded below by g up to constant factor, and $\Theta(g)$ is a term asymptotically bounded below and above by g (up to constant factors).

Our first major contribution is a bound on the expected regret of UCB-V with this choice of the exploration function that scales in an improved fashion with b. In particular, in Theorem 4 we show that for $c = 1$ and $\zeta = 1.2$,

$$\mathbb{E}[\hat{R}_n] \leq 10 \sum_{k:\mu_k<\mu^*} \left(\frac{\sigma_k^2}{\Delta_k} + 2b \right) \log(n). \tag{4}$$

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, a major difference to the bound (3). Although, this is unfortunate, it is possible to show that the dependence on b is unavoidable (see Section A.1).

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 the result of Auer et al. [3], our regret bound also relies on the analysis of the sampling times of suboptimal arms (Theorem 2). Compared to the analysis by Auer et al. [3], 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 such that if the tuning parameters are below these limits the loss goes up considerably (Theorems 5 and 6).

The second major contribution of the paper is the 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. We think that the concentration of regret results obtained can be important in the analysis of algorithms that nest sequences of bandits, such as the UCT algorithm proposed by Kocsis and Szepesvári [9], which recently was proven to be very efficient in computer go (e.g., Gelly et al. [6]).

In order to analyze the risk, we 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:[3]

$$\mathbb{E}[R_n] = \mathbb{E}[\hat{R}_n], \tag{5}$$

but the randomness of the rewards influences the pseudo-regret only indirectly (i.e., only through $\{T_k(n)\}$). In order to analyze the risk, in Sections 5.2 and 7

[3]This is a standard result that can be shown using Wald's identity exploiting that the rewards coming from different arms are independent.

we develop high-probability bounds for the pseudo-regret. Similar results can be obtained for the cumulative regret (see Remark 2 p.23).

Interestingly, this analysis revealed the following unexpected tradeoff: If one aims for logarithmic expected regret (or, more generally, for subpolynomial regret) then the regret will not concentrate exponentially fast around its mean when with positive probability the optimal arm yields rewards smaller than some suboptimal arm's expected reward (Theorem 10). In order to explain what happens let us consider the case of two arms that satisfy this condition. Assume that the first arm is 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, the first at $\Theta(\log n)$ and the second at $\Omega(\Theta_2 n)$. The second mode corresponds to the case when the algorithm starts in an unlucky manner in the sense that the rewards obtained when testing the first (optimal) arm are all small in an initial phase. In this case the algorithm may get stuck with the suboptimal arm for a long time. Hence, the probability mass associated with the second mode will decay only polynomially with n and the decay-rate will depend on Δ_2. (The probability that the regret is above a threshold larger than the second mode decays exponentially.) The decay rate of the mass in the second mode can be increased by increasing exploration rate. However, then the expected regret will increase. Our regret tail bound (Theorem 9) makes the dependence on the algorithm's parameters explicit in this tradeoff. The theoretical findings of this part are illustrated in a series of experiments which are described in Section 6.

In the final part of the paper (Section 7) 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. Thus, knowing the horizon decreases the risk significantly. We conjecture that the knowledge of the time horizon indeed represents a significant advantage in this sense.

2. Notation

We let $\lfloor x \rfloor$ denote the largest integer smaller or equal to $x \in \mathbb{R}$ and let $\lceil x \rceil$ denote the smallest integer larger than x. Further, for u, v reals, $u \wedge v$ ($u \vee v$) denotes the minimum (resp., maximum) of u and v.

3. 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 $\overline{X}_{k,t}$ (resp.,

$V_{k,t}$) be the empirical estimate of the expected payoff (resp., 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 k^* is the index of an *optimal arm*:

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

In the paper we will use the convention that quantities related to the optimal arm will be denoted by putting $*$ in the upper index.

In the following, we assume that the rewards are bounded. In particular, we make the simplifying assumption that all the rewards are almost surely in $[0, b]$ for some $b > 0$ known to the decision maker. (We loose generality only because we assume that the bound b is the same for all the arms. However, our results can be easily generalized to the case when these bounds differ between the arms.) For easy reference we summarize our assumptions on the reward sequence here:

Assumption A1 Let $K > 2$ and let $\nu_1, \dots, \nu_K$ be distributions over the 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, i.e., for any $t \ge 1$, the vectors $(X_{1,1}, \dots, X_{1,t}), \dots, (X_{K,1}, \dots, X_{K,t})$ are independent. The decision maker does not know the distributions of the arms, but knows b.

3.1. 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

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\,\mathcal{E}_{s,t}}{s}} + c\frac{3b\,\mathcal{E}_{s,t}}{s} \tag{6}$$

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

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

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). After this initial phase the arms will

[4] The i.i.d. assumption can be relaxed, see e.g., Lai and Yakowitz [11].

be tried multiple times. The more an arm k has been tested, the closer the bound (6) 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 arms having the largest expected rewards with an increasing frequency.

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. Hence an algorithm that only uses sample-means might get stuck with not choosing the optimal arm any more. UCB policies (in general) prevent this situation by using upper confidence bounds on the mean rewards. The confidence level with which these bounds hold determine the amount of exploration of the policy and ultimately the performance of the algorithm.

The UCB-V policy uses the function, $\mathcal{E}$, to facilitate exploration. Indeed, assuming that for any fixed s, $\mathcal{E}_{s,t}$ increases without bounds in t, we see that if an arm is not tried for a long time then after a while the last term of (6) will start to dominate the other terms and will also dominate the bound associated with the arms drawn very often. This will then allow the algorithm to draw this arm again and thus the algorithm will have a chance to develop a better estimate of the arm's expected payoff. In particular, this holds for all the optimal arms, too and will allow the algorithm to recover even when the optimal arm(s) start in an unlucky way. 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 and as a result 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 k can only be chosen if its confidence bound is larger than Δ_k, its expected payoff loss. Since the 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 quantity $B_{k,s,t}$ comes from a novel tail bound on the sample average of i.i.d. random variables with bounded support. Unlike previous similar bounds (e.g., based on Bennett's and Bernstein's inequalities) that used the true (but unknown) variance our bound uses the empirical variance. The bound relies on the exponential concentration of the empirical variance around the true 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 $1 - 3e^{-x}$,

$$|\overline{X}_t - \mu| \leq \sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}. \tag{7}$$

Furthermore, introducing

$$\beta(x,t) = 3 \inf_{1<\alpha\leq 3} \Big(\frac{\log t}{\log \alpha} \wedge t \Big) e^{-x/\alpha}, \tag{8}$$

where $u \wedge v$ denotes the minimum of u and v, we have for any $t \in \mathbb{N}$ and $x > 0$, with probability at least $1 - \beta(x,t)$

$$|\overline{X}_s - \mu| \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s} \tag{9}$$

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

Proof. See Section A.3.

Remark 1. The uniformity in time is the only difference between the two assertions of the previous theorem. When we use (9), 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 (7) that is uniform in time. In particular, this means that if $1 \leq S \leq t$ is an integer-valued random variable then (9) still holds with probability at least $1 - \beta(x,t)$ and when in (9) s is replaced with S.

Note that (7) 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.

There are essentially two types of exploration functions leading to interesting properties of the resulting algorithms in terms of expected regret, high-probability bounds on the regret and tunability with respect to the total number of plays:

- the ones in which $\mathcal{E}_{s,t}$ depends only on t (see Sections 4 and 5.2).
- the ones in which $\mathcal{E}_{s,t}$ depends only on s (see Section 7).

To understand why we do not consider $\mathcal{E}_{s,t}$ depending on both s and t, recall that in $\mathcal{E}_{s,t}$ variable s plays the role of the number of pulls of an arm. Hence we always have $s < t$. Further, for suboptimal arms we will hopefully have $s \ll t$. Normally, the contribution of s to the exploration function should be in the same order as the contribution of t. Thus, when $\mathcal{E}_{s,t}$ already depends on t, the dependence on s will not alter the behavior (and hence the performance) of the algorithm in a significant way.

3.2. 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 the suboptimal arms are drawn. In this section we derive such bounds, generalizing and improving upon the previous analysis of Auer et al. [3]. The improvement is a necessary step to get *tight* bounds in the case when the exploration function scales logarithmically with t, i.e., for the class of most interesting 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. *The followings hold:* (i) *After K plays, each arm has been pulled once.* (ii) *Pick an arm k and a time $n \in \mathbb{N}^+$. For any $\tau \in \mathbb{R}$ and any integer $u > 1$, it holds that*

$$T_k(n) \le u + \sum_{t=u+K-1}^{n} \Big(\mathbb{1}_{\{\exists s : u \le s \le t-1 \text{ s.t. } B_{k,s,t} > \tau\}} + \mathbb{1}_{\{\exists s^* : 1 \le s^* \le t-1 \text{ s.t. } \tau \ge B_{k^*,s^*,t}\}} \Big). \tag{10}$$

Hence, also

$$\begin{aligned} \mathbb{E}\left[T_k(n)\right] \quad \le \quad & u + \sum_{t=u+K-1}^{n} \sum_{s=u}^{t-1} \mathbb{P}\big(B_{k,s,t} > \tau\big) \\ & + \sum_{t=u+K-1}^{n} \mathbb{P}\big(\exists s : 1 \le s \le t-1 \text{ s.t. } B_{k^*,s,t} \le \tau\big). \end{aligned} \tag{11}$$

Further, it holds that

$$\begin{aligned} \mathbb{P}\big(T_k(n) > u\big) \quad \le \quad & \sum_{t=u+1}^{n} \mathbb{P}\big(B_{k,u,t} > \tau\big) \\ & + \mathbb{P}\big(\exists s : 1 \le s \le n-u \text{ s.t. } B_{k^*,s,u+s} \le \tau\big). \end{aligned} \tag{12}$$

Note that even though the above statements hold for any arm, the bounds are trivial for the optimal arms. Besides, (10) and (11) hold independently of the form of the quantity $B_{k,s,t}$.

Proof. Part *(i)* 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.

Let us thus turn to the proof of Part *(ii)*. To obtain (10), 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

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

Putting these inequalities together proves (10). Taking the expectation of both sides of (10) and using a union bound, we obtain (11).

Finally, inequality (12) comes from a direct argument that uses that the exploration function $\mathcal{E}_{s,t}$ is a nondecreasing function with respect to t: In order to prove this inequality consider an event such that the following statements hold:

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

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

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

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

$$\begin{aligned} \big\{T_k(n) > u\big\} \subset \Big(&\big\{\exists t : u+1 \le t \le n \text{ s.t. } B_{k,u,t} > \tau\big\} \\ &\cup \big\{\exists s : 1 \le s \le n-u \text{ s.t. } B_{k^*,s,u+s} \le \tau\big\}\Big). \end{aligned} \tag{13}$$

By taking probabilities of both sides and using a union-bound argument, we get the announced result. □

4. The expected regret of UCB-V

In this section, we assume that the exploration function does not depend on s (still, $\mathcal{E} = (\mathcal{E}_t)_{t\ge 0}$ 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 the constant c in $B_{k,s,t}$ should not be taken too small if one does not want to suffer a polynomial regret instead of a logarithmic one. We will derive bounds on the expected regret and conclude by specifying natural constraints on c and $\mathcal{E}_t$.

4.1. Upper bounds on the expected regret

Theorem 3. *We have*

$$\begin{aligned} \mathbb{E}[R_n] \quad \le \quad \sum_{k:\Delta_k>0} \Bigg\{ &1 + 8(c \vee 1)\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right)\mathcal{E}_n \\ &+ n e^{-\mathcal{E}_n}\left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) + \sum_{t=16\mathcal{E}_n}^{n} \beta\big((c\wedge 1)\mathcal{E}_t, t\big)\Bigg\}\Delta_k, \end{aligned} \tag{14}$$

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 (8) and Remark 1).

Note that by (5) the theorem gives a bound on the expected regret, $\mathbb{E}\left[\hat{R}_n\right]$.

We need the following Lemma that will be useful later, too:

Lemma 1. *Let* $u = \left\lceil 8(c \vee 1)\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right)\mathcal{E}_n \right\rceil$. *Then for any* s, t *such that* $u \le s \le t \le n$, $t \ge 2$, *it holds that*

$$\mathbb{P}(B_{k,s,t} > \mu^*) \le 2e^{-s\Delta_k^2/(8\sigma_k^2 + 4b\Delta_k/3)}. \tag{15}$$

Note that for any suboptimal arm k the probability decays exponentially in s for s large enough, independently of the value of t and n. Intuitively, this makes sense as the main term in $B_{k,s,t}$ is $\overline{X}_{k,s}$, which estimates $\mu_k < \mu^*$.

Proof (of Lemma 1). From the definition of $B_{k,s,t}$ (cf. (6)) we obtain

$$\begin{aligned}
&\mathbb{P}(B_{k,s,t} > \mu^*) \\
&\quad \le \mathbb{P}\Big(\overline{X}_{k,s} + \sqrt{\tfrac{2V_{k,s}\mathcal{E}_t}{s}} + 3bc\tfrac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\Big) \\
&\quad \le \mathbb{P}\Big(\overline{X}_{k,s} + \sqrt{\tfrac{2[\sigma_k^2 + b\Delta_k/2]\mathcal{E}_t}{s}} + 3bc\tfrac{\mathcal{E}_t}{s} > \mu_k + \Delta_k\Big) + \mathbb{P}\big(V_{k,s} \ge \sigma_k^2 + b\Delta_k/2\big).
\end{aligned}$$

In order to bound the second term note that $V_{k,s} = 1/s\sum_{j=1}^{s}(X_{k,j} - \mu_k)^2 - (\mu_k - \overline{X}_{k,s})^2$, hence $\mathbb{P}\big(V_{k,s} \ge \sigma_k^2 + b\Delta_k/2\big) \le \mathbb{P}\big(\frac{\sum_{j=1}^{s}(X_{k,j}-\mu_k)^2}{s} - \sigma_k^2 \ge b\Delta_k/2\big)$. Let $\mathcal{E}_n' = (c \vee 1)\mathcal{E}_n$. In order to bound the first term note that since $u \le s$, $t \le n$ and thanks to the choice of u we have

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

where the last inequality holds as it is equivalent to $(x-1)^2 \ge 0$ with $x = \sqrt{\frac{2\sigma_k^2 + b\Delta_k}{2\sigma_k^2 + 4b\Delta_k}}$. Hence,

$$\begin{aligned}
&\mathbb{P}(B_{k,s,t} > \mu^*) \\
&\quad \le \mathbb{P}\big(\overline{X}_{k,s} - \mu_k > \Delta_k/2\big) + \mathbb{P}\Big(\tfrac{\sum_{j=1}^{s}(X_{k,j}-\mu_k)^2}{s} - \sigma_k^2 \ge b\Delta_k/2\Big) \\
&\quad \le 2e^{-s\Delta_k^2/(8\sigma_k^2 + 4b\Delta_k/3)},
\end{aligned}$$

where in the last step we used Bernstein's inequality (see (46)) twice. □

Proof (of Theorem 3). Because $R_n = \sum_k \Delta_k T_k(n)$ it suffices to bound $\mathbb{E}[T_k(n)]$, where k is the index of a suboptimal arm. Thus, pick such an index k. We use (11) to bound $\mathbb{E}[T_k(n)]$ with $\tau = \mu^*$ and $u = \left\lceil 8\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k}\right)\mathcal{E}_n' \right\rceil$ with $\mathcal{E}_n' = (c \vee 1)\mathcal{E}_n$, as in Lemma 1:

$$\mathbb{E}[T_k(n)] \;\le\; u + \sum_{t=u+1}^{n}\sum_{s=u}^{t-1}\mathbb{P}\big(B_{k,s,t} > \mu^*\big) + \sum_{t=u+1}^{n}\sum_{s=1}^{t-1}\mathbb{P}\big(B_{k^*,s,t} \le \mu^*\big). \tag{16}$$

Via the help of Lemma 1, the inner sum of the first double sum is bounded as follows:

$$\sum_{s=u}^{t-1} \mathbb{P}(B_{k,s,t} > \mu^*) \le 2 \sum_{s=u}^{\infty} e^{-s\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} = 2\frac{e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}{1-e^{-\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)}}$$
$$\le \left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} \le \left(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{4b}{\Delta_k}\right) e^{-\mathcal{E}_n'}. \quad (17)$$

Here we have used that $1-e^{-x} \ge 2x/3$ that holds when $0 \le x \le 3/4$. The other term of (16) is bounded by using the uniform, empirical variance-estimate-based deviation bound (9) of Theorem 1. Putting the so obtained bounds together we get

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

This gives the announced result since by assumption $u \ge 16\mathcal{E}_n$. □

In order to balance the terms in (14) the exploration function should be chosen to be proportional to $\log t$, yielding the following upper estimate of the payoff of arm k provided that this arm was chosen s times up to time t:

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2\zeta V_{k,s} \log t}{s}} + c\frac{3b\zeta \log t}{s}. \quad (18)$$

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 that depends on ζ only such that for $n \ge 2$*

$$\mathbb{E}[R_n] \quad \le \quad c_\zeta \sum_{k:\Delta_k>0} \left(\frac{\sigma_k^2}{\Delta_k} + 2b\right) \log n. \quad (19)$$

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

Proof. Inequality (19) follows directly from Theorem 3 once we bound the four terms between the brackets in (14). To obtain the logarithmic regret, the third term of (14) requires $\zeta \ge 1$ while the fourth term requires $\zeta > 1$.

The proof of the numerical assertion is tedious. First it uses that

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

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 of (14) using $1 \le \frac{\log n}{\log 91} \le a_1 \frac{2b}{\Delta_k}(\log n)$, with $a_1 = 1/(2\log 91) \approx 0.11$.

The second term between the brackets in (14) is bounded by $a_2\left(\frac{\sigma_k^2}{\Delta_k^2}+\frac{2b}{\Delta_k}\right)\log n$, with $a_2 = 8\times 1.2 = 9.6$. For the third term, we use that for $n > 91$, we have $24n^{-0.2} < a_3 \log n$, with $a_3 = \frac{24}{91^{0.2}\times\log 91} \approx 0.21$. By tedious computations, the fourth term can be bounded by $a_4\frac{2b}{\Delta_k}(\log n)$, with $a_4 \approx 0.07$. This gives the desired result since $a_1 + a_2 + a_3 + a_4 \leq 10$. □

As promised, 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).

4.2. 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:

- The distribution of the rewards of arm 1 is concentrated on 0 and 1 with equal probabilities.
- The other arms provide a reward equal to $\frac{1}{2} - \varepsilon_n$ deterministically.

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

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

$$B_{1,s,t} = \overline{X}_{1,s} + \sqrt{\frac{2V_{1,s}\zeta\log t}{s}} + \tilde{b}\frac{\log t}{s} \leq \frac{1}{2} + \left(\overline{X}_{1,s} - \frac{1}{2}\right) + \sqrt{\frac{\zeta\log n}{2s}} + \tilde{b}\frac{\log n}{s}. \tag{20}$$

On the other hand, for any $0 \leq \tilde{s} \leq t$ and arm $k > 1$, we have

$$B_{k,\tilde{s},t} = \frac{1}{2} - \varepsilon_n + \tilde{b}\frac{\log t}{\tilde{s}} \geq \frac{1}{2} - \varepsilon_n. \tag{21}$$

So the algorithm will continue to choose arm 1, i.e., will behave badly, as long as for some $s < n$, we have $B_{1,s,t} < 1/2 - \varepsilon_n$. Now, we will choose ε_n and s so that this happens with a non-negligible probability.

To do this, we need a lower bound on the deviations of $\overline{X}_{1,s}$ from $1/2$, which is provided by the following lemma.

Lemma 2. *Let $\overline{X}_s$ denote the mean of s independent Bernoulli random variables with parameter $1/2$. There is a constant $C > 0$ such that for any $s > 1$ and any $1 \le \kappa \le s^{1/3}/(8 \log s)$,*

$$\mathbb{P}\left(\overline{X}_s - \frac{1}{2} \le -\sqrt{\frac{\kappa \log s}{2s}}\right) \ge \frac{C s^{-\kappa}}{\sqrt{\kappa \log s}}.$$

Proof (of Lemma 2). From Stirling's formula

$$n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n+1)} < n! < n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n)}, \tag{22}$$

for ℓ such that $(s+\ell)/2 \in \{0, 1, \ldots, s\}$, we have

$$\begin{aligned}
&\mathbb{P}\big(\overline{X}_s - \tfrac{1}{2} = -\tfrac{\ell}{2s}\big) \\
&\quad = \big(\tfrac{1}{2}\big)^s \tbinom{s}{\frac{s+\ell}{2}} \\
&\quad \ge \big(\tfrac{1}{2}\big)^s \frac{(\frac{s}{e})^s \sqrt{2\pi s} e^{\frac{1}{12s+1}}}{(\frac{s+\ell}{2e})^{\frac{s+\ell}{2}} (\frac{s-\ell}{2e})^{\frac{s-\ell}{2}} \sqrt{\pi(s+\ell)}\sqrt{\pi(s-\ell)} e^{\frac{1}{6(s+\ell)}} e^{\frac{1}{6(s-\ell)}}} \\
&\quad = \frac{1}{(1+\frac{\ell}{s})^{\frac{s+\ell}{2}} (1-\frac{\ell}{s})^{\frac{s-\ell}{2}}} \sqrt{\frac{2s}{\pi(s^2-\ell^2)}} e^{\frac{1}{12s+1} - \frac{1}{6(s+\ell)} - \frac{1}{6(s-\ell)}} \\
&\quad \ge \sqrt{\frac{2}{\pi s}} e^{-\frac{s-\ell}{2}\log(1-\frac{\ell^2}{s^2})} e^{-\ell \log(1+\frac{\ell}{s})} e^{-\frac{1}{6(s+\ell)} - \frac{1}{6(s-\ell)}} \\
&\quad \ge \sqrt{\frac{2}{\pi s}} e^{-\frac{\ell^2}{2s} - \frac{\ell^3}{2s^2} - \frac{1}{6(s+\ell)} - \frac{1}{6(s-\ell)}},
\end{aligned} \tag{23}$$

where the last inequality uses $\log(1+t) \le t$ for any $t > 0$. Let $\ell_0 = \sqrt{2\kappa s \log s} + \sqrt{s/(2\kappa \log s)}$. For ℓ such that $\sqrt{2\kappa s \log s} \le \ell \le \ell_0$, since $\ell_0 \le 2\sqrt{2\kappa s \log s} \le s^{2/3}$ and $s \ge 2$, we have

$$\begin{aligned}
\mathbb{P}\big(\overline{X}_s - \tfrac{1}{2} = -\tfrac{\ell}{2s}\big) &\ge \sqrt{\frac{2}{\pi s}} e^{-\frac{\ell_0^2}{2s} - \frac{\ell_0^3}{2s^2} - \frac{1}{3(s-\ell_0)}} \\
&\ge \sqrt{\frac{2}{\pi s}} s^{-\kappa} e^{-1 - \frac{1}{4\kappa \log s} - \frac{1}{2} - \frac{1}{3(s-s^{2/3})}} \\
&\ge \tfrac{1}{30}\sqrt{\tfrac{2}{s}} s^{-\kappa}.
\end{aligned}$$

By summing the probabilities corresponding to $\sqrt{2\kappa s \log s} \le \ell \le \ell_0$, we obtain

$$\mathbb{P}\left(\overline{X}_s - \frac{1}{2} \le -\sqrt{\frac{\kappa \log s}{2s}}\right) \ge \frac{s^{-\kappa}}{30\sqrt{\kappa \log s}}.$$

□

Let $\zeta' = (1+\zeta)/2$ and $\kappa > 1/(1-\zeta)$ such that $n^{\zeta'/\kappa}$ is an integer larger than $(8\zeta' \log n)^3$ (for a fixed $\zeta < 1$, such a κ exists as soon as n is sufficiently large). We consider $s = n^{\zeta'/\kappa}$ so that from (20) and Lemma 2, we obtain

$$\mathbb{P}\left(B_{1,s,t} \le \tfrac{1}{2} - (\sqrt{\zeta'} - \sqrt{\zeta})\sqrt{\frac{\log n}{2n^{\zeta'/\kappa}}} + \tilde{b}\frac{\log n}{n^{\zeta'/\kappa}}\right) \ge C\frac{n^{-\zeta'}}{\sqrt{\zeta' \log n}}. \tag{24}$$

In view of (21), we take $\varepsilon_n = \frac{\sqrt{\zeta'} - \sqrt{\zeta}}{2}\sqrt{\frac{\log n}{2n^{\zeta'/\kappa}}} - 2\tilde{b}\frac{\log n}{n^{\zeta'/\kappa}}$ such that with probability at least $C\frac{n^{-\zeta'}}{\sqrt{\zeta' \log n}}$, we draw the optimal arm no more than $s = n^{\zeta'/\kappa}$ times.

Up to multiplicative constants, this leads to an expected number of draws of suboptimal arms larger than $(n - n^{\zeta'/\kappa})\frac{n^{-\zeta'}}{\sqrt{\log n}} \approx \frac{n^{1-\zeta'}}{\sqrt{\log n}}$ and an expected regret larger than $(n - n^{\zeta'/\kappa})\varepsilon_n n^{-\zeta'} \approx n^{1-\zeta'-\zeta'/(2\kappa)} > n^{(1-\zeta)/2-1/(2\kappa)}$ up to logarithmic factors. Since the exponent is positive, 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 $c\zeta < 1/3$, there exist probability distributions of the rewards such that the UCB-V algorithm suffers a polynomial loss.*

Proof. See Section A.4.

The construction used in the proof is a 2-armed bandit problem, where the optimal arm has a Bernoulli payoff with a parameter ε adjusted to $c\zeta$ and the suboptimal arm deterministically gives a payoff of $\varepsilon/2$. The parameter ε is chosen such that with a polynomially decaying probability it holds that the optimal arm during its first $O(\log n)$ pulls always returns 0 and as a result it is not pulled more than $\Omega(\log n)$ times during the first n steps. This results in a polynomial regret.

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

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

This choice corresponds to the critical exploration function $\mathcal{E}_t = \zeta \log t$ with $\zeta = 1$ and to $c = 1/3$, 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 than these.

5. Risk bounds

Decision makers may care not only about a good expected return, but also about the distribution of the return. One desired property of a good algorithm is to guarantee high returns with high probability, alternatively to guarantee that the probability of a large regret is small. Motivated by this, in the next section we study the tail distribution of 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 the regret of UCB-V. These results are illustrated by computer experiments in Section 6.

5.1. Risk bounds for UCB1

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

$$B_{k,s,t} \triangleq \overline{X}_{k,s} + b\sqrt{\frac{\rho \log t}{s}}. \tag{25}$$

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

The user may thus choose a range of possible algorithms between an algorithm (when setting ρ 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 ρ 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 $\rho > 1$. The expected pseudo-regret for UCB1 defined by (25) satisfies:*

$$\mathbb{E}[R_n] \leq \sum_{k:\Delta_k>0} \left[\frac{4b^2}{\Delta_k} \rho \log(n) + \Delta_k \left(\frac{3}{2} + \frac{1}{2(\rho-1)} \right) \right]. \tag{26}$$

The proof parallels the proof of Theorem 3. We start with a Lemma that mimics Lemma 1.

Lemma 3. *Let $n \geq 2$, k be index of some arm and $u = \left\lceil \left(\frac{2b}{\Delta_k}\right)^2 \rho \log n \right\rceil$. Then, for any $u \leq s \leq t \leq n$, we have*

$$\mathbb{P}\left(B_{k,s,t} > \mu^*\right) \leq e^{-s\Delta_k^2/(2b^2)} \tag{27}$$

Proof (of Lemma 3). By the choice of u, s, t, we have $b\sqrt{\frac{\rho \log(t)}{s}} \leq \Delta_k/2$. Therefore, $\mathbb{P}\left(B_{k,s,t} > \mu^*\right) = \mathbb{P}\left(\overline{X}_{k,s} + b\sqrt{\frac{\rho \log(t)}{s}} > \mu_k + \Delta_k\right) \leq \mathbb{P}\left(\overline{X}_{k,s} > \mu_k + \Delta_k/2\right) \leq e^{-s\Delta_k^2/(2b^2)}$, where we used Hoeffding's inequality (cf. [8]). □

Proof (of Theorem 7). Again, because $R_n = \sum_k \Delta_k T_k(n)$ it suffices to bound $\mathbb{E}[T_k(n)]$, where k is the index of a suboptimal arm. Thus, pick such an index k. We use (11) to bound $\mathbb{E}[T_k(n)]$ with $\tau = \mu^*$ and u as in Lemma 3:

$$\mathbb{E}[T_k(n)] \leq u + \sum_{t=u+1}^{n} \sum_{s=u}^{t-1} \mathbb{P}\left(B_{k,s,t} > \mu^*\right) + \sum_{t=u+1}^{n} \sum_{s=1}^{t-1} \mathbb{P}\left(B_{k^*,s,t} \leq \mu^*\right). \tag{28}$$

Therefore, for any $s \geq u$, $\mathbb{P}\big(B_{k,s,t} > \mu^*\big) \leq e^{-u\Delta_k^2/(2b^2)} \leq n^{-2\rho}$ and we deduce that $\sum_{t=u+1}^{n}\sum_{s=u}^{t-1}\mathbb{P}\big(B_{k,s,t} > \mu^*\big) \leq n^{2(1-\rho)}/2$. The first sum in (28) is thus bounded by $n^{2(1-\rho)}/2 \leq 1/2$ whenever $n \geq 1$.

For the second sum, we have $\mathbb{P}\big(B_{k^*,s,t} \leq \mu^*\big) \leq t^{-2\rho}$, again from Hoeffding's inequality. Thus

$$\sum_{t=u+1}^{n}\sum_{s=1}^{t-1}\mathbb{P}\big(B_{k^*,s,t} \leq \mu^*\big) \leq \sum_{t=u+1}^{n} t^{1-2\rho} \leq \int_u^\infty t^{1-2\rho}dt = \frac{u^{-2(\rho-1)}}{2(\rho-1)}$$

for $\rho > 1$. Thus (28) implies that $\mathbb{E}\left[T_k(n)\right] \leq \left(\frac{2b}{\Delta_k}\right)^2 \rho\log(n) + \frac{3}{2} + \frac{1}{2(\rho-1)}$ holds for all $n \geq 1$. The bound on the expected regret follows. □

Theorem 8. *Assume that $\rho > 1/2$. Let $v_k = (2b/\Delta_k)^2$, $r_0 = \sum_k \Delta_k(1 + \rho v_k \log n)$. Then, for any $x \geq 1$, we have*

$$\mathbb{P}\big(R_n > r_0 x\big) \leq \sum_{k:\Delta_k>0}\left\{n^{-2\rho x+1} + \frac{\big((1+\rho v_k \log n)x\big)^{-2\rho+1}}{2\rho-1}\right\}. \tag{29}$$

Proof. We have:

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

Define $u_k = \lfloor(1+\rho v_k \log n)x\rfloor$. Hence, $\mathbb{P}\big(T_k(n) > (1+\rho v_k\log n)x\big) \leq \mathbb{P}\big(T_k(n) > u_k\big)$. We use (12) with $u = u_k$ and $\tau = \mu^*$ to bound $\mathbb{P}\big(T_k(n) > u_k\big)$:

$$\mathbb{P}\big(T_k(n) > u_k\big) \leq \sum_{t=u_k+1}^{n}\mathbb{P}\big(B_{k,u_k,t} > \mu^*\big) + \sum_{s=1}^{n-u_k}\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big). \tag{30}$$

Since $u_k \geq \lceil\rho v_k \log n\rceil$, we can apply Lemma 3. This gives $\mathbb{P}\big(B_{k,u_k,t} > \mu^*\big) \leq e^{-u_k\Delta_k^2/(2b^2)}$, which can be further bounded by $e^{-2x\rho\log n} = n^{-2\rho x}$ since $u_k \geq x\rho v_k \log n$. Hence, the first sum in (30) is bounded by $n^{-2\rho x+1}$.

Now, Hoeffding's inequality gives $\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big) \leq (u_k+s)^{-2\rho}$. Thus the second sum in (30) is bounded by $\sum_{s=1}^{n-u_k}\mathbb{P}\big(B_{k^*,s,u_k+s} \leq \mu^*\big) \leq \sum_{s=1}^{n-u_k}(u_k+s)^{-2\rho} \leq \int_{u_k}^\infty t^{-2\rho}dt = \frac{u_k^{1-2\rho}}{2\rho-1} \leq \frac{((1+v_k\rho\log n)x)^{1-2\rho}}{2\rho-1}$. Collecting the terms gives (29). □

The second term of (29) in Theorem 8 is only polynomial in x. In fact, this bound cannot be improved in the sense that there exist distributions of the rewards for which for some constant $C > 0$, for any z large enough, $\mathbb{P}\big(R_n > z\big) \geq 1/(Cz^C)$. See Theorem 10 for the analogous statement for UCB-V.

Theorems 7 and 8 show that the more we explore (i.e. larger ρ is), the smaller the tail of the regret is. However, this comes at the price of a larger expected regret. The next section is devoted to proving similar results for UCB-V.

5.2. 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.23, 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

$$\tilde{\beta}_n(t) \triangleq 3 \min_{\substack{\alpha\geq 1,\ M\in\mathbb{N},\\ s_0=0<s_1<\cdots<s_M=n\\ \text{s.t.}\ \ s_{j+1}\leq\alpha(s_j+1)}} \sum_{j=0}^{M-1} e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j}+t+1}{\alpha}}. \tag{31}$$

This function will appear naturally in the tail-bound of the pseudo-regret of UCB-V. Although $\tilde{\beta}_n(t)$ has a complicated definition, up to second order logarithmic terms it is of the order $e^{-(c\wedge 1)\mathcal{E}_t}$ when $\mathcal{E}_t = \Theta(\log t)$. This can be seen by considering (disregarding rounding issues) the geometric grid $s_j = \alpha^j$ with α close to 1 and noting that with $C_\alpha = (c\wedge 1)/\alpha$, we have $\sum_{j:\alpha^j\leq t} e^{-C_\alpha \log(t+\alpha^j)} = \Theta(\frac{\log t}{\log\alpha} e^{-C_\alpha \log t})$ and $\sum_{j:\alpha^j>t} e^{-C_\alpha \log(t+\alpha^j)} = \Theta(e^{-C_\alpha \log t})$.

One of the main results of the paper is the following tail-bound for the pseudo-regret of UCB-V:

Theorem 9. *Let*

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

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

$$\mathbb{P}\big(R_n > r_0 x\big) \leq \sum_{k:\Delta_k>0} \left\{ 2ne^{-(c\vee 1)\mathcal{E}_n x} + \tilde{\beta}_n(\lfloor v_k\mathcal{E}_n x\rfloor)\right\}. \tag{32}$$

Proof. The proof parallels the proof of Theorem 8. First note that

$$\begin{aligned} \mathbb{P}\big(R_n > r_0 x\big) &= \mathbb{P}\Big\{ \textstyle\sum_{k:\Delta_k>0} \Delta_k T_k(n) > \sum_{k:\Delta_k>0} \Delta_k(1 + v_k\mathcal{E}_n)x\Big\} \\ &\leq \textstyle\sum_{k:\Delta_k>0} \mathbb{P}\Big\{T_k(n) > (1 + v_k\mathcal{E}_n)x\Big\}. \end{aligned} \tag{33}$$

We use (12) with $\tau = \mu^*$ and $u = \lfloor(1 + v_k\mathcal{E}_n)x\rfloor$. Since $u \geq \lceil v_k\mathcal{E}_n\rceil$ we can apply Lemma 1 to get $\mathbb{P}(B_{k,u,t} > \mu^*) \leq 2e^{-u\Delta_k^2/(8\sigma_k^2+4b\Delta_k/3)} \leq 2e^{-(c\vee 1)\mathcal{E}_n x}$, where the last inequality follows because $u \geq v_k\mathcal{E}_n x$. To bound the other probability in (12), we use $\alpha \geq 1$ and the grid $s_0, \ldots, s_M$ of $\{1, \ldots, n\}$ realizing the minimum

of (31) when $t = u$. Let $I_j = \{s_j + 1, \dots, s_{j+1}\}$. Then

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

where the last line comes from (47) with the roles $'n' = s_{j+1}$, 't' restricted to I_j and 'x'$= (c \wedge 1)\mathcal{E}_{s_j+u+1}/\alpha$. □

In particular, when $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$, the last term dominates the first in (32), and 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. *Assume that $c = 1$ and $\mathcal{E}_t = \zeta \log t$, where $\zeta > 1$. Then there exist $\kappa_1 > 0$ and $\kappa_2 > 0$ that depend only on b, K, $\sigma_1, \dots, \sigma_K$ and $\Delta_1, \dots, \Delta_K$ such that for any $\varepsilon > 0$, $n \ge 3$ and $z > \kappa_1 \log n$, it holds that*

$$\mathbb{P}\big(R_n > z\big) \le \frac{(\kappa_2\zeta)^\zeta}{\varepsilon}\frac{\log(z/\kappa_1)}{z^{\zeta(1-\varepsilon)}}.$$

Proof. It suffices to prove the result for $\varepsilon \le 1/2$, since for larger ε, the property holds by possibly considering a twice larger constant κ_2. For $\kappa_3 > 0$ and $\kappa_4 > 0$ well chosen and depending only on $\theta \triangleq (b, K, \sigma_1, \dots, \sigma_K, \Delta_1, \dots, \Delta_K)$, Theorem 9 gives

$$\mathbb{P}(R_n > \kappa_3\mathcal{E}_n x) \le 2nKe^{-\mathcal{E}_n x} + K\tilde{\beta}_n(\lfloor \kappa_4\mathcal{E}_n x\rfloor).$$

Defining $x = z/(\kappa_3\mathcal{E}_n)$ and $z' = \lfloor \kappa_4\mathcal{E}_n x\rfloor = \lfloor \kappa_4/\kappa_3 z\rfloor$, this rewrites into

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

For $\kappa_1 \triangleq 2\kappa_3$, $n \ge 3$ and $z > \kappa_1 \log n$, we have $ne^{-z/\kappa_3} \le e^{-z/\kappa_1}$ so the first

term of the r.h.s is upper bounded by

$$\begin{aligned} 2Ke^{-z/\kappa_1} &\leq 2K\frac{\log(z/\kappa_1)}{(z/\kappa_1)^\zeta}\sup_{s>\kappa_1\log 3} e^{-s/\kappa_1}\frac{(s/\kappa_1)^\zeta}{\log(s/\kappa_1)} \\ &\leq 2K\frac{\log(z/\kappa_1)}{(z/\kappa_1)^\zeta}\sup_{u>\log 3} e^{-u}\frac{u^\zeta}{\log(\log 3)} \\ &\leq 2K\frac{\log(z/\kappa_1)}{(z/\kappa_1)^\zeta}\frac{\zeta^\zeta}{\log(\log 3)} \\ &\leq \frac{\log(z/\kappa_1)}{z^\zeta}(\kappa_2'\zeta)^\zeta \end{aligned}$$

for an appropriate choice of κ_2' that depends only on θ. To upper bound $\tilde{\beta}_n(z')$, we consider a geometric grid with increment $\alpha = 1/(1-\varepsilon)$ and split the sum defining $\tilde{\beta}_n(z')$ (cf. (31)) into two parts: for indexes j with $s_j \leq z'$ we use

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

whereas for indexes j with $s_j > z'$, we use $e^{-\frac{(c\wedge 1)\mathcal{E}_{s_j+z'+1}}{\alpha}} \leq e^{-\frac{\mathcal{E}_{s_j}}{\alpha}} \leq e^{-j\zeta\frac{\log\alpha}{\alpha}}$. The first part of the sum has at most $1+(\log z')/\log[1/(1-\varepsilon)]$ terms, which is of order $(\log(z/\kappa_1))/\varepsilon$ when $\varepsilon \leq 1/2$. Let j_0 be the smallest index with $s_j > z'$. We bound the second part of the sum as follows:

$$\frac{e^{-j_0\zeta(\log\alpha)/\alpha}}{1-e^{-\zeta(\log\alpha)/\alpha}} \leq \frac{1}{1-(1-\varepsilon)^{\zeta(1-\varepsilon)}}\left(\frac{z'}{1-\varepsilon}\right)^{\zeta(1-\varepsilon)} \leq \frac{2^\zeta(z')^{\zeta(1-\varepsilon)}}{1-\sqrt{1-\varepsilon}} \leq \frac{2^{\zeta+1}(z')^{\zeta(1-\varepsilon)}}{\varepsilon},$$

where the second to last inequality uses $2^{-\zeta} \leq (1-\varepsilon)^{\zeta(1-\varepsilon)} \leq (1-\varepsilon)^{1/2}$ which holds since $\zeta > 1$ and $\varepsilon \leq 1/2$. Combining the bounds gives the final result. □

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 it improves as ζ grows, is pretty slow. For comparison, remember that a zero-mean martingale M_n with increments bounded by 1 would satisfy $\mathbb{P}(M_n > z) \leq \exp(-2z^2/n)$. The slow concentration for UCB-V happens because the first $\Omega(\log(t))$ choices of the optimal arm can be unlucky (yielding small rewards) in which case the optimal arm will not be selected any more during the first t steps. As a result, the distribution of the regret will be of a mixture form with a mode whose position scales linearly with time and whose associated mass decays only at a polynomial rate. The rate of this decay is in turn controlled by ζ. The following result shows that the polynomial rate obtained in Corollary 1 cannot be replaced by an exponential rate when there is a chance for the optimal arm to be unlucky.[5]

[5] An entirely analogous result holds for UCB1.

Theorem 10. *Assume that the optimal arm is unique. Consider $\mathcal{E}_t = \zeta \log t$ with $c\zeta > 1$. Let $\tilde{\mu} = \sup\{v \in \mathbb{R} : \mathbb{P}(X_{k^*,1} < v) = 0\}$ be the essential infimum of the optimal arm's distribution and let $\tilde{k}$ be the index of a second best arm. The followings hold:*

1. *If $\tilde{\mu} > \mu_{\tilde{k}}$ then the pseudo-regret has exponentially small tails.*
2. *If, on the contrary, $\tilde{\mu} < \mu_{\tilde{k}}$ then the pseudo-regret assumes a polynomial tail only.*

When there are multiple optimal arms and the minimum of the essential infimums of the optimal arms' payoffs is above the mean payoff of a second best arm then the first part of the result continues to hold. On the other hand, when the maximum of the essential infimums is below the mean payoff of a second best arm then the second part continues to hold.

Proof. First consider the case when $\tilde{\mu} > \mu_{\tilde{k}}$. Let μ' be such that $\mu_{\tilde{k}} < \mu' < \tilde{\mu}$ and let $\delta_k = \mu' - \mu_k$. The bound on the tail probability of R_n is bounded in terms of the tail-probabilities of $T_k(n)$, where k ranges over the indexes of suboptimal arms as in (33). Fix such a k. The tail of $T_k(n)$ is bounded by using (12) with $\tau = \mu'$ and where $u = \left\lceil 8(c \vee 1)\big(\frac{\sigma_k^2}{\delta_k^2} + \frac{2b}{\delta_k}\big)\mathcal{E}_n \right\rceil$. This value of τ makes the last probability in (12) vanish. The first term is controlled as in the proof of Theorem 9. Precisely, for $v'_k \triangleq 8(c \vee 1)\big(\frac{\sigma_k^2}{\delta_k^2} + \frac{2b}{\delta_k}\big)$, $r'_0 \triangleq \sum_{k:\Delta_k > 0} \Delta_k\big(1 + v'_k \mathcal{E}_n\big)$ and any $x \geq 1$ we have

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

which proves that R_n has exponential tails in this case.

Now consider the case when $\tilde{\mu} < \mu_{\tilde{k}}$. We prove the result for a special distribution first and then argue that the general case follows along similar lines. Consider the following payoff distributions:

- the optimal arm with index 1 concentrates its rewards on $\tilde{\mu}$ and b such that its expected reward is strictly larger than $\mu_{\tilde{k}}$,
- all suboptimal arms are deterministic to the extent that they always provide a reward equal to $\mu_{\tilde{k}}$.

Let q be any positive integer. Consider the event:

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

Let $c_2 \triangleq 3bc\zeta$ and $\eta \triangleq \mu_{\tilde{k}} - \tilde{\mu}$. On Γ 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 $k > 1$, $0 \leq s \leq t$, we have

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

This means that the optimal arm cannot be played more than q times during the first $e^{\eta q/c_2}$ plays. Hence, $R_n \geq \Delta_{\tilde{k}}\big(e^{\eta q/c_2} - q\big)$. Now, take q large enough so that $e^{\eta q/c_2} - q \geq \frac{1}{2} e^{\eta q/c_2}$ so that $R_n \geq \frac{1}{2}\Delta_{\tilde{k}}\, e^{\eta q/c_2}$. Further, let $w > 0$ be such that $n \triangleq e^{w^{-1}e^{\eta q/c_2}} = \lceil e^{\eta q/c_2} \rceil$. As $w \log n = e^{\eta q/c_2}$ we get

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

where $C = \frac{c_2}{\eta} \log(1/p)$ and $p = \mathbb{P}(X_{1,1} = \tilde{\mu})$. Since w increases with q and the inequality holds for any sufficiently large q (the threshold depends only on c_2 and η), we have thus shown that the pseudo-regret cannot have a tail thinner than polynomial.

The proof for the general case is essentially the same. The main difference is that Γ has to be redefined as the event when $X_{1,1}, X_{1,2}, \ldots, X_{1,q}$ are below μ'' with $\tilde{\mu} < \mu'' < \mu_{\tilde{k}}$, and when, for the second optimal arm, the empirical means stay close to the associated expected mean $\mu_{\tilde{k}}$. The rest of the proof, which is omitted here in the interest of saving some space, follows the same steps as above.

□

Remark 2. Theorem 9 and Corollary 1 provide tail bounds for the pseudo-regret, $R_n = \sum_{k=1}^K T_k(n)\Delta_k$, instead of the regret,

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

The following considerations show that when the optimal arm is unique, similar concentration bounds hold for the regret: Assume that $c = 1$ and $\mathcal{E}_t = \zeta \log t$ with $\zeta > 1$. By slightly modifying the analysis in Theorem 9 and Corollary 1, one can derive that there exists $C'' > 0$ such that for any $z > C'' \log n$, with probability at least $1 - z^{-1}$, the number of draws of suboptimal arms is bounded by Cz for some $C > 0$ (in this remark, the constants C, C' and C'' depend on b, K, $\sigma_1, \ldots, \sigma_K$ and $\Delta_1, \ldots, \Delta_K$ and may differ from line to line). So the algorithm draws the optimal arm at least $n - Cz$ times. This means that $n - Cz$ terms cancel out in the sum defining the regret. For the Cz remaining terms, one can use Hoeffding's inequality and union bounds to prove that with probability $1 - Cz^{-1}$, for any suboptimal arm k,

$$\frac{\sum_{t=1}^{T_k(n)} (X_{k,t} - \mu_k)}{\sqrt{T_k(n)}} \leq \max_{1 \leq s \leq Cz} \frac{\sum_{t=1}^{s} (X_{k,t} - \mu_k)}{\sqrt{s}} \leq C' \sqrt{\log z},$$

hence, by the Cauchy-Schwarz inequality

$$\hat{R}_n - R_n \leq C' \sum_{k \neq k^*} \sqrt{T_k(n) \log z} \leq C' \sqrt{K-1} \sqrt{Cz \log z}.$$

Therefore, with probability at least $1 - z^{-1}$, we simultaneously have $\hat{R}_n \leq R_n + C'\sqrt{z \log z}$ and $R_n \leq Cz$. Since $\sqrt{z \log z} = \mathrm{o}(z)$, the regret $\hat{R}_n$ has similar

tails than the pseudo-regret R_n. Thus, we conclude that for $z > C \log n$, with probability at least $1 - z^{-1}$, $\hat{R}_n \le C' z$.

6. Numerical experiments

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 may take a lot of trials to identify the optimal arm. In particular, if the experiments start in an unlucky way, the algorithms will keep choosing 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, ρ (cf. Equation (25)), 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 (18)). 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 $\alpha \ge 0$. Define the *value at risk* for the risk level α as the upper α-percentile of the regret:

$$R_n(\alpha) = \inf\{r : \mathbb{P}(R_n \ge r) \le \alpha\}.$$

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 $\alpha = 0.01$, 0.16 and 0.5 (the latter value corresponding to the median). These figures also show the mean regret and (estimated) upper percentiles of Gaussians fitted to the respective regret distributions. (The labels of the percentile curves for the Gaussians are marked by pasting "(n)" after the respective α-values. The percentiles were estimated by drawing $10,000$ values

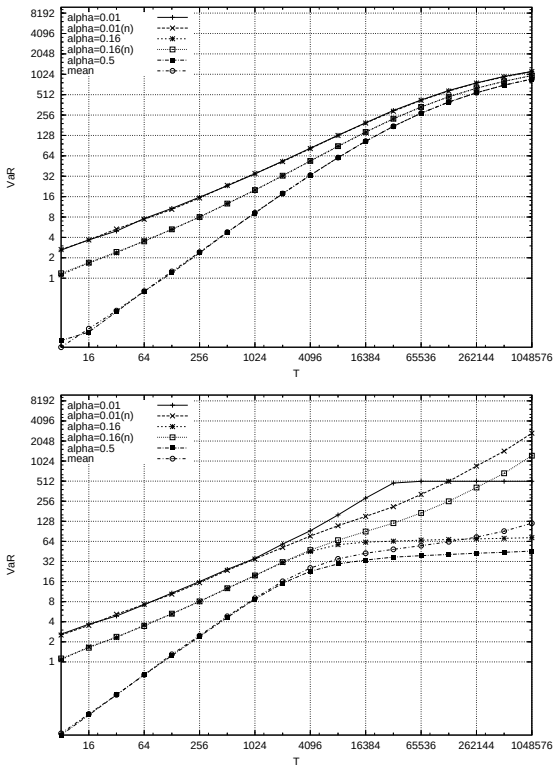

Figure 1: Value at risk as a function of time when the expected payoff of the suboptimal arm is $p = 0.48$. The upper figure depicts results for UCB1, while the lower one depicts results for UCB-V. Note the logarithmic scale of the time axis. For more details see the text.

from the respective Gaussians.) If the regret is normally distributed, we can expect a good match between the respective percentile curves.

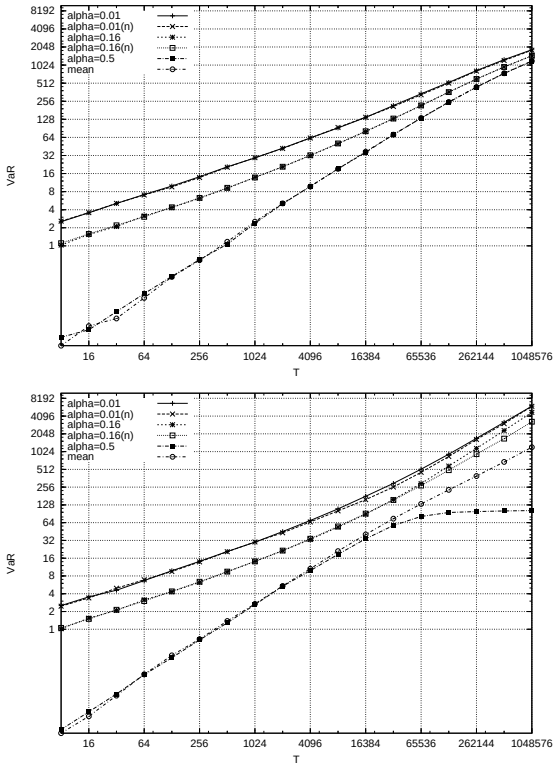

Figure 2: Value at risk as a function of time when the expected payoff of the suboptimal arm is $p = 0.495$. The upper figure depicts results for UCB1, while the lower one figure depicts results for UCB-V. For more details see the text.

As expected, in the case of the "easier" problem UCB-V outperforms UCB1 by a large margin except for the smallest value α (which partially confirms the results on the scaling of the expected regret with the variance of the suboptimal arms). For UCB1, uniformly over time, the distribution of regret is well approximated by Gaussians. 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 shift the vast majority of the probability mass to the optimal arm by the end of the experiment. At the

same time, for UCB-V the shift happens at around $T = 8,192$. Note that in the 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 of UCB-V for $\alpha = 0.16$ and smaller is significantly larger than that for UCB1.

In order to illustrate what "goes wrong" with UCB-V for 20 independent runs we show in Figure 3 the time evolution of the proportion of time-steps when the suboptimal arm is chosen. That is, the figure shows the time evolution of $T_{\text{bad}}(t)/t$ for 20 different runs, where $T_{\text{bad}}(t) = \sum_{s=1}^{t} \mathbb{I}_{\{I_s \text{ is the bad arm}\}}$. We see that in quite a few runs the suboptimal arm is preferred for a long time, though ultimately all curves converge to 0.

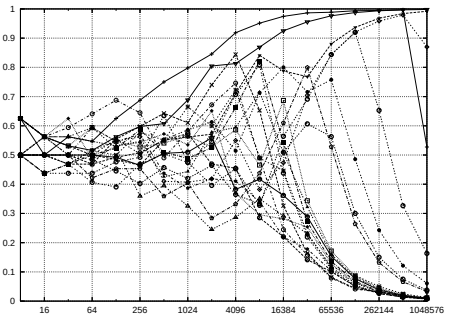

Figure 3: $T_{\text{bad}}(t)/t$, the proportion of times of using the suboptimal arm in the first t time-steps as a function of time for 20 independent runs. The bandit problem has parameter $p = 0.495$ and the algorithm is UCB-V.

Based on Figure 3 one may suspect that the distribution of $T_{\text{bad}}(t)/t$ is bimodal. This is confirmed by Figure 4 which shows this distribution as a function of time. Note that at around time $T = 2,048$ ($\log_2(T) = 11$) the probability mass indeed becomes bimodal. At this time, the probability mass is split into two with a larger mass shifting towards the (desired) mode with value 0, while a smaller, but still substantial mass drifting towards 1. The mass of this second mode is continuously decreasing, albeit at a slow rate. The slow rate of this decay 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), but 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., $\mathbb{P}(T_{\text{bad}}(t) \geq \alpha t)$), as a function of time and

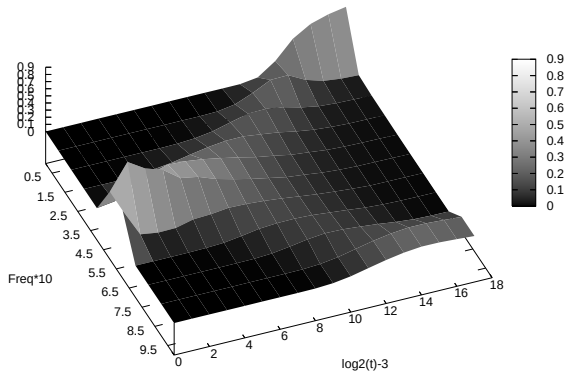

Figure 4: The distribution of $T_{\mathrm{bad}}(t)/t$, the frequency of using the suboptimal arm, plotted against time. The bandit problem has parameter $p = 0.495$ and the algorithm is UCB-V.

for various values of α, see Figure 5. The figure reinforces that in the initial

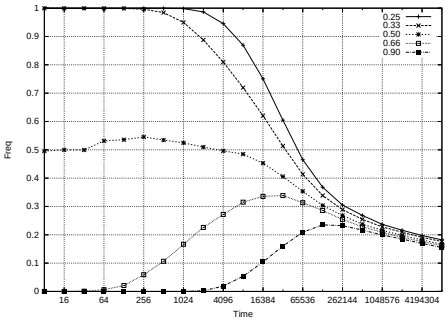

Figure 5: The probability of choosing the suboptimal arm more than α-fraction of time plotted against time and various values of α. Note that the experiment was continued up to $T = 2^{24}$ steps to show the beginning of the asymptotic phase.

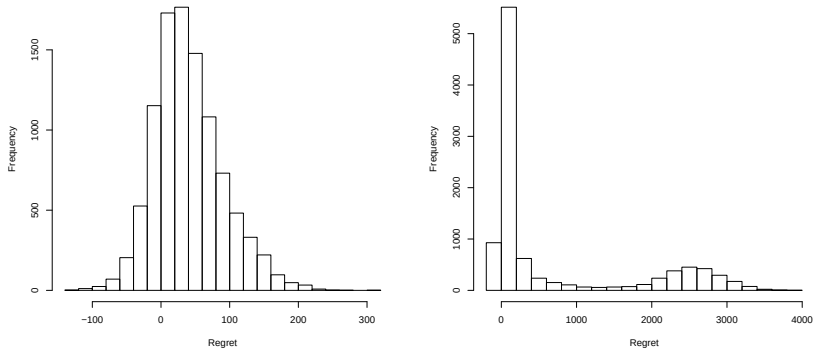

Figure 6: Distribution of the regret for UCB-V at time $T_1 = 16,384$ (l.h.s. figure) and $T_2 = 524,288$ (r.h.s. figure). The bandit problem has parameter $p = 0.495$.

phase $T_{\text{bad}}(t)$ is concentrated around $0.5t$. At the time when the two modes appear most of the mass drifts towards zero, though at the same time some mass is drifting towards t as indicated by the large spread of $\mathbb{P}(T_{\text{bad}}(t) \geq \alpha t)$. The fact that all curves are converging to each other reveals that the 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 (not shown here), it looks as if asymptotically the curves followed a polynomial curve.

To show that also the regret 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 contain 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.

7. 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 a UCB algorithm which with high probability plays any suboptimal arm only a few times. Hence, the algorithm is "Probably Approximately Correct", explaining the algorithm's name. Note that in this setting, the quantity $B_{k,s,t}$ does not depend on the time t so in what follows we will write $B_{k,s}$ instead of $B_{k,s,t}$. Besides, in order to simplify the discussion, we take $c = 1$.

Theorem 11. *Let $\beta \in (0,1)$. Consider a sequence $(\mathcal{E}_s)_{s\geq 0}$ that takes values in $\mathbb{R} \cup \{+\infty\}$ and satisfies $\mathcal{E}_s \geq 2$ and*

$$4K \sum_{s\geq 7} e^{-\mathcal{E}_s} \leq \beta. \tag{34}$$

Let k be the index of some suboptimal arm and let u_k be the smallest integer satisfying

$$\frac{u_k}{\mathcal{E}_{u_k}} > \frac{8\sigma_k^2}{\Delta_k^2} + \frac{26b}{3\Delta_k} \tag{35}$$

with the understanding that if no integer index satisfies this inequality then $u_k = +\infty$. Then with probability at least $1-\beta$ it holds that no suboptimal arm k is played more than u_k times by PAC-UCB.

When $\mathcal{E}_s$ takes only finite values the existence of a finite u_k is guaranteed if $\mathcal{E}_s = o(s)$. Note that infinite values of $\mathcal{E}_s$ are allowed only for technical reasons. In particular, this will be needed when we apply this theorem in a finite horizon setting in which case we will use $\mathcal{E}_s = +\infty$ for s bigger than the horizon.

Proof. See Section A.5.

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

$$\mathcal{E}_s = \log(Ks^q\beta^{-1}) \vee 2, \tag{36}$$

up to some additive constant ensuring that (34) 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 value is independent of the total number of plays. Hence, we get the following upper bound on the pseudo-regret:

$$R_n = \textstyle\sum_{k=1}^K T_k(n)\Delta_k \le \sum_{k:\Delta_k>0} \mathcal{T}_{k,\beta}\Delta_k. \tag{37}$$

One should notice that the previous bound holds with an even set of probability at least $1-\beta$. On the complementing event no small upper bound is possible: there exist situations when with probability of at least $\Omega(\beta)$, the regret is of order n, while (37) still holds with probability greater than $1-\beta$. Hence, without any additional assumptions the following bound cannot be essentially improved:

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

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 the followings:*

- *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*

$$\mathbb{E}[R_n] \le \textstyle\sum_{k:\Delta_k>0} \big(\frac{24\sigma_k^2}{\Delta_k} + 30b\big)\log n. \tag{38}$$

Proof. See Section A.6.

8. 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:

$$\mathbb{P}\Big\{R_n > 24\textstyle\sum_{k:\Delta_k>0}\Big(\frac{\sigma_k^2}{\Delta_k} + 2b\Big)\log n\Big\} \quad \le \quad \frac{C}{n}. \tag{39}$$

The behavior of this algorithm should be contrasted to the one with $\mathcal{E}_{s,t} = 3\log t$: 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 that adapts to the time horizon explores and exploits at any time during the interval $[0, n]$. However, in view of Corollary 1 and Theorem 10, its regret satisfies

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

a significantly worse behavior than what is shown (39). The open question is: is there an algorithm that does not need to know the time horizon and which has a logarithmic expected regret and a concentration property similar to (39)? We conjecture that the answer is no.

Acknowledgements

This work was supported in part by the Agence Nationale de la Recherche project "Modèles Graphiques et Applications" (Jean-Yves Audibert). Csaba Szepesvári greatly acknowledges the support received from the Alberta Ingenuity Fund, iCore and NSERC.

A. Proofs of the results

A.1. Lower bound for UCB-V

The aim of this section is to prove that both terms in (4) are unavoidable. Precisely, we have the following result:

Theorem 13. *Fix b and any constant $C > 0$. Then there is no algorithm that would satisfy either*

$$\mathbb{E}[\hat{R}_n] \leq C \sum_{k:\mu_k<\mu^*} b \log(n), \tag{40}$$

or

$$\mathbb{E}[\hat{R}_n] \leq C \sum_{k:\mu_k<\mu^*} \frac{\sigma_k^2}{\Delta_k} \log(n) \tag{41}$$

uniformly for all reward distributions with support in $[0, b]$.

Proof. We apply a lower bound developed by Lai and Robbins [10]. Let δ_a be the Dirac distribution supported on $a \in \mathbb{R}$. Let $\nu_p = (1-p)\delta_0 + p\delta_b$ be a Bernoulli-like distribution parameterized by $p \in (0, 1)$. Consider a bandit policy. For $(p_1, p_2) \in (0, 1)^2$ let $\mathcal{R}_n(p_1, p_2)$ denote the expected regret of this policy when it is applied to a two-armed bandit problem in which the reward distributions for the two arms are respectively ν_{p_1} and ν_{p_2}. If for some $a > 0$, $(p_1, p_2) \in (0, 1)^2$, $\mathcal{R}_n(p_1, p_2) = \mathrm{o}(n^a)$ does not hold then the logarithmic regret bounds, (40), (41), cannot hold. Therefore let us assume that $\mathcal{R}_n(p_1, p_2) = \mathrm{o}(n^a)$ holds for any $a > 0$ and $(p_1, p_2) \in (0, 1)^2$. Then, from Lai and Robbins [10], Theorem 1 we conclude that for any $(p_1, p_2) \in (0, 1)^2$ with $p_1 > p_2$, we have

$$\liminf_{n\to+\infty} \frac{\mathcal{R}_n(p_1, p_2)}{\log n} \geq \frac{b(p_1 - p_2)}{p_1 \log(\frac{p_1}{p_2}) + (1-p_1)\log(\frac{1-p_1}{1-p_2})}.$$

Let $\Theta(p_1, p_2)$ denote the r.h.s. of this inequality. Let us consider $p_1 = (1+\delta)/2$ and $p_2 = (1-\delta)/2$ with $\delta \in (0,1)$. Then we have $\Theta(p_1, p_2) = b/\log[(1+\delta)/(1-\delta)]$. Since the logarithmic term goes to 0 when δ goes to 0, there is no algorithm which can satisfy (40) for all reward distributions in $\{\nu_p : p \in (0,1)\}$. Besides, we have $\sigma_2^2 = b^2(1-\delta^2)/4$, $\Delta_2 = b\delta$ and

$$\frac{\Delta_2}{\sigma_2^2}\Theta(p_1, p_2) = \frac{4\delta}{(1-\delta^2)\log[(1+\delta)/(1-\delta)]}.$$

Since the last r.h.s. goes to infinity when δ goes to 1, there is no algorithm which can satisfy (41) for all reward distributions in $\{\nu_p : p \in (0,1)\}$. □

A.2. Lower bound for UCB1

Proposition 1. *There exist arm rewards in $[0,b]$ such that UCB1 (defined by the bias factor (1)) has an expected regret $\mathbb{E}[R_n] = \Omega(b^2 \log n)$, while UCB-V with $c = 1$ and $\zeta = 1.2$ satisfies $\mathbb{E}[R_n] \leq 20b \log n$.*

Proof. Consider the 2-armed deterministic bandit problem such that arm 1 yields the reward Δ, and arm 2 yields the reward 0. In this case, Theorem 4 gives the desired property of UCB-V. For UCB1, in order to obtain a lower bound on the regret, we look for a lower bound on $T_2(n)$.

First consider the "balance equation"

$$\Delta + b\sqrt{\frac{2\log(n+1)}{n-p(n)}} = b\sqrt{\frac{2\log(n+1)}{p(n)}}, \tag{42}$$

where $p(n)$ is considered as a function of $n \geq 1$. Note that solving (42) 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, as a consequence, decreases its value.[6] Thus we expect that $T_2(n)$ will be close to $p(n)$. For that purpose, let us prove the following result.

Lemma 4. *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 $n \geq 3$, $T_2(n) \geq p(n-1)$.*

[6] The same holds for UCB-V. However, the corresponding "balance" equation for UCB-V looks different.

PROOF (OF LEMMA 4). 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

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

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

Now, let us prove the second part of the lemma. The proof by contradiction: Let $n \geq 3$ be the first time when $T_2(n) < p(n-1)$, 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 of UCB1 satisfies $R_n = T_2(n)\Delta = \Omega(\frac{b^2}{\Delta}\log n)$. □

A.3. 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 5. *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 $1 \leq t \leq n$,*

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{E}[U^2]\, x} + b''_+ x/3, \tag{43}$$

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

$$t(\overline{U}_t - \mu) \leq \sqrt{2n\mathbb{V}\mathrm{ar}(U)\, x} + b' x/3. \tag{44}$$

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

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

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

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

Consider $a = \sqrt{2nvx} + x/3$. To prove (44), it remains to check that

$$nvh[a/(nv)] \ge x. \tag{45}$$

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 $r \ge 0$, 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 $\log(1 + r') \ge r'/(1 + r')$ for any $r' \ge 0$. The proof of (44) is finished since $\varphi(\sqrt{2x/(nv)}) \ge 0$ implies (45).

To prove (43), we need to modify the martingale argument underlying Freedman's result. Precisely, let $g(r) \triangleq (e^r - 1 - r)/r^2$. Then we replace

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

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

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

By following Freedman's arguments, we get

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

Now if $b'' \le 0$, this minimum is upper bounded by

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

which leads to (43) when $b'' \le 0$. When $b'' > 0$, the minimum is reached for $\lambda b'' = \log\big(\frac{b''a + n\mathbb{E}[U^2]}{n\mathbb{E}[U^2]}\big)$. The computations then are similar to the one developed to obtain (44). □

Remark 3. Lemma 5 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 $w > 0$,

$$\mathbb{P}\big(\overline{U}_n - \mu > w\big) \le e^{-\frac{nw^2}{2\mathbb{V}\mathrm{ar}(U) + (2b'w)/3}}. \tag{46}$$

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

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

Compared with (44) 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 5 allows the time indices to vary in an interval. This form follows from a martingale argument due to Freedman [5].

PROOF (OF THEOREM 1). Given Lemma 5, the proof essentially reduces to an application of the "square-root trick". For the first part of the theorem, we will prove the following result: for any $x > 0$ and $n \in \mathbb{N}$, with probability at least $1 - 3e^{-x}$, for any $0 \le t \le n$,

$$|\overline{X}_t - \mu| < \tfrac{\sqrt{2nV_t x}}{t} + \tfrac{3bnx}{t^2}. \tag{47}$$

Note that this is slightly stronger than the first part of Theorem 1. We prove this result since we need it in the proof of the second part of the Theorem.

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 5, (44) with the choices $U_i = X_i$, $U_i = -X_i$, and Lemma 5, (43) with the choice $U_i = -(X_i - \mathbb{E}[X_1])^2$ yield that with probability at least $1 - 3e^{-x}$, for any $0 \le t \le n$, we simultaneously have

$$|\overline{X}_t - \mu| \le \sigma \tfrac{\sqrt{2nx}}{t} + \tfrac{x}{3t} \tag{48}$$

and

$$\sigma^2 \le V_t + (\mu - \overline{X}_t)^2 + \tfrac{\sqrt{2n\mathcal{V}\,x}}{t}. \tag{49}$$

Let $L \triangleq nx/t^2$. We claim that from (48) and (49), it follows that

$$\sigma \le \sqrt{V_t} + 1.8\sqrt{L}. \tag{50}$$

Since the random variable X_1 takes its values in $[0, 1]$, we necessarily have $\sigma \le 1/2$. Hence, when $1.8\sqrt{L} \ge 1/2$ then (50) is trivially satisfied, so from now on we may assume that $1.8\sqrt{L} \le 1/2$, i.e., $L \le (3.6)^{-2}$. Noting that $\mathcal{V} \le \sigma^2$, plugging (48) into (49) for $0 \le t \le n$ we obtain

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

or $\sigma^2 - 1.77\sqrt{L}\sigma - (V_t + \frac{L}{100}) \le 0$. 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: $\sigma \le \frac{1.77}{2}\sqrt{L} + \sqrt{V_t + 0.8L} \le \sqrt{V_t} + 1.8\sqrt{L}$, finishing the proof of (50). Plugging (50) into (48), we obtain

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

which, given the definition of L, proves (47), and thus the first part of Theorem 1.

Let us now consider the second part of the theorem: Fix $t_1 \le t_2$, $t_1, t_2 \in \mathbb{N}$ and let $\alpha \ge t_2/t_1$. From (47) it follows that simultaneously for $t \in \{t_1, \ldots, t_2\}$ we have with probability at least $1 - 3e^{-x/\alpha}$ that

$$\begin{aligned} t|\overline{X}_t - \mu| &< \sqrt{2t_2 V_t x/\alpha} + 3x/\alpha \\ &\le \sqrt{2tV_t x} + 3x. \end{aligned} \tag{51}$$

To finish the proof we will use this inequality for a sequence of suitably chosen intervals $[t_1, t_2]$ that form a partition of $[4, n]$. (It suffices to consider a partition of $[4, n]$, because for $1 - 3e^{-x/\alpha} \geq 0$ the r.h.s. of (51) is always greater than 3. Thus, for $t \leq 3$ inequality (51) holds with probability one.) For the rigorous reasoning, introduce

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

and let $s_0, \ldots, s_M$ be a grid realizing the above minimum. Then we have

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

where the last inequality follows since $s_0, \ldots, s_M$ forms a complete geometric grid of $\{3, 4, \ldots, n\}$ with step-size α. This finishes the proof of Theorem 1. □

Remark 4. Any PAC empirical bound on $|\overline{X}_t - \mu|$ leads to a corresponding UCB policy. The tighter the bound is, the more efficient (in terms of expected regret) the UCB policy is. Theorem 1 is essentially obtained by using Bernstein's inequality for both the empirical mean and variance. There is a small cost to consider the variance when it is high. Indeed, in the worst case, the variance is equal to $b^2/4$, so that (47) leads to that with probability at least $1 - 3e^{-x}$, $|\overline{X}_t - \mu| < b\sqrt{\frac{x}{2t}} + \frac{3bx}{t}$. This inequality has to be compared with Hoeffding's inequality (the one used in UCB1), which, for the same level of confidence, $1 - 3e^{-x}$, reads $|\overline{X}_t - \mu| < b\sqrt{\frac{x - \log 3}{2t}}$. This is much tighter than the former inequality when x/t is not very small. Note that Theorem 1 (and therefore UCB-V policy) can be (numerically) improved by using (48) and $\sigma \leq b/2$. This gives: for any $t \in \mathbb{N}$ and $x > 0$, with probability at least $1 - 3e^{-x}$,

$$|\overline{X}_t - \mu| \leq \left(\sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}\right) \wedge \left(b\sqrt{\frac{x}{2t}} + \frac{bx}{3t}\right),$$

and with probability at least $1 - \beta(x, t)$, for any $s \in \{1, 2, \ldots, t\}$,

$$|\overline{X}_s - \mu| \leq \left(\sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s}\right) \wedge \left(b\sqrt{\frac{x}{2s}} + \frac{bx}{3s}\right).$$

Finally, the last term in $B_{k,s,t}$, which corresponds to the bx/s term in the previous inequality, does play a role when the variance is very small. It cannot

be eliminated as can be seen by considering the case when X_1 is a Bernoulli of parameter λ/s with $\lambda > 0$. Indeed, in this case, $s\overline{X}_s$ has the distribution $\mathrm{Bin}(s, \lambda/s)$, which converges in law to Poisson(λ) for s tending to infinity. Now, it is known that there are no positive constants c_1 and c_2 such that the inequality $\mathbb{P}(|Z - \mathbb{E}Z| \geq c_1\sqrt{x\mathbb{V}\mathrm{ar}\, Z}) \leq c_2 e^{-x}$ holds for all Poisson distributions (because of the left "tail" of the Poisson distributions).

Remark 5. One may also write a one-sided version of Theorem 1 taking into account the previous remark, namely for any $t \in \mathbb{N}$ and $x > 0$, with probability at least $1 - 2e^{-x}$,

$$\overline{X}_t - \mu \leq \left(\sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}\right) \wedge \left(b\sqrt{\frac{x}{2t}} + \frac{bx}{3t}\right), \tag{52}$$

and with probability at least $1 - 2\inf_{1<\alpha\leq 3}\left(\frac{\log t}{\log \alpha} \wedge t\right)e^{-x/\alpha}$, for any $s \in \{1, 2, \ldots, t\}$,

$$\overline{X}_s - \mu \leq \left(\sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s}\right) \wedge \left(b\sqrt{\frac{x}{2s}} + \frac{bx}{3s}\right). \tag{53}$$

To prove (52), we use (49) and the one-sided version of (48), which holds simultaneously with probability $1 - 2e^{-x}$. We claim that these inequalities imply that either (50) holds or $\overline{X}_s - \mu \leq 0$. In both cases, (52) follows. Inequality (53) follows from a similar argument.

A.4. Proof of Theorem 6

We will prove the following, slightly stronger result.

Theorem 14. *Consider $\mathcal{E}_t = \zeta \log t$. For any $\zeta > 0$ and $p \in (0, 1)$, if $c\zeta < \frac{p}{-3\log(1-p)}$, there exist probability distributions of the rewards such that the mean reward of the optimal arm is pb and the UCB-V algorithm suffers a polynomial loss.*

Theorem 14 implies Theorem 6 by letting $p \to 0$.

Proof (of Theorem 14). For $c\zeta < \frac{p}{-3\log(1-p)}$, there exists $\varepsilon \in (0, 1)$ such that

$$c\zeta = \varepsilon^2 \frac{p}{-3\log(1-p)} \tag{54}$$

Consider the following two-armed bandit problem: Let $\{X_{1,t}\}$ be an i.i.d. sequence with $\mathbb{P}(X_{1,t} = b) = p = 1 - \mathbb{P}(X_{1,t} = 0)$. Let $\{X_{2,t}\}$ be the deterministic sequence given by $X_{2,t} = pb\varepsilon$. Arm 1 is then the optimal arm and its mean reward is pb. Fix $n \in \mathbb{N}$. Let $T = \lceil \gamma \log n \rceil$ with $\gamma = -\varepsilon/\log(1-p)$. We consider large values of n for which $n > T$.

Claim: Consider an event when during the first T 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., $T_1(n) \leq T$.

PROOF. The claim is proved by contradiction. Assume that on the considered event, the optimal arm is pulled more than T times. Then, at some time $t_1 \leq n$, the optimal arm is drawn for the $(T+1)$-th time, hence $B_{1,T,t_1} \geq B_{2,T_2(t_1-1),t_1}$. Now, since $V_{1,T} = 0$ and $\overline{X}_{1,T} = 0$, we have

$$B_{1,T,t_1} = \frac{3c\zeta b \log(t_1)}{T} \leq \frac{3c\zeta b}{\gamma} \leq pb\varepsilon,$$

where in the last inequality we used (54) and the definition of γ. Further, $B_{2,T_2(t_1-1),t_1} = pb\varepsilon + 3c\zeta \log(t_1)/T_2(t_1-1) > pb\varepsilon$, hence we get the desired contradiction. □

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

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

Further, when this event holds the regret is at least $(n-T)pb(1-\varepsilon)$. Thus, the expected regret is at least $(1-p)pb(1-\varepsilon)n^{1-\varepsilon}(1-\gamma(\log n)/n)$, which is indeed polynomial in n since $1-\varepsilon > 0$. □

A.5. Proof of Theorem 11

PROOF. Without the loss of generality (by a scaling argument), we may assume that $b = 1$. We prove the theorem by first proving three claims.

Claim: Consider the event $\mathcal{A}$ on which

$$\forall s \geq 7 \quad \forall k \in \{1, \ldots, K\} \quad \begin{cases} \left|\overline{X}_{k,s} - \mu_k\right| < \sigma_k \sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{\mathcal{E}_s}{3s} \\ \sigma_k \leq \sqrt{V_{k,s}} + 1.8\sqrt{\frac{\mathcal{E}_s}{s}} \\ \sqrt{V_{k,s}} \leq \sigma_k + \sqrt{\frac{\mathcal{E}_s}{2s}} \end{cases} \tag{55}$$

This event holds with probability at least $1-\beta$.

PROOF. The arguments that we will use to prove the first two inequalities are similar to the ones used in the proof of Theorem 1. The main difference here is that we want the third inequality to hold simultaneously with the first two inequalities. We apply Lemma 5 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 $k \in \{1, \ldots, K\}$, 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.3, the first two of these inequalities imply the first two inequalities of (55). The last inequality of (55) is obtained by taking the square root in the above third inequality.

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 $B_{k,s} \triangleq \overline{X}_{k,s} + \sqrt{\frac{2V_{k,s}\mathcal{E}_s}{s}} + \frac{3\mathcal{E}_s}{s}$. We have the following claim:
Claim: On the event $\mathcal{A}$, for any $s \geq 1$ and $k \in \{1,\dots,K\}$, the following inequalities hold:

$$\mu_k \leq B_{k,s}, \tag{56}$$

$$B_{k,s} \leq \mu_k + 2\sigma_k\sqrt{\frac{2\mathcal{E}_s}{s}} + \frac{13\mathcal{E}_s}{3s} \tag{57}$$

Proof. Inequality (56) is obtained by plugging the second inequality of (55) into the first one of (55) and by noting that since $\mathcal{E}_s \geq 2$, (56) is trivial for $s \leq 6$. Introduce $L_s = \frac{\mathcal{E}_s}{s}$. To prove (57), we use the first and third inequalities of (55) to obtain

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

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

Claim: The choice of u_k in Theorem 11 guarantees that

$$\mu_k + 2\sigma_k\sqrt{\frac{2\mathcal{E}_{u_k}}{u_k}} + \frac{13\mathcal{E}_{u_k}}{3u_k} < \mu^*. \tag{58}$$

Proof. For the sake of compactness, for a moment we drop the arm indices, so that (58) is equivalent to

$$2\sigma\sqrt{\frac{2\mathcal{E}_u}{u}} + \frac{13\mathcal{E}_u}{3u} < \Delta. \tag{59}$$

Let $r = u/\mathcal{E}_u$. Given that $r \geq 0$, we have

$$\begin{array}{rcll} (59) & \Leftrightarrow & r - \frac{13}{3\Delta} > \frac{2\sigma}{\Delta}\sqrt{2r} & \\ & \Leftrightarrow & r > \frac{13}{3\Delta} & \text{and} \quad \big(r - \frac{13}{3\Delta}\big)^2 > \frac{8\sigma^2}{\Delta^2}r \\ & \Leftrightarrow & r > \frac{13}{3\Delta} & \text{and} \quad r^2 - \big(\frac{8\sigma^2}{\Delta^2} + \frac{26}{3\Delta}\big)r + \frac{169}{9\Delta^2} > 0 \end{array}$$

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

□

Let $\mathcal{B} = \{\exists k : T_k(\infty) > u_k\}$ be the event that arm k is pulled more than u_k times. By adapting the argument used in the proof Theorem 2 to prove (13) one can show that

$$\mathcal{B} \subset \big\{\exists k \text{ s.t. } B_{k,u_k} > \tau\big\} \bigcup \big\{\exists s \geq 1 \text{ s.t. } B_{k^*,s} \leq \tau\big\}.$$

Taking $\tau = \mu^*$ and using (58), (56) and (57), we get

$$\mathcal{B} \subset \big\{\exists k \text{ s.t. } B_{k,u_k} > \mu_k + 2\sigma_k\sqrt{\tfrac{2\mathcal{E}_{u_k}}{u_k}} + \tfrac{13\mathcal{E}_{u_k}}{3u_k}\big\} \bigcup \big\{\exists s \geq 1 \text{ s.t. } B_{k^*,s} \leq \mu^*\big\} \subset \overline{\mathcal{A}},$$

where $\overline{\mathcal{A}}$ denotes the complementer of $\mathcal{A}$. Taking probabilities we get $\mathbb{P}(\mathcal{B}) \leq \mathbb{P}(\overline{\mathcal{A}}) \leq \beta$, thus finishing the proof. □

A.6. Proof of Theorem 12

Proof. Consider the following sequence $\tilde{\mathcal{E}}_s = \log[Kn(s+1)]$ for $s \leq n$ and $\tilde{\mathcal{E}}_s = \infty$ 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 $(\tilde{\mathcal{E}}_s)_{s\geq 0}$ 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

$$T_k(n) \leq 1 + \big(\tfrac{8\sigma_k^2}{\Delta_k^2} + \tfrac{26b}{3\Delta_k}\big)\log[KnT_k(n)], \tag{60}$$

which gives the first assertion.

For the second assertion, first note that since $R_n \leq n$, (38) is non-trivial only when $30\log n < n$. So the bound is trivial when $n \leq 100$. Besides, from the first assertion of Theorem 2, we have $T_k(n) = 1$ if $K \geq n$, in which case (38) is trivial. For $n > 100$ and $K < n$, (60) gives

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

hence

$$\mathbb{E}[T_k(n)] \leq 4\log(n/7) + 1 + \big(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{26b}{\Delta_k}\big)\log n \leq \big(\tfrac{24\sigma_k^2}{\Delta_k^2} + \tfrac{30b}{\Delta_k}\big)\log n.$$

□

References

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

[2] J.-Y. Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. http://certis.enpc.fr/~audibert/ThesePack.zip.

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

[4] P. Auer, N. Cesa-Bianchi, and J. Shawe-Taylor. Exploration versus exploitation challenge. In *2nd PASCAL Challenges Workshop*. Pascal Network, 2006.

[5] D.A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

[6] S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in monte-carlo go. Technical report, INRIA RR-6062, 2006.

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

[8] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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

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

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

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

[13] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.