Machine Learning, 47, 235–256, 2002

Finite-time Analysis of the Multiarmed Bandit Problem*

PETER AUER pauer@igi.tu-graz.ac.at
University of Technology Graz, A-8010 Graz, Austria

NICOLÒ CESA-BIANCHI cesa-bianchi@dti.unimi.it
DTI, University of Milan, via Bramante 65, I-26013 Crema, Italy

PAUL FISCHER fischer@ls2.informatik.uni-dortmund.de
Lehrstuhl Informatik II, Universität Dortmund, D-44221 Dortmund, Germany

Editor: Jyrki Kivinen

Abstract. Reinforcement learning policies face the exploration versus exploitation dilemma, i.e. the search for a balance between exploring the environment to find profitable actions while taking the empirically best action as often as possible. A popular measure of a policy's success in addressing this dilemma is the regret, that is the loss due to the fact that the globally optimal policy is not followed all the times. One of the simplest examples of the exploration/exploitation dilemma is the multi-armed bandit problem. Lai and Robbins were the first ones to show that the regret for this problem has to grow at least logarithmically in the number of plays. Since then, policies which asymptotically achieve this regret have been devised by Lai and Robbins and many others. In this work we show that the optimal logarithmic regret is also achievable uniformly over time, with simple and efficient policies, and for all reward distributions with bounded support.

Keywords: bandit problems, adaptive allocation rules, finite horizon regret

1. Introduction

The exploration versus exploitation dilemma can be described as the search for a balance between exploring the environment to find profitable actions while taking the empirically best action as often as possible. The simplest instance of this dilemma is perhaps the multi-armed bandit, a problem extensively studied in statistics (Berry & Fristedt, 1985) that has also turned out to be fundamental in different areas of artificial intelligence, such as reinforcement learning (Sutton & Barto, 1998) and evolutionary programming (Holland, 1992).

In its most basic formulation, a K-armed bandit problem is defined by random variables $X_{i,n}$ for $1 \leq i \leq K$ and $n \geq 1$, where each i is the index of a gambling machine (i.e., the "arm" of a bandit). Successive plays of machine i yield rewards $X_{i,1}, X_{i,2}, \ldots$ which are

*A preliminary version appeared in *Proc. of 15th International Conference on Machine Learning, pages 100–108. Morgan Kaufmann, 1998*

independent and identically distributed according to an unknown law with unknown expectation μ_i. Independence also holds for rewards across machines; i.e., $X_{i,s}$ and $X_{j,t}$ are independent (and usually not identically distributed) for each $1 \le i < j \le K$ and each $s, t \ge 1$.

A *policy*, or *allocation strategy*, A is an algorithm that chooses the next machine to play based on the sequence of past plays and obtained rewards. Let $T_i(n)$ be the number of times machine i has been played by A during the first n plays. Then the *regret* of A after n plays is defined by

$$\mu^* n - \mu_j \sum_{j=1}^{K} \mathbb{E}[T_j(n)] \quad \text{where} \quad \mu^* \stackrel{\text{def}}{=} \max_{1 \le i \le K} \mu_i$$

and $\mathbb{E}[\cdot]$ denotes expectation. Thus the regret is the expected loss due to the fact that the policy does not always play the best machine.

In their classical paper, Lai and Robbins (1985) found, for specific families of reward distributions (indexed by a single real parameter), policies satisfying

$$\mathbb{E}[T_j(n)] \le \left(\frac{1}{D(p_j \| p^*)} + o(1) \right) \ln n \tag{1}$$

where $o(1) \to 0$ as $n \to \infty$ and

$$D(p_j \| p^*) \stackrel{\text{def}}{=} \int p_j \ln \frac{p_j}{p^*}$$

is the Kullback-Leibler divergence between the reward density p_j of any suboptimal machine j and the reward density p^* of the machine with highest reward expectation μ^*. Hence, under these policies the optimal machine is played exponentially more often than any other machine, at least asymptotically. Lai and Robbins also proved that this regret is the best possible. Namely, for any allocation strategy and for any suboptimal machine j, $\mathbb{E}[T_j(n)] \ge (\ln n)/D(p_j \| p^*)$ asymptotically, provided that the reward distributions satisfy some mild assumptions.

These policies work by associating a quantity called *upper confidence index* to each machine. The computation of this index is generally hard. In fact, it relies on the entire sequence of rewards obtained so far from a given machine. Once the index for each machine is computed, the policy uses it as an estimate for the corresponding reward expectation, picking for the next play the machine with the current highest index. More recently, Agrawal (1995) introduced a family of policies where the index can be expressed as simple function of the total reward obtained so far from the machine. These policies are thus much easier to compute than Lai and Robbins', yet their regret retains the optimal logarithmic behavior (though with a larger leading constant in some cases).[1]

In this paper we strengthen previous results by showing policies that achieve logarithmic regret uniformly over time, rather than only asymptotically. Our policies are also simple to implement and computationally efficient. In Theorem 1 we show that a simple variant of Agrawal's index-based policy has finite-time regret logarithmically bounded for arbitrary sets of reward distributions with bounded support (a regret with better constants is proven

in Theorem 2 for a more complicated version of this policy). A similar result is shown in Theorem 3 for a variant of the well-known randomized ε-greedy heuristic. Finally, in Theorem 4 we show another index-based policy with logarithmically bounded finite-time regret for the natural case when the reward distributions are normally distributed with unknown means and variances.

Throughout the paper, and whenever the distributions of rewards for each machine are understood from the context, we define

$$\Delta_i \stackrel{\text{def}}{=} \mu^* - \mu_i$$

where, we recall, μ_i is the reward expectation for machine i and μ^* is any maximal element in the set $\{\mu_1, \ldots, \mu_K\}$.

2. Main results

Our first result shows that there exists an allocation strategy, UCB1, achieving logarithmic regret uniformly over n and without any preliminary knowledge about the reward distributions (apart from the fact that their support is in $[0, 1]$). The policy UCB1 (sketched in figure 1) is derived from the index-based policy of Agrawal (1995). The index of this policy is the sum of two terms. The first term is simply the current average reward. The second term is related to the size (according to Chernoff-Hoeffding bounds, see Fact 1) of the one-sided confidence interval for the average reward within which the true expected reward falls with overwhelming probability.

Theorem 1. *For all $K > 1$, if policy* UCB1 *is run on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then its expected regret after any number n of plays is at most*

$$\left[8 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln n}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, \ldots, \mu_K$ are the expected values of $P_1, \ldots, P_K$.

Deterministic policy: UCB1.
Initialization: Play each machine once.
Loop:

- Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

Figure 1. Sketch of the deterministic policy UCB1 (see Theorem 1).

Deterministic policy: UCB2.
Parameters: $0 < \alpha < 1$.
Initialization: Set $r_j = 0$ for $j = 1, \ldots, K$. Play each machine once.
Loop:

1. Select machine j maximizing $\bar{x}_j + a_{n,r_j}$, where $\bar{x}_j$ is the average reward obtained from machine j, a_{n,r_j} is defined in (3), and n is the overall number of plays done so far.
2. Play machine j exactly $\tau(r_j + 1) - \tau(r_j)$ times.
3. Set $r_j \leftarrow r_j + 1$.

Figure 2. Sketch of the deterministic policy UCB2 (see Theorem 2).

To prove Theorem 1 we show that, for any suboptimal machine j,

$$I\!E[T_j(n)] \leq \frac{8}{\Delta_j^2} \ln n \tag{2}$$

plus a small constant. The leading constant $8/\Delta_i^2$ is worse than the corresponding constant $1/D(p_j \parallel p^*)$ in Lai and Robbins' result (1). In fact, one can show that $D(p_j \parallel p^*) \geq 2\Delta_j^2$ where the constant 2 is the best possible.

Using a slightly more complicated policy, which we call UCB2 (see figure 2), we can bring the main constant of (2) arbitrarily close to $1/(2\Delta_j^2)$. The policy UCB2 works as follows.

The plays are divided in epochs. In each new epoch a machine i is picked and then played $\tau(r_i + 1) - \tau(r_i)$ times, where τ is an exponential function and r_i is the number of epochs played by that machine so far. The machine picked in each new epoch is the one maximizing $\bar{x}_i + a_{n,r_i}$, where n is the current number of plays, $\bar{x}_i$ is the current average reward for machine i, and

$$a_{n,r} = \sqrt{\frac{(1+\alpha)\ln(en/\tau(r))}{2\tau(r)}} \tag{3}$$

where

$$\tau(r) = \lceil (1+\alpha)^r \rceil.$$

In the next result we state a bound on the regret of UCB2. The constant c_α, here left unspecified, is defined in (18) in the appendix, where the theorem is also proven.

Theorem 2. *For all $K > 1$, if policy* UCB2 *is run with input $0 < \alpha < 1$ on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in* $[0, 1]$, *then its expected regret after any number*

$$n \geq \max_{i:\mu_i < \mu^*} \frac{1}{2\Delta_i^2}$$

of plays is at most

$$\sum_{i\,:\,\mu_i<\mu^*}\left(\frac{(1+\alpha)(1+4\alpha)\ln\left(2e\Delta_i^2 n\right)}{2\Delta_i}+\frac{c_\alpha}{\Delta_i}\right) \tag{4}$$

where $\mu_1,\ldots,\mu_K$ *are the expected values of* $P_1,\ldots,P_K$.

Remark. By choosing α small, the constant of the leading term in the sum (4) gets arbitrarily close to $1/(2\Delta_i^2)$; however, $c_\alpha\to\infty$ as $\alpha\to 0$. The two terms in the sum can be traded-off by letting $\alpha=\alpha_n$ be slowly decreasing with the number n of plays.

A simple and well-known policy for the bandit problem is the so-called ε-greedy rule (see Sutton, & Barto, 1998). This policy prescribes to play with probability $1-\varepsilon$ the machine with the highest average reward, and with probability ε a randomly chosen machine. Clearly, the constant exploration probability ε causes a linear (rather than logarithmic) growth in the regret. The obvious fix is to let ε go to zero with a certain rate, so that the exploration probability decreases as our estimates for the reward expectations become more accurate. It turns out that a rate of $1/n$, where n is, as usual, the index of the current play, allows to prove a logarithmic bound on the regret. The resulting policy, ε_n-GREEDY, is shown in figure 3.

Theorem 3. *For all* $K>1$ *and for all reward distributions* $P_1,\ldots,P_K$ *with support in* $[0,1]$, *if policy* ε_n-GREEDY *is run with input parameter*

$$0<d\le\min_{i:\mu_i<\mu^*}\Delta_i,$$

Randomized policy: ε_n-GREEDY.
Parameters: $c>0$ and $0<d<1$.
Initialization: Define the sequence $\varepsilon_n\in(0,1]$, $n=1,2,\ldots$, by

$$\varepsilon_n\stackrel{\text{def}}{=}\min\left\{1,\frac{cK}{d^2n}\right\}$$

Loop: For each $n=1,2,\ldots$

- Let i_n be the machine with the highest current average reward.
- With probability $1-\varepsilon_n$ play i_n and with probability ε_n play a random arm.

Figure 3. Sketch of the randomized policy ε_n-GREEDY (see Theorem 3).

*then the probability that after any number $n \geq cK/d$ of plays ε_n-*GREEDY *chooses a suboptimal machine j is at most*

$$\frac{c}{d^2 n} + 2\left(\frac{c}{d^2}\ln\frac{(n-1)d^2e^{1/2}}{cK}\right)\left(\frac{cK}{(n-1)d^2e^{1/2}}\right)^{c/(5d^2)} + \frac{4e}{d^2}\left(\frac{cK}{(n-1)d^2e^{1/2}}\right)^{c/2}.$$

Remark. For c large enough (e.g. $c > 5$) the above bound is of order $c/(d^2 n) + o(1/n)$ for $n \to \infty$, as the second and third terms in the bound are $O(1/n^{1+\varepsilon})$ for some $\varepsilon > 0$ (recall that $0 < d < 1$). Note also that this is a result stronger than those of Theorems 1 and 2, as it establishes a bound on the instantaneous regret. However, unlike Theorems 1 and 2, here we need to know a lower bound d on the difference between the reward expectations of the best and the second best machine.

Our last result concerns a special case, i.e. the bandit problem with normally distributed rewards. Surprisingly, we could not find in the literature regret bounds (not even asymptotical) for the case when both the mean and the variance of the reward distributions are unknown. Here, we show that an index-based policy called UCB1-NORMAL, see figure 4, achieves logarithmic regret uniformly over n without knowing means and variances of the reward distributions. However, our proof is based on certain bounds on the tails of the χ^2 and the Student distribution that we could only verify numerically. These bounds are stated as Conjecture 1 and Conjecture 2 in the Appendix.

The choice of the index in UCB1-NORMAL is based, as for UCB1, on the size of the one-sided confidence interval for the average reward within which the true expected reward falls with overwhelming probability. In the case of UCB1, the reward distribution was unknown, and we used Chernoff-Hoeffding bounds to compute the index. In this case we know that

Deterministic policy: UCB1-NORMAL.
Loop: For each $n = 1, 2, \ldots$

- If there is a machine which has been played less then $\lceil 8 \log n \rceil$ times then play this machine.
- Otherwise play machine j that maximizes

$$\bar{x}_j + \sqrt{16 \cdot \frac{q_j - n_j \bar{x}_j^2}{n_j - 1} \cdot \frac{\ln(n-1)}{n_j}}$$

  where $\bar{x}_j$ is the average reward obtained from machine j, q_j is the sum of squared rewards obtained from machine j, and n_j is the number of times machine j has been played so far.
- Update $\bar{x}_j$ and q_j with the obtained reward x_j.

Figure 4. Sketch of the deterministic policy UCB1-NORMAL (see Theorem 4).

the distribution is normal, and for computing the index we use the sample variance as an estimate of the unknown variance.

Theorem 4. *For all $K > 1$, if policy* UCB1-NORMAL *is run on K machines having normal reward distributions $P_1, \ldots, P_K$, then its expected regret after any number n of plays is at most*

$$256(\log n)\left(\sum_{i:\mu_i<\mu^*} \frac{\sigma_i^2}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{2} + 8\,\log n\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, \ldots, \mu_K$ and $\sigma_1^2, \ldots, \sigma_K^2$ are the means and variances of the distributions $P_1, \ldots, P_K$.

As a final remark for this section, note that Theorems 1–3 also hold for rewards that are not independent across machines, i.e. $X_{i,s}$ and $X_{j,t}$ might be dependent for any s, t, and $i \neq j$. Furthermore, we also do not need that the rewards of a single arm are i.i.d., but only the weaker assumption that $\mathbb{E}[X_{i,t} \mid X_{i,1}, \ldots, X_{i,t-1}] = \mu_i$ for all $1 \leq t \leq n$.

3. Proofs

Recall that, for each $1 \leq i \leq K$, $\mathbb{E}[X_{i,n}] = \mu_i$ for all $n \geq 1$ and $\mu^* = \max_{1\leq i\leq K} \mu_i$. Also, for any fixed policy A, $T_i(n)$ is the number of times machine i has been played by A in the first n plays. Of course, we always have $\sum_{i=1}^{K} T_i(n) = n$. We also define the r.v.'s $I_1, I_2, \ldots$, where I_t denotes the machine played at time t.

For each $1 \leq i \leq K$ and $n \geq 1$ define

$$\bar{X}_{i,n} = \frac{1}{n}\sum_{t=1}^{n} X_{i,t}.$$

Given $\mu_1, \ldots, \mu_K$, we call *optimal* the machine with the least index i such that $\mu_i = \mu^*$. In what follows, we will always put a superscript "$*$" to any quantity which refers to the optimal machine. For example we write $T^*(n)$ and $\bar{X}_n^*$ instead of $T_i(n)$ and $\bar{X}_{i,n}$, where i is the index of the optimal machine.

Some further notation: For any predicate Π we define $\{\Pi(x)\}$ to be the indicator fuction of the event $\Pi(x)$; i.e., $\{\Pi(x)\} = 1$ if $\Pi(x)$ is true and $\{\Pi(x)\} = 0$ otherwise. Finally, $\mathrm{Var}[X]$ denotes the variance of the random variable X.

Note that the regret after n plays can be written as

$$\sum_{j:\mu_j<\mu^*} \Delta_j\, \mathbb{E}[T_j(n)] \tag{5}$$

So we can bound the regret by simply bounding each $\mathbb{E}[T_j(n)]$.

We will make use of the following standard exponential inequalities for bounded random variables (see, e.g., the appendix of Pollard, 1984).

Fact 1 (Chernoff-Hoeffding bound)**.** *Let $X_1, \ldots, X_n$ be random variables with common range* $[0, 1]$ *and such that* $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] = \mu$. *Let* $S_n = X_1 + \cdots + X_n$. *Then for all* $a \geq 0$

$$\mathbb{P}\{S_n \geq n\mu + a\} \leq e^{-2a^2/n} \quad \text{and} \quad \mathbb{P}\{S_n \leq n\mu - a\} \leq e^{-2a^2/n}$$

Fact 2 (Bernstein inequality)**.** *Let $X_1, \ldots, X_n$ be random variables with range* $[0, 1]$ *and*

$$\sum_{t=1}^{n} \mathrm{Var}[X_t \mid X_{t-1}, \ldots, X_1] = \sigma^2.$$

Let $S_n = X_1 + \cdots + X_n$. *Then for all* $a \geq 0$

$$\mathbb{P}\{S_n \geq \mathbb{E}[S_n] + a\} \leq \exp\left\{-\frac{a^2/2}{\sigma^2 + a/2}\right\}.$$

Proof of Theorem 1: Let $c_{t,s} = \sqrt{(2 \ln t)/s}$. For any machine i, we upper bound $T_i(n)$ on any sequence of plays. More precisely, for each $t \geq 1$ we bound the indicator function of $I_t = i$ as follows. Let ℓ be an arbitrary positive integer.

$$\begin{aligned}
T_i(n) &= 1 + \sum_{t=K+1}^{n} \{I_t = i\} \\
&\leq \ell + \sum_{t=K+1}^{n} \{I_t = i,\ T_i(t-1) \geq \ell\} \\
&\leq \ell + \sum_{t=K+1}^{n} \Big\{\bar{X}^*_{T^*(t-1)} + c_{t-1,T^*(t-1)} \leq \bar{X}_{i,T_i(t-1)} \\
&\qquad\qquad + c_{t-1,T_i(t-1)},\ T_i(t-1) \geq \ell\Big\} \\
&\leq \ell + \sum_{t=K+1}^{n} \left\{ \min_{0<s<t} \bar{X}^*_s + c_{t-1,s} \leq \max_{\ell \leq s_i < t} \bar{X}_{i,s_i} + c_{t-1,s_i} \right\} \\
&\leq \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} \left\{\bar{X}^*_s + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i}\right\}. \qquad (6)
\end{aligned}$$

Now observe that $\bar{X}^*_s + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i}$ implies that at least one of the following must hold

$$\bar{X}^*_s \leq \mu^* - c_{t,s} \qquad (7)$$

$$\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \qquad (8)$$

$$\mu^* < \mu_i + 2c_{t,s_i}. \qquad (9)$$

We bound the probability of events (7) and (8) using Fact 1 (Chernoff-Hoeffding bound)

$$\mathbb{P}\{\bar{X}^*_s \leq \mu^* - c_{t,s}\} \leq e^{-4 \ln t} = t^{-4}$$

$$I\!P\left\{\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i}\right\} \leq e^{-4\ln t} = t^{-4}.$$

For $\ell = \lceil (8\ln n)/\Delta_i^2 \rceil$, (9) is false. In fact

$$\mu^* - \mu_i - 2c_{t,s_i} = \mu^* - \mu_i - 2\sqrt{2(\ln t)/s_i} \geq \mu^* - \mu_i - \Delta_i = 0$$

for $s_i \geq (8\ln n)/\Delta_i^2$. So we get

$$\begin{aligned} I\!E[T_i(n)] &\leq \left\lceil \frac{8\ln n}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=\lceil (8\ln n)/\Delta_i^2 \rceil}^{t-1} \\ &\quad \times \left(I\!P\{\bar{X}_s^* \leq \mu^* - c_{t,s}\} + I\!P\left\{\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i}\right\}\right) \\ &\leq \left\lceil \frac{8\ln n}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t}\sum_{s_i=1}^{t} 2t^{-4} \\ &\leq \frac{8\ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \end{aligned}$$

which concludes the proof. □

Proof of Theorem 3: Recall that, for $n \geq cK/d^2$, $\varepsilon_n = cK/(d^2 n)$. Let

$$x_0 = \frac{1}{2K}\sum_{t=1}^{n} \varepsilon_t.$$

The probability that machine j is chosen at time n is

$$I\!P\{I_n = j\} \leq \frac{\varepsilon_n}{K} + \left(1 - \frac{\varepsilon_n}{K}\right) I\!P\left\{\bar{X}_{j,T_j(n-1)} \geq \bar{X}^*_{T^*(n-1)}\right\}$$

and

$$\begin{aligned} &I\!P\left\{\bar{X}_{j,T_j(n)} \geq \bar{X}^*_{T^*(n)}\right\} \\ &\leq I\!P\left\{\bar{X}_{j,T_j(n)} \geq \mu_j + \frac{\Delta_j}{2}\right\} + I\!P\left\{\bar{X}^*_{T^*(n)} \leq \mu^* - \frac{\Delta_j}{2}\right\}. \end{aligned} \tag{10}$$

Now the analysis for both terms on the right-hand side is the same. Let $T_j^R(n)$ be the number of plays in which machine j was chosen *at random* in the first n plays. Then we have

$$\begin{aligned} &I\!P\left\{\bar{X}_{j,T_j(n)} \geq \mu_j + \frac{\Delta_j}{2}\right\} \\ &= \sum_{t=1}^{n} I\!P\left\{T_j(n) = t \wedge \bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} \end{aligned} \tag{11}$$

$$
\begin{aligned}
&= \sum_{t=1}^{n} I\!P\left\{T_j(n) = t \mid \bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} \cdot I\!P\left\{\bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} \\
&\leq \sum_{t=1}^{n} I\!P\left\{T_j(n) = t \mid \bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} \cdot e^{-\Delta_j^2 t/2} \\
&\quad \text{by Fact 1 (Chernoff-Hoeffding bound)} \\
&\leq \sum_{t=1}^{\lfloor x_0 \rfloor} I\!P\left\{T_j(n) = t \mid \bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} \\
&\quad \text{since } \textstyle\sum_{t=x+1}^{\infty} e^{-\kappa t} \leq \frac{1}{\kappa} e^{-\kappa x} \\
&\leq \sum_{t=1}^{\lfloor x_0 \rfloor} I\!P\left\{T_j^R(n) \leq t \mid \bar{X}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}\right\} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} \\
&\leq x_0 \cdot I\!P\left\{T_j^R(n) \leq x_0\right\} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} \qquad (12)
\end{aligned}
$$

where in the last line we dropped the conditioning because each machine is played at random independently of the previous choices of the policy. Since

$$I\!E\left[T_j^R(n)\right] = \frac{1}{K} \sum_{t=1}^{n} \varepsilon_t$$

and

$$\mathrm{Var}\left[T_j^R(n)\right] = \sum_{t=1}^{n} \frac{\varepsilon_t}{K}\left(1 - \frac{\varepsilon_t}{K}\right) \leq \frac{1}{K} \sum_{t=1}^{n} \varepsilon_t,$$

by Bernstein's inequality (2) we get

$$I\!P\left\{T_j^R(n) \leq x_0\right\} \leq e^{-x_0/5}. \qquad (13)$$

Finally it remains to lower bound x_0. For $n \geq n' = cK/d^2$, $\varepsilon_n = cK/(d^2 n)$ and we have

$$
\begin{aligned}
x_0 &= \frac{1}{2K} \sum_{t=1}^{n} \varepsilon_t \\
&= \frac{1}{2K} \sum_{t=1}^{n'} \varepsilon_t + \frac{1}{2K} \sum_{t=n'+1}^{n} \varepsilon_t \\
&\geq \frac{n'}{2K} + \frac{c}{d^2} \ln \frac{n}{n'} \\
&\geq \frac{c}{d^2} \ln \frac{n d^2 e^{1/2}}{cK}.
\end{aligned}
$$

Thus, using (10)–(13) and the above lower bound on x_0 we obtain

$$\begin{aligned}\mathbb{P}\{I_n = j\} &\leq \frac{\varepsilon_n}{K} + 2x_0 e^{-x_0/5} + \frac{4}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} \\ &\leq \frac{c}{d^2 n} + 2\left(\frac{c}{d^2} \ln \frac{(n-1)d^2 e^{1/2}}{cK}\right)\left(\frac{cK}{(n-1)d^2 e^{1/2}}\right)^{c/(5d^2)} \\ &\quad + \frac{4e}{d^2}\left(\frac{cK}{(n-1)d^2 e^{1/2}}\right)^{c/2}.\end{aligned}$$

This concludes the proof. □

4. Experiments

For practical purposes, the bound of Theorem 1 can be tuned more finely. We use

$$V_j(s) \stackrel{\text{def}}{=} \left(\frac{1}{s}\sum_{\tau=1}^{s} X_{j,\tau}^2\right) - \bar{X}_{j,s}^2 + \sqrt{\frac{2\ln t}{s}}$$

as un upper confidence bound for the variance of machine j. As before, this means that machine j, which has been played s times during the first t plays, has a variance that is at most the sample variance plus $\sqrt{(2\ln t)/s}$. We then replace the upper confidence bound $\sqrt{2\ln(n)/n_j}$ of policy UCB1 with

$$\sqrt{\frac{\ln n}{n_j}\min\{1/4, V_j(n_j)\}}$$

(the factor $1/4$ is an upper bound on the variance of a Bernoulli random variable). This variant, which we call UCB1-TUNED, performs substantially better than UCB1 in essentially all of our experiments. However, we are not able to prove a regret bound.

We compared the empirical behaviour policies UCB1-TUNED, UCB2, and ε_n-GREEDY on Bernoulli reward distributions with different parameters shown in the table below.

	1	2	3	4	5	6	7	8	9	10
1	0.9	0.6								
2	0.9	0.8								
3	0.55	0.45								
11	0.9	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
12	0.9	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6
13	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
14	0.55	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45

Rows 1–3 define reward distributions for a 2-armed bandit problem, whereas rows 11–14 define reward distributions for a 10-armed bandit problem. The entries in each row denote the reward expectations (i.e. the probabilities of getting a reward 1, as we work with Bernoulli distributions) for the machines indexed by the columns. Note that distributions 1 and 11 are "easy" (the reward of the optimal machine has low variance and the differences $\mu^* - \mu_i$ are all large), whereas distributions 3 and 14 are "hard" (the reward of the optimal machine has high variance and some of the differences $\mu^* - \mu_i$ are small).

We made experiments to test the different policies (or the same policy with different input parameters) on the seven distributions listed above. In each experiment we tracked two performance measures: (1) the percentage of plays of the optimal machine; (2) the actual regret, that is the difference between the reward of the optimal machine and the reward of the machine played. The plot for each experiment shows, on a semi-logarithmic scale, the behaviour of these quantities during 100,000 plays averaged over 100 different runs. We ran a first round of experiments on distribution 2 to find out good values for the parameters of the policies. If a parameter is chosen too small, then the regret grows linearly (exponentially in the semi-logarithmic plot); if a parameter is chosen too large then the regret grows logarithmically, but with a large leading constant (corresponding to a steep line in the semi-logarithmic plot).

Policy UCB2 is relatively insensitive to the choice of its parameter α, as long as it is kept relatively small (see figure 5). A fixed value 0.001 has been used for all the remaining experiments. On other hand, the choice of c in policy ε_n-GREEDY is difficult as there is no value that works reasonably well for all the distributions that we considered. Therefore, we have roughly searched for the best value for each distribution. In the plots, we will also show the performance of ε_n-GREEDY for values of c around this empirically best value. This shows that the performance degrades rapidly if this parameter is not appropriately tuned. Finally, in each experiment the parameter d of ε_n-GREEDY was set to

$$\Delta = \mu^* - \max_{i:\mu_i<\mu^*} \mu_i.$$

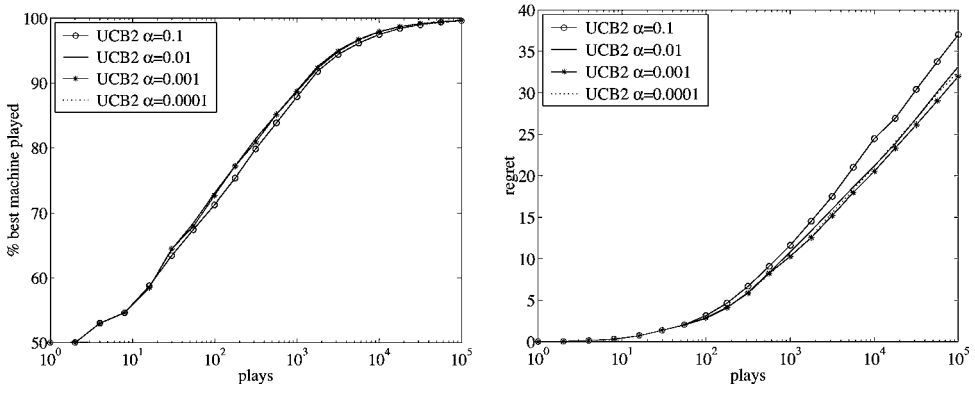

Figure 5. Search for the best value of parameter α of policy UCB2.

4.1. *Comparison between policies*

We can summarize the comparison of all the policies on the seven distributions as follows (see Figs. 6–12).

- An optimally tuned ε_n-GREEDY performs almost always best. Significant exceptions are distributions 12 and 14: this is because ε_n-GREEDY explores uniformly over all machines, thus the policy is hurt if there are several nonoptimal machines, especially when their reward expectations differ a lot. Furthermore, if ε_n-GREEDY is not well tuned its performance degrades rapidly (except for distribution 13, on which ε_n-GREEDY performs well a wide range of values of its parameter).
- In most cases, UCB1-TUNED performs comparably to a well-tuned ε_n-GREEDY. Furthermore, UCB1-TUNED is not very sensitive to the variance of the machines, that is why it performs similarly on distributions 2 and 3, and on distributions 13 and 14.
- Policy UCB2 performs similarly to UCB1-TUNED, but always slightly worse.

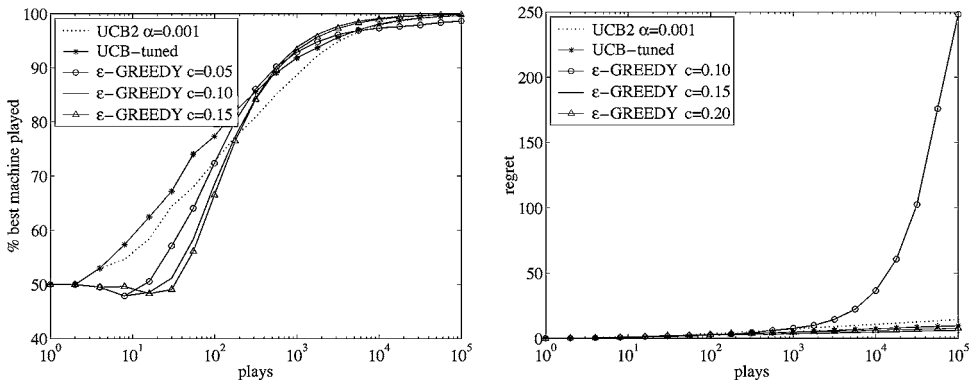

Figure 6. Comparison on distribution 1 (2 machines with parameters 0.9, 0.6).

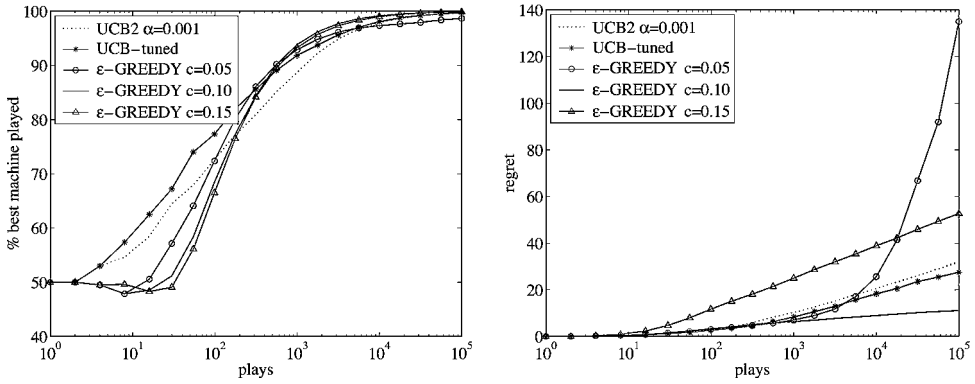

Figure 7. Comparison on distribution 2 (2 machines with parameters 0.9, 0.8).

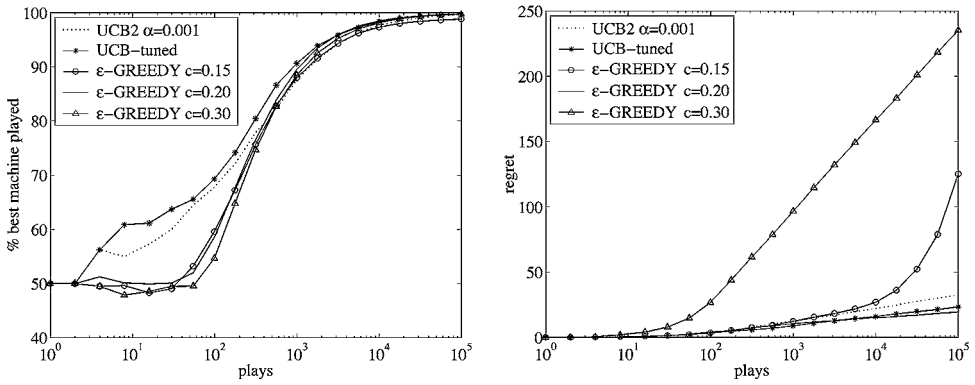

Figure 8. Comparison on distribution 3 (2 machines with parameters 0.55, 0.45).

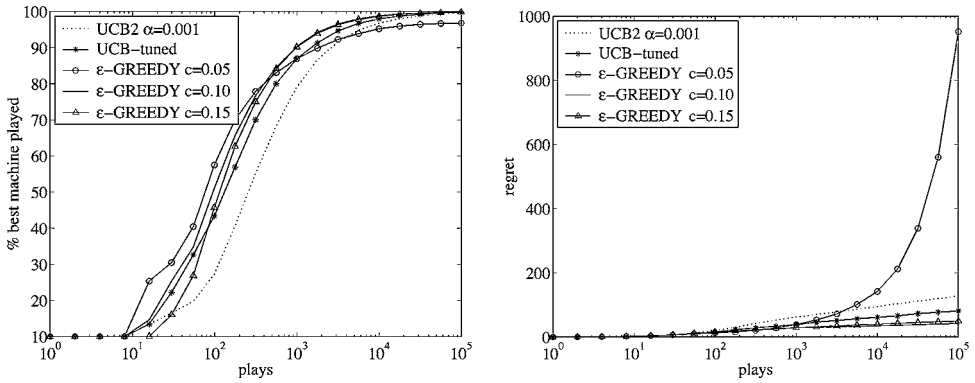

Figure 9. Comparison on distribution 11 (10 machines with parameters 0.9, 0.6, . . . , 0.6).

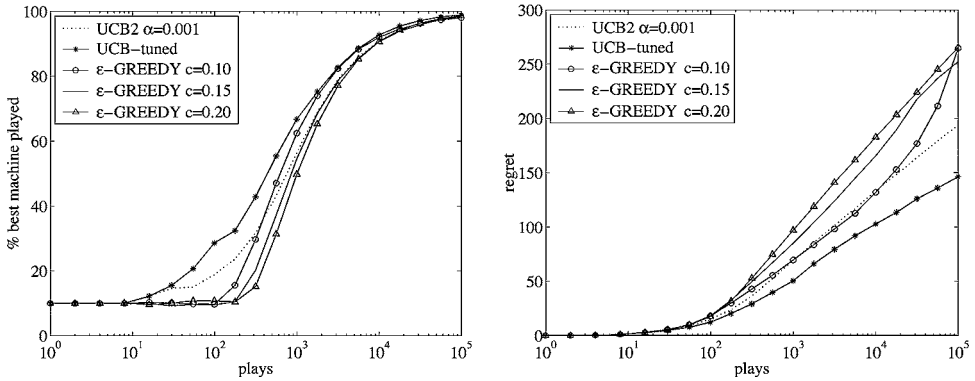

Figure 10. Comparison on distribution 12 (10 machines with parameters 0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6, 0.6).

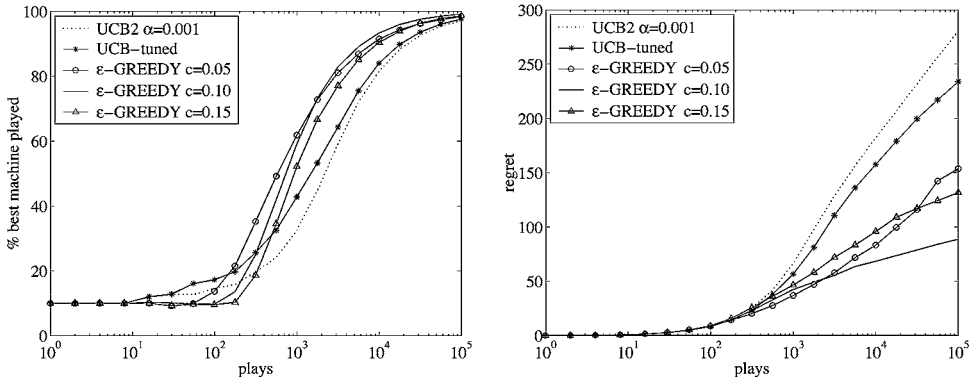

Figure 11. Comparison on distribution 13 (10 machines with parameters 0.9, 0.8, . . . , 0.8).

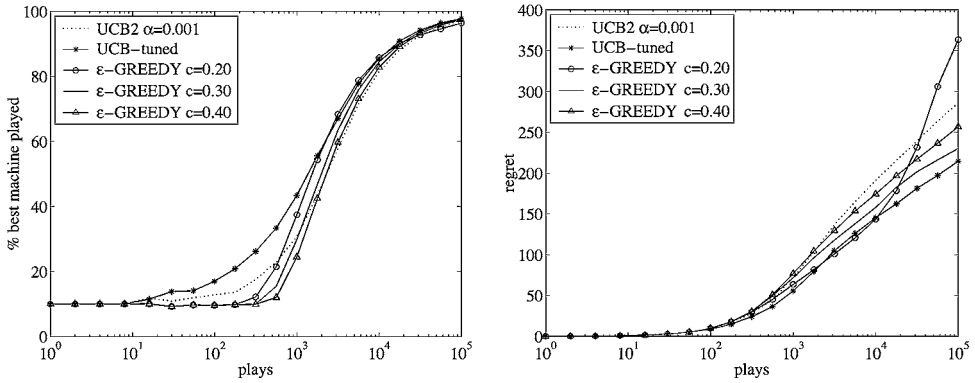

Figure 12. Comparison on distribution 14 (10 machines with parameters 0.55, 0.45, . . . , 0.45).

5. Conclusions

We have shown simple and efficient policies for the bandit problem that, on any set of reward distributions with known bounded support, exhibit uniform logarithmic regret. Our policies are deterministic and based on upper confidence bounds, with the exception of ε_n-GREEDY, a randomized allocation rule that is a dynamic variant of the ε-greedy heuristic. Moreover, our policies are robust with respect to the introduction of moderate dependencies in the reward processes.

This work can be extended in many ways. A more general version of the bandit problem is obtained by removing the stationarity assumption on reward expectations (see Berry & Fristedt, 1985; Gittins, 1989 for extensions of the basic bandit problem). For example, suppose that a stochastic reward process $\{X_{i,s} : s = 1, 2, \ldots\}$ is associated to each machine $i = 1, \ldots, K$. Here, playing machine i at time t yields a reward $X_{i,s}$ and causes the current

state s of i to change to $s+1$, whereas the states of other machines remain frozen. A well-studied problem in this setup is the maximization of the total expected reward in a sequence of n plays. There are methods, like the Gittins allocation indices, that allow to find the optimal machine to play at each time n by considering each reward process independently from the others (even though the globally optimal solution depends on all the processes). However, computation of the Gittins indices for the average (undiscounted) reward criterion used here requires preliminary knowledge about the reward processes (see, e.g., Ishikida & Varaiya, 1994). To overcome this requirement, one can learn the Gittins indices, as proposed in Duff (1995) for the case of finite-state Markovian reward processes. However, there are no finite-time regret bounds shown for this solution. At the moment, we do not know whether our techniques could be extended to these more general bandit problems.

Appendix A: Proof of Theorem 2

Note that

$$\tau(r) \leq (1+\alpha)^r + 1 \leq \tau(r-1)(1+\alpha) + 1 \tag{14}$$

for $r \geq 1$. Assume that $n \geq 1/(2\Delta_j^2)$ for all j and let $\tilde{r}_j$ be the largest integer such that

$$\tau(\tilde{r}_j - 1) \leq \frac{(1+4\alpha)\ln\left(2en\Delta_j^2\right)}{2\Delta_j^2}.$$

Note that $\tilde{r}_j \geq 1$. We have

$$\begin{aligned} T_j(n) &\leq 1 + \sum_{r\geq 1}(\tau(r) - \tau(r-1)) \text{ \{machine } j \text{ finishes its } r\text{-th epoch\}} \\ &\leq \tau(\tilde{r}_j) + \sum_{r>\tilde{r}_j}(\tau(r) - \tau(r-1)) \text{ \{machine } j \text{ finishes its } r\text{-th epoch\}} \end{aligned}$$

Now consider the following chain of implications

$$\begin{aligned} &\text{machine } j \text{ finishes its } r\text{-th epoch} \\ &\Rightarrow \exists i \geq 0,\ \exists t \geq \tau(r-1) + \tau(i) \text{ such that} \\ &\quad \left(\bar{X}_{j,\tau(r-1)} + a_{t,r-1}\right) \geq \left(\bar{X}^*_{\tau(i)} + a_{t,i}\right) \\ &\Rightarrow \exists t \geq \tau(r-1) \text{ such that } \left(\bar{X}_{j,\tau(r-1)} + a_{t,r-1}\right) \geq \mu^* - \alpha\Delta_j/2 \\ &\quad \text{or} \quad \exists i \geq 0,\ \exists t' \geq \tau(r-1) + \tau(i) \text{ such that} \\ &\quad \left(\bar{X}^*_{\tau(i)} + a_{t',i}\right) \leq \mu^* - \alpha\Delta_j/2 \\ &\Rightarrow \bar{X}_{j,\tau(r-1)} + a_{n,r-1} \geq \mu^* - \alpha\Delta_j/2 \\ &\quad \text{or} \quad \exists i \geq 0 \text{ such that } \bar{X}^*_{\tau(i)} + a_{\tau(r-1)+\tau(i),i} \leq \mu^* - \alpha\Delta_j/2 \end{aligned}$$

where the last implication hold because $a_{t,r}$ is increasing in t. Hence

$$I\!E[T_j(n)] \leq \tau(\tilde{r}_j) + \sum_{r>\tilde{r}_j}(\tau(r) - \tau(r-1))\, I\!P\left\{\bar{X}_{j,\tau(r-1)} + a_{n,r-1} \geq \mu^* - \alpha\Delta_j/2\right\}$$

$$+ \sum_{r>\tilde{r}_j} \sum_{i\geq 0} (\tau(r) - \tau(r-1))$$
$$\cdot \; I\!P\left\{\bar{X}^*_{\tau(i)} + a_{\tau(r-1)+\tau(i),i} \leq \mu^* - \alpha\Delta_j/2\right\}. \tag{15}$$

The assumption $n \geq 1/(2\Delta_j^2)$ implies $\ln(2en\Delta_j^2) \geq 1$. Therefore, for $r > \tilde{r}_j$, we have

$$\tau(r-1) > \frac{(1+4\alpha)\ln\left(2en\Delta_j^2\right)}{2\Delta_j^2} \tag{16}$$

and

$$\begin{aligned}
a_{n,r-1} &= \sqrt{\frac{(1+\alpha)\ln(en/\tau(r-1))}{2\tau(r-1)}} \\
&\leq \Delta_j \sqrt{\frac{(1+\alpha)\ln(en/\tau(r-1))}{\left(1+4\alpha\right)\ln(2en\Delta_j^2)}} \quad \text{using (16) above} \\
&\leq \Delta_j \sqrt{\frac{\left(1+\alpha\right)\ln(2en\Delta_j^2)}{\left(1+4\alpha\right)\ln(2en\Delta_j^2)}} \\
&\quad \text{using } \tau(r-1) > 1/2\Delta_j^2 \text{ derived from (16)} \\
&\leq \Delta_j \sqrt{\frac{1+\alpha}{1+4\alpha}}.
\end{aligned} \tag{17}$$

We start by bounding the first sum in (15). Using (17) and Fact 1 (Chernoff-Hoeffding bound) we get

$$\begin{aligned}
&I\!P\left\{\bar{X}_{j,\tau(r-1)} + a_{n,r-1} \geq \mu^* - \alpha\Delta_j/2\right\} \\
&= I\!P\{\bar{X}_{j,\tau(r-1)} + a_{n,r-1} \geq \mu_j + \Delta_j - \alpha\Delta_j/2\} \\
&\leq \exp\Big\{-2\tau(r-1)\Delta_j^2\big(1-\alpha/2-\sqrt{(1+\alpha)/(1+4\alpha)}\big)^2\Big\} \\
&\leq \exp\left\{-2\tau(r-1)\Delta_j^2\,(1-\alpha/2-(1-\alpha))^2\right\} \\
&= \exp\left\{-\tau(r-1)\Delta_j^2\alpha^2/2\right\}
\end{aligned}$$

for $\alpha < 1/10$. Now let $g(x) = (x-1)/(1+\alpha)$. By (14) we get $g(x) \leq \tau(r-1)$ for $\tau(r-1) \leq x \leq \tau(r)$ and $r \geq 1$. Hence

$$\begin{aligned}
&\sum_{r>\tilde{r}_j} (\tau(r) - \tau(r-1)) I\!P\left\{\bar{X}_{j,\tau(r-1)} + a_{n,r-1} \geq \mu^* - \alpha\Delta_j/2\right\} \\
&\leq \sum_{r>\tilde{r}_j} (\tau(r) - \tau(r-1)) \exp\Big\{-\tau(r-1)\Delta_j^2\alpha^2\Big\} \\
&\leq \int_0^\infty e^{-cg(x)}dx
\end{aligned}$$

where $c = (\Delta_j \alpha)^2 < 1$. Further manipulation yields

$$\int_0^\infty \exp\left\{-\frac{c}{1+\alpha}(x-1)\right\} dx = e^{c/(1+\alpha)} \frac{1+\alpha}{c} \le \frac{(1+\alpha)e}{(\Delta_j \alpha)^2}.$$

We continue by bounding the second sum in (15). Using once more Fact 1, we get

$$\begin{aligned}
&\sum_{r>\tilde{r}_j} \sum_{i\ge 0} (\tau(r) - \tau(r-1)) I\!P\left\{\bar{X}^*_{\tau(i)} + a_{\tau(r-1)+\tau(i),i} \le \mu^* - \alpha\Delta_j/2\right\} \\
&\le \sum_{i\ge 0} \sum_{r>\tilde{r}_j} (\tau(r) - \tau(r-1)) \\
&\quad \cdot \exp\left\{ -\tau(i)\frac{(\alpha\Delta_j)^2}{2} - (1+\alpha)\ln\left(e\frac{\tau(r-1)+\tau(i)}{\tau(i)}\right)\right\} \\
&\le \sum_{i\ge 0} \exp\{-\tau(i)(\alpha\Delta_j)^2/2\} \\
&\quad \cdot \left[\sum_{r>\tilde{r}_j} (\tau(r) - \tau(r-1)) \exp\left\{ -(1+\alpha)\ln\left(1+\frac{\tau(r-1)}{\tau(i)}\right)\right\}\right] \\
&= \sum_{i\ge 0} \exp\{-\tau(i)(\alpha\Delta_j)^2/2\} \\
&\quad \cdot \left[\sum_{r>\tilde{r}_j} (\tau(r) - \tau(r-1))\left(1+\frac{\tau(r-1)}{\tau(i)}\right)^{-(1+\alpha)}\right] \\
&\le \sum_{i\ge 0} \exp\{-\tau(i)(\alpha\Delta_j)^2/2\}\left[\int_0^\infty \left(1+\frac{x-1}{(1+\alpha)\tau(i)}\right)^{-(1+\alpha)} dx\right] \\
&= \sum_{i\ge 0} \tau(i)\exp\{-\tau(i)(\alpha\Delta_j)^2/2\}\left[\frac{1+\alpha}{\alpha}\left(1-\frac{1}{(1+\alpha)\tau(i)}\right)^{-\alpha}\right] \\
&\le \sum_{i\ge 0} \tau(i)\exp\{-\tau(i)(\alpha\Delta_j)^2/2\}\left[\frac{1+\alpha}{\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{-\alpha}\right] \quad \text{as } \tau(i) \ge 1 \\
&= \left(\frac{1+\alpha}{\alpha}\right)^{1+\alpha} \sum_{i\ge 0} \tau(i)\exp\{-\tau(i)(\alpha\Delta_j)^2/2\}.
\end{aligned}$$

Now, as $(1+\alpha)^{x-1} \le \tau(i) \le (1+\alpha)^x + 1$ for $i \le x \le i+1$, we can bound the series in the last formula above with an integral

$$\begin{aligned}
&\sum_{i\ge 0} \tau(i)\exp\{-\tau(i)(\alpha\Delta_j)^2/2\} \\
&\le 1 + \int_1^\infty ((1+\alpha)^x + 1)\exp\{-(1+\alpha)^{x-1}(\alpha\Delta_j)^2/2\}\, dx
\end{aligned}$$

$$\leq 1 + \int_1^\infty \frac{z+1}{z\ln(1+\alpha)} \exp\left\{ -\frac{z(\alpha\Delta_j)^2}{2(1+\alpha)} \right\} dz$$

by change of variable $z = (1+\alpha)^x$

$$= 1 + \frac{1}{\ln(1+\alpha)} \left[\frac{e^{-\lambda}}{\lambda} + \int_\lambda^\infty \frac{e^{-x}}{x}\, dx \right]$$

where we set

$$\lambda = \frac{(\alpha\Delta_j)^2}{2(1+\alpha)}.$$

As $0 < \alpha, \Delta_j < 1$, we have $0 < \lambda < 1/4$. To upper bound the bracketed formula above, consider the function

$$F(\lambda) = e^{-\lambda} + \lambda \int_\lambda^\infty \frac{e^{-x}}{x}\, dx$$

with derivatives

$$F'(\lambda) = \int_\lambda^\infty \frac{e^{-x}}{x}\, dx - 2e^{-\lambda} \qquad F''(\lambda) = 2\lambda e^{-\lambda} - \int_\lambda^\infty \frac{e^{-x}}{x}\, dx.$$

In the interval $(0, 1/4)$, F' is seen to have a zero at $\lambda = 0.0108\ldots$. As $F''(\lambda) < 0$ in the same interval, this is the unique maximum of F, and we find $F(0.0108\ldots) < 11/10$. So we have

$$\frac{e^{-\lambda}}{\lambda} + \int_\lambda^\infty \frac{e^{-x}}{x}\, dx < \frac{11}{10\lambda} = \frac{11(1+\alpha)}{5(\alpha\Delta_j)^2}$$

Piecing everything together, and using (14) to upper bound $\tau(\tilde{r}_j)$, we find that

$$\begin{aligned} \mathbb{E}[T_j(n)] &\leq \tau(\tilde{r}_j) + \frac{(1+\alpha)e}{(\Delta_j\alpha)^2} + \left(\frac{1+\alpha}{\alpha}\right)^{1+\alpha} \left[1 + \frac{11(1+\alpha)}{5(\alpha\Delta_j)^2 \ln(1+\alpha)} \right] \\ &\leq \frac{(1+\alpha)(1+4\alpha)\ln\left(2en\Delta_j^2\right)}{2\Delta_j^2} + \frac{c_\alpha}{\Delta_j^2} \end{aligned}$$

where

$$c_\alpha = 1 + \frac{(1+\alpha)e}{\alpha^2} + \left(\frac{1+\alpha}{\alpha}\right)^{1+\alpha} \left[1 + \frac{11(1+\alpha)}{5\alpha^2\ln(1+\alpha)} \right]. \tag{18}$$

This concludes the proof. □

Appendix B: Proof of Theorem 4

The proof goes very much along the same lines as the proof of Theorem 1. It is based on the two following conjectures which we only verified numerically.

Conjecture 1. Let X be a Student random variable with s degrees of freedom. Then, for all $0 \le a \le \sqrt{2(s+1)}$,

$$I\!P\{X \ge a\} \le e^{-a^2/4}.$$

Conjecture 2. Let X be a χ^2 random variable with s degrees of freedom. Then

$$I\!P\{X \ge 4s\} \le e^{-(s+1)/2}.$$

We now proceed with the proof of Theorem 4. Let

$$Q_{i,n} = \sum_{t=1}^{n} X_{i,t}^2.$$

Fix a machine i and, for any s and t, set

$$c_{t,s} = \sqrt{16 \cdot \frac{Q_{i,s} - s\bar{X}_{i,s}^2}{s-1} \cdot \frac{\ln t}{s}}$$

Let $c_{t,s}^*$ be the corresponding quantity for the optimal machine. To upper bound $T_i(n)$, we proceed exactly as in the first part of the proof of Theorem 1 obtaining, for any positive integer ℓ,

$$T_i(n) \le \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} \left[\{\bar{X}_s^* \le \mu^* - c_{t,s}^*\} + \left\{\bar{X}_{i,s_i} \ge \mu_i + c_{t,s_i}\right\} + \left\{\mu^* < \mu_i + 2c_{t,s_i}\right\} \right].$$

The random variable $(\bar{X}_{i,s_i} - \mu_i)/\sqrt{(Q_{i,s_i} - s_i\bar{X}_{i,s_i}^2)/(s_i(s_i-1))}$ has a Student distribution with $s_i - 1$ degrees of freedom (see, e.g., Wilks, 1962, 8.4.3 page 211). Therefore, using Conjecture 1 with $s = s_i - 1$ and $a = 4\sqrt{\ln t}$, we get

$$I\!P\left\{\bar{X}_{i,s_i} \ge \mu_i + c_{t,s_i}\right\} = I\!P\left\{ \frac{\bar{X}_{i,s_i} - \mu_i}{\sqrt{\left(Q_{i,s_i} - s_i\bar{X}_{i,s_i}^2\right)/(s_i(s_i-1))}} \ge 4\sqrt{\ln t} \right\} \le t^{-4}$$

for all $s_i \ge 8 \ln t$. The probability of $\bar{X}_s^* \le \mu^* - c_{t,s}^*$ is bounded analogously. Finally, since $(Q_{i,s_i} - s_i\bar{X}_{i,s_i}^2)/\sigma_i^2$ is χ^2-distributed with $s_i - 1$ degrees of freedom (see, e.g., Wilks, 1962,

8.4.1 page 208). Therefore, using Conjecture 2 with $s = s_i - 1$ and $a = 4s$, we get

$$\begin{aligned} I\!P\left\{\mu^* < \mu_i + 2c_{t,s_i}\right\} &= I\!P\left\{\left(Q_{i,s_i} - s_i \bar{X}_{i,s_i}^2\right)/\sigma_i^2 > (s_i - 1)\frac{\Delta_i^2}{\sigma_i^2}\frac{s_i}{64\ln t}\right\} \\ &\leq I\!P\left\{\left(Q_{i,s_i} - s_i \bar{X}_{i,s_i}^2\right)/\sigma_i^2 > 4(s_i - 1)\right\} \\ &\leq e^{-s_i/2} \leq t^{-4} \end{aligned}$$

for

$$s_i \geq \max\left\{256\frac{\sigma_i^2}{\Delta_i^2}, 8\right\}\ln t.$$

Setting

$$\ell = \left\lceil \max\left\{256\frac{\sigma_i^2}{\Delta_i^2}, 8\right\}\ln t \right\rceil$$

completes the proof of the theorem. □

Acknowledgments

The support from ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II), is gratefully acknowledged.

Note

1. Similar extensions of Lai and Robbins' results were also obtained by Yakowitz and Lowe (1991), and by Burnetas and Katehakis (1996).

References

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

Berry, D., & Fristedt, B. (1985). *Bandit problems*. London: Chapman and Hall.

Burnetas, A., & Katehakis, M. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, *17:2*, 122–142.

Duff, M. (1995). Q-learning for bandit problems. In *Proceedings of the 12th International Conference on Machine Learning* (pp. 209–217).

Gittins, J. (1989). *Multi-armed bandit allocation indices*, Wiley-Interscience series in Systems and Optimization. New York: John Wiley and Sons.

Holland, J. (1992). *Adaptation in natural and artificial systems*. Cambridge: MIT Press/Bradford Books.

Ishikida, T., & Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications*, *83:1*, 113–154.

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

Pollard, D. (1984). *Convergence of stochastic processes*. Berlin: Springer.

Sutton, R., & Barto, A. (1998). *Reinforcement learning, an introduction*. Cambridge: MIT Press/Bradford Books.

Wilks, S. (1962). *Matematical statistics*. New York: John Wiley and Sons.

Yakowitz, S., & Lowe, W. (1991). Nonparametric bandit methods. *Annals of Operations Research*, *28*, 297–312.

Received September 29, 2000
Revised May 21, 2001
Accepted June 20, 2001
Final manuscript June 20, 2001