UCB REVISITED: IMPROVED REGRET BOUNDS FOR THE STOCHASTIC MULTI-ARMED BANDIT PROBLEM

PETER AUER AND RONALD ORTNER

ABSTRACT. In the stochastic multi-armed bandit problem we consider a modification of the UCB algorithm of Auer et al. [4]. For this modified algorithm we give an improved bound on the regret with respect to the optimal reward. While for the original UCB algorithm the regret in K-armed bandits after T trials is bounded by $const \cdot \frac{K\log(T)}{\Delta}$, where Δ measures the distance between a suboptimal arm and the optimal arm, for the modified UCB algorithm we show an upper bound on the regret of $const \cdot \frac{K\log(T\Delta^2)}{\Delta}$.

1. INTRODUCTION

In the stochastic multi-armed bandit problem, a learner has to choose in trials $t = 1, 2, \ldots$ an *arm* from a given set A of $K := |A|$ arms. In each trial t the learner obtains random reward $r_{i,t} \in [0, 1]$ for choosing arm i. It is assumed that for each arm i the random rewards $r_{i,t}$ are independent and identically distributed random variables with mean r_i which is unknown to the learner. Further, it is assumed that the rewards $r_{i,t}$ and $r_{j,t'}$ for distinct arms i, j are independent for all $i \neq j \in A$ and all $t, t' \in \mathbb{N}$. The learner's aim is to compete with the arm giving highest mean reward $r^* := \max_{i\in A} r_i$.

When the learner has played each arm at least once, he faces the so-called *exploration vs. exploitation dilemma*: Shall he stick to an arm that gave high reward so far (*exploitation*) or rather probe other arms further (*exploration*)?

Date: August 2, 2011.

2000 *Mathematics Subject Classification.* 68T05, 62M05, 91A60.

Key words and phrases. multi-armed bandit problem, regret.

The authors would like to thank an anonymous COLT reviewer as well as Philippe Rigollet for pointing out errors in earlier versions of this paper. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 216886, the PASCAL2 Network of Excellence, and n° 216529, Personal Information Navigator Adapting Through Viewing, PinView, and the Austrian Federal Ministry of Science and Research. This publication only reflects the authors' views.

When exploiting the best arm so far, the learner takes the risk that the arm with the highest mean reward is currently underestimated. On the other hand, exploration may simply waste time with playing suboptimal arms. The multi-armed bandit problem is considered to be the simplest instance of this dilemma, that also appears in more general *reinforcement learning* problems such as learning in Markov decision processes [11]. As the multi-armed bandit and its variants also have applications as diverse as routing in networks, experiment design, pricing, and placing ads on webpages, to name a few (for references and further applications see e.g. [8]), the problem has attracted attention in areas like statistics, economics, and computer science.

The seminal work of Lai and Robbins [9] introduced the idea of using *upper confidence values* for dealing with the exploration-exploitation dilemma in the multi-armed bandit problem. The arm with the best estimate $\hat{r}^*$ so far serves as a benchmark, and other arms are played only if the upper bound of a suitable confidence interval is at least $\hat{r}^*$. That way, within T trials each suboptimal arm can be shown to be played at most $\left(\frac{1}{D_{\mathrm{KL}}} + o(1)\right) \log T$ times in expectation, where D_{KL} measures the distance between the reward distributions of the optimal and the suboptimal arm by the Kullback-Leibler divergence, and $o(1) \to 0$ as $T \to \infty$. This bound was also shown to be asymptotically optimal [9].

The original algorithm suggested by Lai and Robbins considers the whole history for computing the arm to choose. Only later, their method was simplified by Agrawal [1]. Also for this latter approach the optimal asymptotic bounds given by Lai and Robbins remain valid, yet with a larger leading constant in some cases.

More recently, Auer et al. [4] introduced the simple, yet efficient UCB algorithm, that is also based on the ideas of Lai and Robbins [9]. After playing each arm once for initialization, UCB chooses at trial t the arm i that maximizes[1]

$$\hat{r}_i + \sqrt{\frac{2\log t}{n_i}}, \tag{1}$$

where $\hat{r}_i$ is the average reward obtained from arm i, and n_i is the number of times arm i has been played up to trial t. The value in (1) can be interpreted as the upper bound of a confidence interval, so that the true mean reward of each arm i with high probability is below this *upper confidence bound*.

[1]Subsequently, $\log$ denotes the natural logarithm, while e stands for its base, i.e., Euler's number.

In particular, the upper confidence value of the optimal arm will be higher than the true optimal mean reward r^* with high probability. Consequently, as soon as a suboptimal arm i has been played sufficiently often so that the length of the confidence interval $\sqrt{\frac{2\log t}{n_i}}$ is small enough to guarantee that

$$\hat{r}_i + \sqrt{\frac{2\log t}{n_i}} < r^*,$$

arm i will not be played anymore with high probability. As it also holds that with high probability

$$\hat{r}_i < r_i + \sqrt{\frac{2\log t}{n_i}},$$

arm i is not played as soon as

$$2\sqrt{\frac{2\log t}{n_i}} < r^* - r_i,$$

that is, as soon as arm i has been played

$$\left\lceil \frac{8\log t}{(r^* - r_i)^2} \right\rceil$$

times. This informal argument can be made stringent to show that each suboptimal arm i in expectation will not be played more often than

$$const \cdot \frac{\log T}{\Delta_i^2} \tag{2}$$

times within T trials, where $\Delta_i := r^* - r_i$ is the distance between the optimal mean reward and r_i. Unlike the bounds of Lai and Robbins [9] and Agrawal [1] this bound holds uniformly over time, and not only asymptotically.

1.1. **Comparison to the nonstochastic setting.** Beside the number of times a suboptimal arm is chosen, another common measure for the quality of a bandit algorithm is the *regret* the algorithm suffers with respect to the optimal arm. That is, we define the (expected) *regret* of an algorithm after T trials as

$$r^*T - \sum_{i\in A} r_i \,\mathbb{E}\left[N_i\right],$$

where N_i denotes the number of times the algorithm chooses arm i within the first T trials. In view of (2), the expected regret of UCB after T trials can be upper bounded by

$$\sum_{i:r_i<r^*} const \cdot \frac{\log T}{\Delta_i}, \tag{3}$$

as choosing arm i once suffers an expected regret of Δ_i with respect to r^*.

In the different setting of *nonstochastic bandits* [5], the learner has to deal with arbitrary reward sequences that for example may be chosen by an adversary. For the setting where the learner competes with the best arm, Auer et al. [5] gave the algorithm **Exp4** whose regret with respect to the best arm is of order $\sqrt{KT \log K}$.

When comparing the two different bounds for the stochastic and the non-stochastic bandit, it strikes odd that when choosing

$$\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}}$$

for all suboptimal arms i in the stochastic setting, the upper bound on the regret of (3) gives

$$\frac{\log T}{\sqrt{\log K}} \sqrt{KT}.$$

This is worse than the bound in the nonstochastic setting, so that one may conclude that the bounds in the stochastic setting are improvable. Recently, this has been confirmed by an upper bound of order $\sqrt{KT}$ for the algorithm MOSS [2] in the stochastic setting.

Further, this is consistent with the lower bounds on the regret derived by Mannor and Tsitsiklis [10]. For the case where all arms except the optimal arm have the same mean reward (so that all distances Δ_i coincide as above), the regret is lower bounded by

$$const \cdot K \cdot \frac{\log\left(\frac{T\Delta^2}{K}\right)}{\Delta}.$$

In this paper we present a modification of the UCB algorithm, for which we prove an upper bound on the regret of

$$\sum_{i: r_i < r^*} const \cdot \frac{\log(T\Delta_i^2)}{\Delta_i}.$$

Compared to the regret bound for the original UCB algorithm, this bound gives an improvement in particular for arms whose reward is close to the optimum.

2. UCB IMPROVED

We first consider the simpler case when the learner knows the horizon T. The unknown horizon case is dealt with in Section 4 below. We first note that if the learner had access to the values Δ_i, one could directly modify the

Input: A set of arms A, the horizon T.
Initialization: Set $\tilde{\Delta}_0 := 1$, and $B_0 := A$.
For rounds $m = 0, 1, 2, \ldots, \left\lfloor \frac{1}{2} \log_2 \frac{T}{e} \right\rfloor$ **do:**

Arm selection:
If $|B_m| > 1$, choose each arm in B_m until the total number of times it has been chosen is $n_m := \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$. Otherwise choose the single arm in B_m until step T is reached.

Arm elimination:
Delete all arms i from B_m for which

$$\left\{ \hat{r}_i + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} \right\} < \max_{j \in B} \left\{ \hat{r}_j - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} \right\}$$

in order to obtain B_{m+1}. Here $\hat{r}_j$ is the average reward obtained from arm j.

Reset $\tilde{\Delta}_m$**:**
Set $\tilde{\Delta}_{m+1} := \frac{\tilde{\Delta}_m}{2}$.

FIGURE 1. The improved UCB algorithm.

confidence intervals of UCB as given in (1) to $\sqrt{\frac{2\log(t\Delta_i^2)}{n_i}}$, and the proof of the claimed regret bound would be straightforward.

However, since the Δ_i are unknown to the learner, the modified algorithm shown in Figure 1 guesses the values Δ_i by a value $\tilde{\Delta}$, which is initialized to 1 and halved each time the confidence intervals become shorter than $\tilde{\Delta}$. Note that compared to the original UCB algorithm the confidence intervals are shorter, in particular for arms with high estimated reward. Unlike the original UCB algorithm, our modification eliminates arms that perform bad. As the analysis will show, each suboptimal arm is eliminated as soon as $\tilde{\Delta} < \frac{\Delta_i}{2}$, provided that the confidence intervals hold. Similar arm elimination algorithms were already proposed in [6]. However, the analysis of [6] concentrated on PAC bounds for identifying an optimal arm instead of regret bounds as in our case.

3. ONLINE REGRET BOUNDS FOR THE IMPROVED UCB ALGORITHM

Now we show the following improved bound on the expected regret.

Theorem 3.1. *The total expected regret of the improved* UCB *algorithm up to trial T is upper bounded by*

$$\sum_{i\in A:\Delta_i>\lambda}\left(\Delta_i+\frac{32\log(T\Delta_i^2)}{\Delta_i}+\frac{96}{\Delta_i}\right)+\sum_{i\in A:0<\Delta_i\leq\lambda}\frac{64}{\lambda}+\max_{i\in A:\Delta_i\leq\lambda}\Delta_i T$$

for all $\lambda\geq\sqrt{\frac{\mathrm{e}}{T}}$.

Remark 3.2. It is easy to see that the logarithmic term is the main term for suitable λ. For example, setting $\lambda:=\sqrt{\frac{\mathrm{e}}{T}}$, the term $\max_{i\in A:\Delta_i\leq\lambda}\Delta_i T$ is trivially bounded by $\sqrt{\mathrm{e}T}$, which is $\leq\frac{\mathrm{e}}{\Delta_i}$ for $\Delta_i\leq\lambda$.

Remark 3.3. For $\lambda\approx\sqrt{\frac{K\log K}{T}}$ the terms $\frac{K\log(T\lambda^2)}{\lambda}$ and λT of the bound in Theorem 3.1 coincide apart from a factor of $\log\log K$. The regret in this case is bounded according to Theorem 3.1 by

$$\sqrt{KT}\cdot\frac{\log(K\log K)}{\sqrt{\log K}},$$

which apart from the factor $\log K$ in the logarithm corresponds to the bound in the nonstochastic setting. Still, there is room for further improvement as the already mentioned bound of $\sqrt{KT}$ for the MOSS algorithm shows [2].

Proof of Theorem 3.1: In the following we use $*$ to indicate an arbitrary optimal arm. Further, for each suboptimal arm i let $m_i:=\min\{m\,|\,\tilde{\Delta}_m<\frac{\Delta_i}{2}\}$ be the first round in which $\tilde{\Delta}_m<\frac{\Delta_i}{2}$. Note that by definition of $\tilde{\Delta}_m$ and m_i we have

$$2^{m_i}\;=\;\frac{1}{\tilde{\Delta}_{m_i}}\;\leq\;\frac{4}{\Delta_i}\;<\;\frac{1}{\tilde{\Delta}_{m_i+1}}\;=\;2^{m_i+1}. \tag{4}$$

We consider suboptimal arms in $A':=\{i\in A\,|\,\Delta_i>\lambda\}$ for some fixed $\lambda\geq\sqrt{\frac{\mathrm{e}}{T}}$, and analyze the regret in the following cases:

Case (a): *Some suboptimal arm i is not eliminated in round m_i (or before) with an optimal arm $*\in B_{m_i}$.*

Let us consider an arbitrary suboptimal arm i. First note that if

$$\hat{r}_i\leq r_i+\sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} \tag{5}$$

and

$$\hat{r}_*\geq r^*-\sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} \tag{6}$$

hold for $m = m_i$, then under the assumption that $*, i \in B_{m_i}$ arm i will be eliminated in round m_i. Indeed, in the elimination phase of round m_i we have by (4) that $\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}} \leq \frac{\tilde{\Delta}_{m_i}}{2} = \tilde{\Delta}_{m_i+1} < \frac{\Delta_i}{4}$, so that by (5) and (6)

$$\begin{aligned}
\hat{r}_i + \sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}} &\leq r_i + 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}} \\
&< r_i + \Delta_i - 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}} = r_* - 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}} \\
&\leq \hat{r}_* - \sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}},
\end{aligned}$$

and arm i is eliminated as claimed. Now by Chernoff-Hoeffding bounds [7] for each $m = 0, 1, 2, \ldots$

$$\mathbb{P}\left\{\hat{r}_i > r_i + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right\} \leq \frac{1}{T\tilde{\Delta}_m^2}, \tag{7}$$

and

$$\mathbb{P}\left\{\hat{r}_* < r^* - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right\} \leq \frac{1}{T\tilde{\Delta}_m^2}, \tag{8}$$

so that the probability that a suboptimal arm i is *not* eliminated in round m_i (or before) is bounded by $\frac{2}{T\tilde{\Delta}_{m_i}^2}$. Summing up over all arms in A' and bounding the regret for each arm i trivially by $T\Delta_i$ we obtain by (4) a contribution of

$$\sum_{i \in A'} \frac{2\Delta_i}{\tilde{\Delta}_{m_i}^2} \leq \sum_{i \in A'} \frac{8}{\tilde{\Delta}_{m_i}} \leq \sum_{i \in A'} \frac{32}{\Delta_i}$$

to the expected regret.

Case (b): *For each suboptimal arm i: either i is eliminated in round m_i (or before) or there is no optimal arm $*$ in B_{m_i}.*

Case (b1): If an optimal arm $* \in B_{m_i}$ for all arms i in A', then each arm i in A' is eliminated in round m_i (or before) and consequently played not more often than

$$n_{m_i} = \left\lceil \frac{2\log(T\tilde{\Delta}_{m_i}^2)}{\tilde{\Delta}_{m_i}^2} \right\rceil \leq \left\lceil \frac{32\log(T\frac{\Delta_i^2}{4})}{\Delta_i^2} \right\rceil \tag{9}$$

times, giving a contribution of

$$\sum_{i\in A'} \Delta_i \left\lceil \frac{32\log(T\frac{\Delta_i^2}{4})}{\Delta_i^2} \right\rceil < \sum_{i\in A'} \left(\Delta_i + \frac{32\log(T\Delta_i^2)}{\Delta_i} \right)$$

to the expected regret.

Case (b2): Now let us consider the case that the last remaining optimal arm $*$ is eliminated by some suboptimal arm i in $A'' := \{i \in A | \Delta_i > 0\}$ in some round m_*. First note that if (5) and (6) hold in round $m = m_*$, then the optimal arm will not be eliminated by arm i in this round. Indeed, this would only happen if

$$\hat{r}_i - \sqrt{\frac{\log(T\tilde{\Delta}_{m_*}^2)}{2n_{m_*}}} > \hat{r}_* + \sqrt{\frac{\log(T\tilde{\Delta}_{m_*}^2)}{2n_{m_*}}},$$

which however leads by (5) and (6) to the contradiction $r_i > r^*$. Consequently, by (7) and (8) the probability that $*$ is eliminated by a fixed suboptimal arm i in round m_* is upper bounded by $\frac{2}{T\tilde{\Delta}_{m_*}^2}$.

Now if $*$ is eliminated by arm i in round m_*, then $* \in B_{m_j}$ for all j with $m_j < m_*$. Hence by assumption of case (b), all arms j with $m_j < m_*$ were eliminated in round m_j (or before). Consequently, $*$ can only be eliminated in round m_* by an arm i with $m_i \geq m_*$. Further, the maximal regret per step after eliminating $*$ is the maximal Δ_j among the remaining arms j with $m_j \geq m_*$. Let $m_\lambda := \min\{m \,|\, \tilde{\Delta}_m < \frac{\lambda}{2}\}$. Then, taking into account the error probability for elimination of $*$ by *some* arm in A'', the contribution to the expected regret in the considered case is upper bounded by

$$\begin{aligned}
&\sum_{m_*=0}^{\max_{j\in A'} m_j} \sum_{i\in A'':\, m_i\geq m_*} \frac{2}{T\tilde{\Delta}_{m_*}^2} \cdot T \max_{j\in A'':m_j\geq m_*} \Delta_j \\
\leq\ & \sum_{m_*=0}^{\max_{j\in A'} m_j} \sum_{i\in A'':\, m_i\geq m_*} \frac{2}{\tilde{\Delta}_{m_*}^2} \cdot 4\tilde{\Delta}_{m_*} \\
\leq\ & \sum_{i\in A''} \sum_{m_*=0}^{\min\{m_i,m_\lambda\}} \frac{8}{\tilde{\Delta}_{m_*}} = \sum_{i\in A''} \sum_{m_*=0}^{\min\{m_i,m_\lambda\}} \frac{8}{2^{-m_*}} \\
<\ & \sum_{i\in A'} 8\cdot 2^{m_i+1} + \sum_{i\in A''\setminus A'} 8\cdot 2^{m_\lambda+1} \\
\leq\ & \sum_{i\in A'} 8\cdot \frac{8}{\Delta_i} + \sum_{i\in A''\setminus A'} 8\cdot\frac{8}{\lambda} = \sum_{i\in A'} \frac{64}{\Delta_i} + \sum_{i\in A''\setminus A'} \frac{64}{\lambda}.
\end{aligned}$$

Finally, summing up the individual contributions to the expected regret of the considered cases, and taking into account suboptimal arms not in A' gives the claimed bound. □

4. When the Horizon is Unknown

4.1. **Algorithm.** When T is unknown, the learner also has to guess T. Thus, we start the algorithm with $\tilde{T}_0 = 2$ and increase $\tilde{T}$ after reaching $\tilde{T}$ steps by setting $\tilde{T}_{\ell+1} := \tilde{T}_\ell^2$, so that $\tilde{T}_\ell = 2^{2^\ell}$.

4.2. **Analysis.** Fix some $\lambda \geq \sqrt{\frac{\mathrm{e}}{T}}$ and assume that $T > 2$. For arms i with $\Delta_i \leq \lambda$ we bound the regret by $T\Delta_i + \frac{64}{\lambda}$ as in Theorem 3.1. Thus let us consider the regret for an arbitrary arm i in $A' = \{i \in A \,|\, \Delta_i > \lambda\}$.

Let ℓ_i be the minimal ℓ with $\tilde{T}_\ell \Delta_i^2 \geq \mathrm{e}$, so that

$$2^{2^{\ell_i-1}} = \tilde{T}_{\ell_i-1} < \frac{\mathrm{e}}{\Delta_i^2} \leq \tilde{T}_{\ell_i} = 2^{2^{\ell_i}}. \tag{10}$$

Then the regret with respect to arm i before period ℓ_i is bounded according to Theorem 3.1, Remark 3.2, and (10) by

$$\begin{aligned}
&\sum_{\ell=0}^{\ell_i-1}\left(\max_{j\in A':\Delta_j\leq\sqrt{\mathrm{e}/\tilde{T}_\ell}} \Delta_j \tilde{T}_\ell + \frac{64}{\sqrt{\mathrm{e}}}\sqrt{\tilde{T}_\ell}\right) \\
&\leq \sum_{\ell=0}^{\ell_i-1}\left(\sqrt{\frac{\mathrm{e}}{\tilde{T}_\ell}}\cdot\tilde{T}_\ell + \frac{64}{\sqrt{\mathrm{e}}}\sqrt{\tilde{T}_\ell}\right) = \left(\sqrt{\mathrm{e}} + \tfrac{64}{\sqrt{\mathrm{e}}}\right)\sum_{\ell=0}^{\ell_i-1}\sqrt{\tilde{T}_\ell} \\
&= \left(\sqrt{\mathrm{e}} + \tfrac{64}{\sqrt{\mathrm{e}}}\right)\sum_{\ell=0}^{\ell_i-1} 2^{2^{\ell-1}} < 2\left(\sqrt{\mathrm{e}} + \tfrac{64}{\sqrt{\mathrm{e}}}\right)\cdot 2^{2^{\ell_i-2}} \\
&\leq 2\left(\sqrt{\mathrm{e}} + \tfrac{64}{\sqrt{\mathrm{e}}}\right)\sqrt{\tilde{T}_{\ell_i-1}} < \frac{2(\mathrm{e}+64)}{\Delta_i} < \frac{134}{\Delta_i}.
\end{aligned} \tag{11}$$

On the other hand, in periods $\ell \geq \ell_i$ the expected regret with respect to arm i is upper bounded according to Theorem 3.1 by

$$\left(\Delta_i + \frac{32\log(\tilde{T}_\ell\Delta_i^2)}{\Delta_i} + \frac{96}{\Delta_i}\right) \leq \frac{129\log(\tilde{T}_\ell\Delta_i^2)}{\Delta_i}.$$

Summing up over all these periods $\ell \geq \ell_i$ until the horizon T is reached in period $2 \leq L \leq \log_2 \log_2 T$ gives

$$\begin{aligned}
\sum_{\ell=\ell_i}^{L} \frac{129 \log(\tilde{T}_\ell \Delta_i^2)}{\Delta_i} &\leq \frac{129}{\Delta_i} \sum_{\ell=\ell_i}^{L} \log(2^{2^\ell} \Delta_i^2) \\
&\leq \frac{129}{\Delta_i} \left(L \log(\Delta_i^2) + (\log 2) \sum_{\ell=0}^{L} 2^\ell \right) \\
&< \frac{129}{\Delta_i} \big(L \log(\Delta_i^2) + 2^{L+1} \log 2 \big) \\
&\leq \frac{129}{\Delta_i} \big(L \log(\Delta_i^2) + 2 \log T \big) \\
&\leq \frac{258 \log(T \Delta_i^2)}{\Delta_i}.
\end{aligned}$$

Taking into account the periods before ℓ_i according to (11) and summing up over all arms in A' gives the following regret bound for the case when the horizon is unknown.

Theorem 4.1. *The total expected regret for the algorithm described in Subsection 4.1 is upper bounded by*

$$\sum_{i \in A : \Delta_i > \lambda} \left(\frac{258 \log(T \Delta_i^2)}{\Delta_i} + \frac{134}{\Delta_i} \right) + \sum_{i \in A : 0 < \Delta_i \leq \lambda} \frac{64}{\lambda} + \max_{i \in A : \Delta_i \leq \lambda} \Delta_i T.$$

for all $\lambda \geq \sqrt{\frac{\mathrm{e}}{T}}$.

5. Conclusion

We were able to improve on the regret bounds of the original UCB algorithm concerning the dependency on T for small Δ_i. Still, the dependency on the number of arms is not completely satisfactory and requires further investigation. Recently, another attempt to modify UCB in order to obtain improved bounds [2] gave logarithmic regret bounds of order $K \sum_{i : r_i < r^*} \frac{\log(T \Delta_i^2 / K)}{\Delta_i}$. The authors of [2] conjecture that the additional factor K in the bound for their algorithm MOSS (when compared to our bound) is an artefact of their proof that one should be able to remove by improved analysis. Of course, generally our algorithm as well as MOSS would benefit from taking into account also the empirical variance for each arm. For modifications of the original UCB algorithm this has been demonstrated in [3].

REFERENCES

[1] Rajeev Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. in Appl. Probab.*, 27(4):1054–1078, 1995.

[2] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *colt2009. Proceedings of the 22nd Annual Conference on Learning Theory*, pages 217–226, 2009.

[3] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, 2009.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256, 2002.

[5] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.

[6] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, 2006.

[7] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[8] Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704. MIT Press, 2005.

[9] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6:4–22, 1985.

[10] Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, 2004.

[11] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

E-mail address: auer@unileoben.ac.at, rortner@unileoben.ac.at

LEHRSTUHL FÜR INFORMATIONSTECHNOLOGIE, MONTANUNIVERSITÄT LEOBEN
FRANZ-JOSEF-STRASSE 18, A-8700 LEOBEN, AUSTRIA