Continuous Time Associative Bandit Problems[*]

András György[1] and **Levente Kocsis**[1] and **Ivett Szabó**[1,2] and **Csaba Szepesvári**[1,3]

[1]Computer and Automation Research Institute of the Hungarian Academy
of Sciences, Machine Learning Research Group
gya@szit.bme.hu, kocsis@sztaki.hu, ivett7@gmail.hu, szcsaba@sztaki.hu

[2] Budapest University of Technology and Economics, Department of Stochastics

[3]University of Alberta, Department of Computing Science

Abstract

In this paper we consider an extension of the multi-armed bandit problem. In this generalized setting, the decision maker receives some side information, performs an action chosen from a finite set and then receives a reward. Unlike in the standard bandit settings, performing an action takes a random period of time. The environment is assumed to be stationary, stochastic and memoryless. The goal is to maximize the average reward received in one unit time, that is, to maximize the average rate of return. We consider the on-line learning problem where the decision maker initially does not know anything about the environment but must learn about it by trial and error. We propose an "upper confidence bound"-style algorithm that exploits the structure of the problem. We show that the regret of this algorithm relative to the optimal algorithm that has perfect knowledge about the problem grows with the number of decisions n at a rate $\tilde{O}(\log n)$, that is, only slightly faster than the optimal logarithmic rate, and scales polynomially with the parameters of the problem.

1 Introduction

Multi-armed bandit problems find applications in various fields, such as statistics, control, learning theory or economics. They became popular with the seminal paper by Robbins [1952] and since then they enjoy perpetual popularity.

The version of the bandit problem we consider here is motivated by the following example: Imagine that a sequence of tasks arrive for processing in a computer center that has a single supercomputer. For each of the tasks a number of alternative algorithms can be applied to. Some information about the tasks is available that can be used to predict which of the algorithms to try. The processing time depends on the task at hand and also on the algorithm selected and may take

continuous values, hence the time instants when the decisions can take place take continuous values, too. The supercomputer has a fixed cost of running, whilst the centre's income is based on the quality of solutions delivered. At any given time only a single task can be executed on the supercomputer. Admittedly, this assumption looks absurd at the first sight in the context of our example, however, we think that our results can be extended to the more general case when the number of algorithms that can run simultaneously is bounded by a constant without much trouble. Hence we decided to stick to this simplifying assumption.

An allocation rule decides, based on the *side information* available about the task just received, which algorithm to use for processing it, the goal being to maximize the *return rate*. Note that this criterion is different from maximizing the total reward. In fact, since processing a task takes some time during which no other tasks can be processed, the rate maximization problem cannot be solved by selecting the algorithm with the highest expected payoff: Some tasks may look so difficult to solve that the best thing could be to drop them, which results in no payoff, but in exchange the learner does not suffer any loss due to not processing other, possibly more rewarding tasks. (Note that this would not be possible without the presence of side information; in the latter case the problem would simplify to the usual multi-armed bandit problem where one needs to find the best option with highest reward rate.) This example illustrates that a learner whose aim is to *quickly* learn a good allocation strategy for *rate maximization* must solve two problems simultaneously: Predicting the long-term values of the available algorithms given the information about the task to be processed and balancing exploration and exploitation so that the loss due to selecting inferior options (i.e., the regret) is kept at minimum. The problem we consider can be thought of as a minimalistic example where the learner faces these two problems simultaneously.

Bandit problems in continuous time have been studied earlier by a number of authors (see e.g. [Kaspi and Mandelbaum, 1998; Karoui and Karatzas, 1997] and the references therein). These earlier results concern the construction of optimal allocation policies (typically in the form of Gittins indexes) given some parametric form of the distributions of the random variables involved. In contrast, here we consider the agnostic case when no particular parametric form is assumed, but the environment is supposed to be stationary and stochastic. The agnostic (or non-parametric) case has been studied

---

extensively in the discrete time case. In fact, this problem was first considered by Robbins [1952], who introduced a certainty-equivalence rule with forcing. In the same article Robbins showed that this rule is asymptotically consistent in the sense that the frequency of the time instants when the best arm is selected converges to one almost surely. More recently, Agrawal [1995] suggested a number of simple sample-mean based policies and showed that the resulting policies' regret after n decisions is $O(\log n)$. Since it is known that no allocation rule can achieve regret lower than $C_p \log n$ for an appropriate (problem dependent) constant C_p [Lai and Robbins, 1985], Agrawals' policies are unimprovable apart from constant factors. Lately, Auer et al. [2002] strengthened these results by suggesting policies that achieve logarithmic regret *uniformly* over time, rather than just asymptotically. An added benefit of their policies is that they are simple to implement.

We base our algorithm on algorithm UCB1 from [Auer *et al.*, 2002] (see also [Agrawal, 1995]). We assume a stationary memoryless stochastic environment, where the side information is an i.i.d. process taking values in a finite set, the payoff-delay sequences are jointly distributed for any of the options and their distribution depends on the side information (the precise assumptions will be listed in the next section). Like UCB1, our algorithm decides which option to choose based on sample-means corrected by <u>u</u>pper <u>c</u>onfidence <u>b</u>ounds. In our case, however, separate statistics are kept for all option - side-information pairs. Our main result shows that the resulting policy achieves an almost logarithmic regret uniformly in time: the growth rate of the regret can be arbitrarily close to the theoretically unimprovable logarithmic rate, but it is not able to achieve it.

The paper is organized as follows: We define the problem and the proposed algorithm in Section 2. Our main result, a logarithmic regret bound on the algorithm's performance is presented in Section 4. Conclusions are drawn in Section 5.

2 The algorithm

The problem of the previous section is formalized as follows: Let K denote the number of options available, and let $\mathcal{X}$ denote the set of possible values of the side information, which is assumed to be finite. Let $x_1, x_2, \ldots, x_t$ be a random sequence of covariates representing the side information available at the time of the t-th decision, generated independently from a distribution p supported on $\mathcal{X}$. At each decision point the decision maker may select an option I_t from $\mathcal{A} = \{1, \ldots, K\}$ and receives reward $r_t = r_{I_t,t}(x_t)$, where $r_{it}(x_t)$ is the reward the decision maker would have received had it chosen option i. Unlike in classical bandit problems the collection of the reward takes some random time. When option i is selected and the side information equals x, this time is $\delta_{it}(x)$. We assume that for any fixed x and i, $(r_{it}(x), \delta_{it}(x))$ is an i.i.d. sequence, independent of $\{x_t\}$. We further assume that $r_{it}(x) \in [r_{\min}, r_{\max}]$, $\delta_{it}(x) \in [\delta_{\min}, \delta_{\max}]$ with $\delta_{\min} > 0$. (We expect that the boundedness assumptions can be relaxed to $\delta_{it}(x) \geq 0$ and appropriate moment conditions on $\delta_{it}(x)$ and $r_{it}(x)$.) Let

$$r_i(x) = \mathbb{E}\left[r_{i1}(x)\right]$$

and

$$\delta_i(x) = \mathbb{E}\left[\delta_{i1}(x)\right]$$

denote the expected reward and delay, respectively, when option i is chosen at the presence of the side-information x.

The exact protocol of decision making is as follows: Decision making happens in discrete trials. Let $\tau_0 = 0$ and let τ_t denote the time of the beginning of the t-th trial. At the beginning of the tth trial the decision maker receives the side information x_t. Based on the value of x_t and all information received by the decision maker at prior trials, the decision maker must select an option I_t. Upon executing I_t, the decision maker receives a reward $r_t = r_{I_t,t}(x_t)$ and suffers a delay $\delta_t = \delta_{I_t,t}(x_t)$. That is, the next time point available when the decision maker can select an option is $\tau_{t+1} = \tau_t + \delta_{I_t,t}(x_t)$.

The goal of the decision maker is to find a good allocation policy. Formally, an allocation policy maps possible histories to some index in the set $\mathcal{A}$. The gain (average reward rate) delivered by an allocation policy u is given by

$$\lambda^u = \limsup_{n \to \infty} \frac{\mathbb{E}\left[\sum_{t=1}^n r_t^u\right]}{\mathbb{E}\left[\sum_{t=1}^n \delta_t^u\right]},$$

where $\{r_t^u\}$ is the reward sequence and $\{\delta_t^u\}$ is the delay sequence experienced when policy u is used. An optimal allocation policy is one that maximizes this gain. Note that the problem as stated is a special case of semi-Markov decision problems [Puterman, 1994]. The theory of semi-Markov decision problems furnishes us with the necessary tools to characterize optimal allocation policies: Let us define the optimal gain by

$$\lambda^* = \sup_u \lambda^u.$$

A policy u is said to be optimal if it satisfies $\lambda^* = \lambda^u$. It follows from the generic theory that there exist deterministic stationary policies that are optimal. An optimal action for some $x \in \mathcal{X}$ can be determined by ordering the options by their *relative values*. The *relative value* of option i upon observing x is the expected reward that can be collected minus the expected reward that is *not gained* during the time it takes to collect the reward:

$$q_i^*(x) = r_i(x) - \delta_i(x)\lambda^*.$$

Intuitively it should be clear that a policy that always selects options with best relative values should optimize the overall gain. In fact, it follows from the theory of semi-Markov decision problems that this is indeed the case. A stationary deterministic policy $u : \mathcal{X} \to \mathcal{A}$ is optimal if and only if it obeys the constraints

$$r_{u(x)}(x) - \delta_{u(x)}(x)\lambda^* = \max_{i \in \mathcal{A}} \left[r_i(x) - \delta_i(x)\lambda^*\right] \quad (1)$$

simultaneously for all $x \in \mathcal{X}$.

The (total) regret of an allocation policy is defined as the loss suffered due to not selecting an optimal option in each time step. Since we are interested in the expected regret only, our regret definition uses the optimal gain λ^*:

$$R_n = \lambda^* \sum_{t=1}^n \delta_t - \sum_{t=1}^n r_t.$$

The value of the first term is the maximum reward that could be collected during the time of the first n decisions. The expected regret thus compares the expected value of the latter

with the expected value of the actual total payoffs received. It follows that an allocation policy that minimizes the regret will optimize the rate of return.

When $\delta_{it}(x) = 1$, and $\mathcal{X}$ has a single element, the problem reduces to the classical stochastic bandit problem. Since for the stochastic bandit problems the regret is lower bounded by $O(\log n)$, we are seeking policies whose regret grows at most at a logarithmic rate.

The idea underlying our algorithm is to develop upper estimates of the values $q_i^*(x)$ with appropriate confidence bounds. Just like in [Auer *et al.*, 2002], the upper confidence estimates are selected to ensure that for any given x with $p(x) > 0$ all options are ultimately selected infinitely often, but at the same time suboptimal options are selected increasingly rarely.

The algorithm is as follows: Let us consider the t-th decision. If we had a good estimate $\overline{\lambda}_t$ of λ^*, then for any given x we could base our decision on the estimates of the relative values $q_i^*(x)$ of the options given by $\overline{r}_{it}(x) - \overline{\delta}_{it}(x)\overline{\lambda}_t$. Here $\overline{r}_{it}(x)$ denotes the average of rewards during the first n decisions for those time points when the side information is x and option i was selected, and $\overline{\delta}_{it}(x)$ is defined analogously:

$$\overline{r}_{it}(x) = \frac{1}{T_i(x,t)} \sum_{s=1}^{t} \mathbb{I}(I_s = i, x_s = x) r_s,$$

$$\overline{\delta}_{it}(x) = \frac{1}{T_i(x,t)} \sum_{s=1}^{t} \mathbb{I}(I_s = i, x_s = x) \delta_s,$$

where $T_i(x,t)$ denotes the number of times option i was selected when side information x was present in trials $1, 2, \ldots, t$:

$$T_i(x,t) = \sum_{j=1}^{t} \mathbb{I}(I_t = i, x_t = x).$$

The plan is to combine appropriate upper bounds on $r_i(x)$ and lower bounds on $\delta_i(x)$ based on the respective sample averages $\overline{r}_{it}(x)$, $\overline{\delta}_{it}(x)$ and $T_i(x,t)$, to obtain an upper estimate of $q_i^*(x)$. However, in order to have a sample based estimate, we also need an appropriate lower estimate of λ^*. This estimate is defined as follows:

Let $p_t(x)$ be the maximum-likelihood estimate of the density over $\mathcal{X}$: $p_t(x) = 1/t \sum_{s=1}^{t} \mathbb{I}(x_s = x)$. Let $\mathcal{U}$ denote the set of stationary policies: $\mathcal{U} = \{u | u : \mathcal{X} \to \mathcal{A}\}$. Pick any $u \in \mathcal{U}$. Clearly, the gain of u satisfies

$$\lambda^u = r^u / \delta^u,$$

where

$$r^u = \sum_{x \in \mathcal{X}} p(x) r_{u(x)}(x),$$

$$\delta^u = \sum_{x \in \mathcal{X}} p(x) \delta_{u(x)}(x).$$

We compute the the empirical estimate of the gain of policy u from the empirical estimates of $p(x)$, $r_i(x)$ and $\delta_i(x)$:

$$\overline{r}_t^u = \sum_{x \in \mathcal{X}} p_t(x) \overline{r}_{u(x),t}(x), \quad \overline{\delta}_t^u = \sum_{x \in \mathcal{X}} p_t(x) \overline{\delta}_{u(x),t}(x),$$

$$\text{and} \qquad \overline{\lambda}_t^u = \frac{\overline{r}_t^u}{\overline{\delta}_t^u}.$$

Then u_t^{UCB}, the estimate of u^* is defined by

$$u_t^{\text{UCB}} = \operatorname*{argmax}_{u \in \mathcal{U}} (\overline{\lambda}_t^u + c_t(u)).$$

Here $c_t(u)$ is an appropriate confidence bound that is selected such that simultaneously for all policies $u \in \mathcal{U}$, $\overline{\lambda}_t^u$ is in the $c_t(u)$-vicinity of λ^u with high probability. This sequence will be explicitly constructed during the proof where we will also make sure that it depends on known quantities only. In words, $\overline{\lambda}_t^{u_t^{\text{UCB}}}$ is the optimal gain that the decision maker can guarantee itself with high probability given the data seen so far.

In our proposed allocation policy, we ensure that each action is sampled sufficiently often for each value of side information: If $T_i(x_t, t-1) < A_t$ for some action $i \in \mathcal{A}$ and a deterministic function A_t to be defined later, then I_t is chosen to be the most undersampled action, otherwise I_t is chosen according to $u_{t-1}^{\text{UCB}}(x_t)$. That is,

$$
I_t = \begin{cases}
\operatorname{argmin}_{i \in \mathcal{A}} T_i(x_t, t-1) & \text{if } T_j(x_t, t-1) < A_t \\
& \text{for some } j \in \mathcal{A} \text{ (ties are} \\
& \text{broken arbitrarily);} \\
u_{t-1}^{\text{UCB}}(x_t), & \text{if } T_i(x_t, t-1) \geq A_t \\
& \text{for all } i \in \mathcal{A}.
\end{cases}
\tag{2}
$$

3 Some useful lemmata

Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$ be an increasing sequence of σ-algebras.

Lemma 1 (Corollary 16 in [Cesa-Bianchi *et al.*, 2004])
Assume that $\{X_s\}$ is a bounded martingale difference series with filtration $\{\mathcal{F}_s\}$, $-K \leq X_s \leq K$ for all s. Let $M_t = \sum_{s=1}^{t} X_s$ be the associated martingale and $V_t = \sum_{s=1}^{t} \mathbb{E}\left[X_s^2 \mid \mathcal{F}_{s-1}\right]$ be the predictable variance of the process. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(M_t \geq \sqrt{2(V_t + K^2)\log(t/\epsilon)} + \frac{\sqrt{2}}{3}K\log(t/\epsilon) \right) \leq \epsilon.$$

Corollary 2 *Let $I_s \in \{0,1\}$, $|Z_s| \leq K$ be $\mathcal{F}_s$ measurble rndom variables, and assume that $\mathbb{E}\left[I_s Z_s | \mathcal{F}_{s-1}\right] = 0$. Let $M_t = \sum_{s=1}^{t} I_s Z_s$, $T_t = \sum_{s=1}^{t} I_s$. Then for all $\epsilon > 0$, $t > 0$,*

$$\mathbb{P}\left(\frac{M_t}{T_t} > K\phi(t, T_t, \epsilon) \right) \leq \epsilon, \tag{3}$$

where

$$\phi(t, T, \epsilon) = \sqrt{\frac{2(1 + \frac{1}{T})\log(t/\epsilon)}{T}} + \frac{\sqrt{2}}{3}\frac{\log(t/\epsilon)}{T}. \tag{4}$$

Proof. Let $X_s = I_s Z_s$. Then X_s is a bounded martingale difference series. Observe that $\mathbb{E}\left[X_s^2 \mid \mathcal{F}_{s-1}\right] \leq K^2 I_s$ and hence $V_t \leq K^2 T_t$. Hence,

$$\mathbb{P}\left(M_t \geq \sqrt{2(V_t + K^2)\log(t/\epsilon)} + \frac{\sqrt{2}}{3}K\log(t/\epsilon) \right) \geq$$

$$\mathbb{P}\left(M_t \geq K\sqrt{2(T_t + 1)\log(t/\epsilon)} + \frac{\sqrt{2}}{3}K\log(t/\epsilon) \right)$$

and the result follows by Lemma 1. $\qquad \square$

Corollary 3 *For all $\epsilon > 0$, $t > 0$ the following inequalities hold:*

$$\mathbb{P}\left(\overline{r}_{it}(x) - r_i(x) > r_{\max}\phi(t, T_i(x,t),\epsilon)\right) \leq \epsilon,$$
$$\mathbb{P}\left(r_i(x) - \overline{r}_{it}(x) < r_{\max}\phi(t, T_i(x,t),\epsilon)\right) \leq \epsilon,$$
$$\mathbb{P}\left(\overline{\delta}_{it}(x) - \delta_i(x) > \delta_{\max}\phi(t, T_i(x,t),\epsilon)\right) \leq \epsilon,$$
$$\mathbb{P}\left(\delta_i(x) - \overline{\delta}_{it}(x) < \delta_{\max}\phi(t, T_i(x,t),\epsilon)\right) \leq \epsilon,$$

where ϕ is defined by (4).

Proof. Use Corollary 2 with $I_s = \mathbb{I}(x_s = x, I_s = i)$, $Z_s = r_s - r_i(x)$, $\mathcal{F}_{s-1} = \sigma(x_s, r_{s-1}, x_{s-1}, \ldots)$. Since $X_s = (r_s - r_i(x))\mathbb{I}(x_s = x, I_s = i) = (r_{I_s,s}(x_s) - r_i(x))\mathbb{I}(x_s = x, I_s = i) = (r_{i,s}(x) - r_i(x))\mathbb{I}(x_s = x, I_s = i)$, we have $\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = \mathbb{I}(x_s = x, I_s = i)\mathbb{E}[r_{i,s}(x) - r_i(x) \mid \mathcal{F}_{s-1}]$. By the assumed properties of the random sequence $\{r_{i,s}(x)\}_{s \geq 1}$ this expectation is indeed zero. Hence the conditions of the Corollary are satisfied. Hence (3) yields the first part of the result. The second part follows by symmetry. The analogous inequalities for $\overline{\delta}_{it}$ can be shown similarly. $\square$

Finally, we shall use the following lemma due to [Weissman *et al.*, 2003] which avoids a polynomial factor as a function of the number of samples:

Lemma 4 ([Weissman *et al.*, 2003], Theorem 2.1) *Let $x_1, \ldots, x_t \sim p$ be an i.i.d. sequence taking values on a finite set $\mathcal{X}$. Let $p_t(x) = \frac{1}{t}\sum_{s=1}^{t}\mathbb{I}(x_s = x)$ be the empirical distribution underlying $x_1, \ldots, x_t$. Then*

$$\mathbb{P}\left(\|p - p_t\|_1 \geq \epsilon\right) \leq (2^{|\mathcal{X}|} - 2)\exp(-t\epsilon^2/2).$$

Note that here we simplified the result of [Weissman *et al.*, 2003] to fit our purposes better (the exponential rate in [Weissman *et al.*, 2003] depends on the distribution, p).

Now, we are ready to prove the following:

Lemma 5 *Pick any policy u, and for any $\epsilon > 0$ define*

$$c_t(u) = \frac{r_{\max}}{\delta_{\min}}\left(1 + \frac{\delta_{\max}}{\delta_{\min}}\right)\phi_0(t, u, \epsilon), \qquad (5)$$

where

$$\phi_0(t, u, \epsilon) = \phi_1(t, u, \epsilon) + \phi_2(t, \epsilon),$$
$$\phi_1(t, u, \epsilon) = \sum_{x \in \mathcal{X}} p_t(x)\phi(t, T_{u(x)}(x,t), \epsilon_1),$$
$$\phi_2(t, \epsilon) = \sqrt{\frac{2\left(|\mathcal{X}|\log(2) + \log(1/\epsilon_1)\right)}{t}},$$
$$\epsilon_1 = \epsilon/(2|\mathcal{X}| + 1).$$

Then with probability at least $1 - \epsilon$, the following inequalities hold, individually:

$$\frac{\overline{r}_t^u}{\overline{\delta}_t^u} - \lambda^u \leq c_t(u),$$
$$\lambda^u - \frac{\overline{r}_t^u}{\overline{\delta}_t^u} \leq c_t(u)$$

Proof. By straightforward algebra, we get

$$\frac{\overline{r}_t^u}{\overline{\delta}_t^u} - \lambda^u = \frac{\overline{r}_t^u \delta^u - r^u \delta^u + r^u \delta^u - r^u \overline{\delta}_t^u}{\delta^u \overline{\delta}_t^u}$$
$$\leq \frac{1}{\delta_{\min}}(\overline{r}_t^u - r^u)_+ + \frac{r_{\max}}{\delta_{\min}^2}(\delta^u - \overline{\delta}_t^u)_+.$$

Let us bound $\overline{r}_t^u - r^u$:

$$\overline{r}_t^u - r^u = \sum_x p_t(x)\overline{r}_{u(x),t}(x) - \sum_x p(x)r_{u(x)}(x)$$
$$= \sum_x p_t(x)(\overline{r}_{u(x),t}(x) - r_{u(x)}(x))$$
$$+ \sum_x (p(x) - p_t(x))r_{u(x)}(x)$$
$$\leq \sum_x p_t(x)(\overline{r}_{u(x),t}(x) - r_{u(x)}(x))$$
$$+ r_{\max}\sum_x |p(x) - p_t(x)|.$$

Now Lemma 4 can be used to construct a bound the second term, whilst Corollary 3 can be used to construct a bound for the first term. The term $\delta^u - \overline{\delta}_t^u$ can be bound in an analogous way. Altogether, we need $(2|\mathcal{X}| + 1)$ bounds to hold simultaneously, hence we use $\epsilon_1 = \epsilon/(2|\mathcal{X}|+1)$ in these bounds. $\square$

4 Main result

Our main result is the following bound on the expected regret:

Theorem 6 *Let $A_t = \log t f(t)$ such that $\lim_{t \to \infty} f(t) = \infty$ and $\lim_{t \to \infty} A_t/t = 0$, and let $c_t(u)$ be defined by (5) with $\epsilon = \epsilon_t = 1/t^2$. Then the regret of algorithm (2) is bounded as*

$$E[R_n] \leq O(A_n).$$

Remark 7 *Note that the above bound shows that the algorithm can achieve a regret fo order $\tilde{O}(\log n)$, that is, it can get arbitrarily close to the optimal $O(\log n)$ rate, but it is not able to achieve it.*

Proof. The proof follows similar lines to that of Theorem 1 of [Auer *et al.*, 2002]. Let

$$\mathcal{U}^*(x) = \{i \in \mathcal{A} : q_i^*(x) = \max_{j \in \mathcal{A}} q_j^*(x)\},$$

and for each i and x, let

$$u(x, i, t) = \operatorname*{argmax}_{u \in \mathcal{U}: u(x)=i} \overline{\lambda}_t^u + c_t(u)$$

Then

$$\mathbb{E}[R_n]$$
$$= \sum_{t=1}^{n}(\delta_t \lambda^* - r_t)$$
$$\leq \sum_{t=1}^{n}\left(\delta_{\max}\frac{r_{\max}}{\delta_{\min}} - r_{\min}\right)\mathbb{P}(I_t \notin \mathcal{U}^*(x_t))$$
$$= \sum_{t=1}^{n}\left(\delta_{\max}\frac{r_{\max}}{\delta_{\min}} - r_{\min}\right)\sum_{x \in \mathcal{X}}p(x)\mathbb{P}(I_t \notin \mathcal{U}^*(x))$$
$$\leq \sum_{t=A_n}^{n-1}\left(\delta_{\max}\frac{r_{\max}}{\delta_{\min}} - r_{\min}\right)\sum_{x \in \mathcal{X}}\Big\{p(x) \cdot$$
$$\sum_{i \notin \mathcal{U}^*(x)}\mathbb{P}\left(\overline{\lambda}_t^{u(x,i,t)} + c_t(u(x,i,t)) \geq \overline{\lambda}_t^* + c_t(u^*)\right)\Big\}.$$

Now the probability term can be bounded by the union bound, for any policy u, as

$$\mathbb{P}\left(\overline{\lambda}_t^u + c_t(u) \geq \overline{\lambda}_t^* + c_t(u^*)\right)$$

$$\leq \quad \mathbb{P}\left(\overline{\lambda}_t^u - \lambda^u > c_t(u)\right) + \mathbb{P}\left(\overline{\lambda}_t^* - \lambda^* > c_t(u^*)\right)$$

$$+\mathbb{P}\left(\lambda^* - \overline{\lambda}_t^u < 2c_t(u)\right)$$

$$\leq \quad 2\epsilon + \mathbb{P}\left(\lambda^* - \lambda^u < 2c_t(u)\right)$$

where in the last step we used Lemma 5. Assume that the side information x is observed $T(x,t)$ times by trial t. If $T(x,t) > |\mathcal{A}|A_t$, then then $T_i(x,T) \geq A_t$ for all $i \in \mathcal{A}$, as the algorithm enforces a minimal numer of trials for each option. Then

$$c_t(u) \leq \frac{r_{\max}}{\delta_{\min}}\left(1 + \frac{\delta_{\max}}{\delta_{\min}}\right)(\phi(t, A_t, \epsilon/(2|\mathcal{X}| + 1)) + \phi_2(t)) \tag{6}$$

thus $\lambda^* - \lambda^u \geq \min_{u \in \mathcal{U}:\lambda^* > \lambda^u} \lambda^* - \lambda^u > 2c_t(u)$ if $A_t \geq C \log t$ where C is a suitable positive constant. This inequality is satisfied for sufficiently large t ($t = O(A_t)$) as $A_t = \log t f(t)$ where $f(t) \to \infty$ as $t \to \infty$. On the other hand, (6) holds simultaniously for all $u \in \mathcal{U}$ if $T(x,t) \geq |\mathcal{A}|A_t$ for all x. For sufficiently large t, this holds with probability at most $1 - \epsilon$, as Hoeffding's inequality implies

$$\mathbb{P}\left(T(x,t) < A_t\right)$$

$$< \quad \mathbb{P}\left(T(x,t) < p(x)t - \sqrt{t\log(|\mathcal{X}|/\epsilon)/2}\right)$$

$$< \quad \epsilon/|\mathcal{X}|.$$

Thus, for $t < O(A_t)$, the bound becomes $2\epsilon + 1$, for $t > O(A_t)$, it is 3ϵ. Setting $\epsilon_t = 1/t^2$, we obtain that

$$E[R_n] \leq O(A_t)$$

as desired. □

5   Conclusions and further work

We considered a generalization of the multi-armed bandit problem, where performing an action (or collecting the reward) takes a random amount of time. The goal of the decision maker is to maximize the reward per unit time where in each time step some side information is received before the decision is made. In this setting one needs to consider seriously the time needed to perform an action, since spending long times with less rewarding actions seriously limits the performance of any algorithm in a given time period. Therefore, efficient methods must predict simultaneously the expected rewards and durations of all actions, as well as to estimate the long term optimal performance. The latter is essential as each action has a hidden cost associated with it: since actions take time, for their correct evaluation their immediate payoffs must be decremented by the optimal reward lost during the time it takes to execute the action.

In this paper we proposed an algorithm to solve this problem, whose cumulative reward after performing n actions is only $\tilde{O}(\log n)$ less than that of the best policy in hindsight. The algorithm is based on the upper confidence bound idea of Auer et al. [2002]. However, in our case the natural performance measure for a policy is the average return per time

unit, and making an efficiet estimate for this quantity is significantly harder than in the simple bandit case. The reason for this is that all the actions are connected via the averaging, hence even one undersampled action can bias the whole estimate. Another problem with the algorithm as presented is that it needs to enumerate all the policies in order to compute the estimate of the optimal gain.

The extension of the algorithm presented seems possible to certain semi-Markov models when there is a state that is recurrent under all stationary policies. Another interesting avenue for further research is to consider continuous time bandit problems in non-stochastic environments.

References

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

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

[Cesa-Bianchi *et al.*, 2004] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2004.

[Karoui and Karatzas, 1997] N. El Karoui and I. Karatzas. Synchronization and optimality for multi-armed bandit problems in continuous time. *Computational and Applied Mathematics*, 16:117–152, 1997.

[Kaspi and Mandelbaum, 1998] H. Kaspi and A. Mandelbaum. Multi armed bandits in discrete and continuous time. *Annals of Applied Probabability*, 8:1270–1290, 1998.

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

[Puterman, 1994] M.L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

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

[Szepesvári and Littman, 1999] Cs. Szepesvári and M.L. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computation*, 11:2017–2059, 1999.

[Weissman *et al.*, 2003] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M.J. Weinberger. Inequalities for the L^1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Labs., 2003.