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 at the optimal logarithmic rate in the number of decisions 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,

*This research was supported in part by the Ministry of Economy and Transport, Hungary (Grant No. GVOP-3.1.1.-2004-05-0145/3.0), the Hungarian National Science Foundation (Grant No. T047193), and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

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 upper confidence bounds. In our case, however, separate statistics are kept for all option - side-information pairs. Our main result shows that the resulting policy achieves logarithmic regret uniformly in time and hence it is also unimprovable, apart from constant factors.

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 3. Conclusions are drawn in Section 4.

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:

$$\begin{aligned}\overline{r}_{it}(x) &= \frac{1}{T_i(x,t)} \sum_{s=1}^{t} \mathbb{I}\left(I_s = i, x_s = x\right) r_s, \\ \overline{\delta}_{it}(x) &= \frac{1}{T_i(x,t)} \sum_{s=1}^{t} \mathbb{I}\left(I_s = i, x_s = x\right) \delta_s,\end{aligned}$$

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

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

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 $\mathcal{U}$ denote the set of stationary policies: $\mathcal{U} = \{u | u : \mathcal{X} \to \mathcal{A}\}$. Pick any $u \in \mathcal{U}$. Let $\overline{\lambda}_t^u$ denote the empirical estimate of the gain of policy u:

$$\overline{\lambda}_t^u = \frac{\sum_{s=1}^{t} \mathbb{I}\left(I_s = u(x_s)\right) r_s}{\sum_{s=1}^{t} \mathbb{I}\left(I_s = u(x_s)\right) \delta_s}$$

and let $T_u(t)$ denote the number of times when an option 'compatible' with policy u was selected:

$$T_u(t) = \sum_{s=1}^{t} \mathbb{I}\left(I_s = u(x_s)\right).$$

Then $\overline{\lambda}_t$, the estimate of λ^* is defined by

$$\overline{\lambda}_t = \max_{u \in \mathcal{U}} (\overline{\lambda}_t^u - c_{t,T_u(t)}).$$

Here $c_{t,s}$ is an appropriate deterministic sequence that is selected such that simultaneously for all policies $u \in \mathcal{U}$, $\overline{\lambda}_t^u$ is in the $c_{t,T_u(t)}$-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, $\overline{\lambda}_n$ is the optimal gain that the decision maker can guarantee itself with high probability given the data seen so far.

Our proposed allocation policy, $\{u_t^{\mathrm{UCB}}\}$, selects the options $I_t = u_t^{\mathrm{UCB}}(x_t)$ by the rule

$$u_t^{\mathrm{UCB}}(x) = \operatorname*{argmax}_{i \in \mathcal{A}} \left\{ \overline{r}_{it}(x) - \overline{\delta}_{it}(x)\overline{\lambda}_t + \hat{c}_{t,T_i(x,t)} \right\},$$

where, similarly to $c_{t,s}$, $\hat{c}_{t,s}$ is an appropriate deterministic sequence that will be chosen later.

3 Main result

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

Theorem 1 *Let the assumptions of the previous section hold on r_{it}, δ_{it}, x_t. Let R_n be the n-step regret of policy u^{UCB}. Then, for all $n \geq 1$,*

$$\begin{aligned}\mathbb{E}\left[R_n\right] \;\leq\; L^* \Bigg(\left(2 + \frac{2\pi^2}{3(|\mathcal{U}|+1)^2}\right) K|\mathcal{X}| + 2K|\mathcal{X}| \log(n) \\ + \sum_{i:\Delta_i>0} \sum_{x \in \mathcal{X}} \frac{a \log(n(\sqrt{|\mathcal{U}|+1}))}{\Delta_i(x)^2} \Bigg),\end{aligned}$$

where $L^* = \delta_{\max}\lambda^* - r_{\min}$,

$$\Delta_i(x) = \max_{j \in \mathcal{A}} q_j^*(x) - q_i^*(x) \geq 0, \quad i = 1, \ldots, K,$$

and the positive constant a is given by (7) in the proof of the theorem.

The proof follows similar lines to that of Theorem 1 of [Auer *et al.*, 2002], with the main difference being that now we have to handle the estimation error of λ^*. We prove the theorem using a series of propositions.

The first proposition bounds the expected regret in terms of the number of times when some suboptimal option is chosen:

Proposition 2 *The following bound holds for the expected regret of an arbitrary policy, $u = (u_1, u_2, \ldots)$:*

$$\mathbb{E}\left[R_n\right] \leq \sum_{x \in \mathcal{X}} p(x) L^*(x) \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left(u_t(x) \notin \mathcal{U}^*(x)\right)\right], \quad (2)$$

where

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

denotes the set of optimal options at x, and

$$L^*(x) = \max_j (\delta_j(x)\lambda^* - r_j(x))$$

is the loss for the worst choice at x. Further, by $L^*(x) \leq L^*$,

$$\begin{aligned}\mathbb{E}\left[R_n\right] &\leq L^* \sum_{x \in \mathcal{X}} p(x) \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left(u_t(x) \notin \mathcal{U}^*(x)\right)\right] \\ &= L^* \sum_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left(u_t(x) \notin \mathcal{U}^*(x), x_t = x\right)\right].\end{aligned}$$

Proof. Let us consider the t-th term, $\mathbb{E}[\delta_t\lambda^* - r_t]$, of the expected regret. We have $\mathbb{E}[\delta_t\lambda^* - r_t] = \sum_{i\in\mathcal{A}} \mathbb{E}[(\delta_t\lambda^* - r_t)\mathbb{I}(I_t = i)]$. Using $I_t = u_t(x_t)$ and that u_t depends only on the past, i.e., if $\mathcal{F}_t$ is the sigma algebra of $x_1, r_1, \delta_1, \ldots, x_t, r_t, \delta_t$ then $I_t = i$ is $\mathcal{F}_{t-1}$ measurable, we get that

$$\begin{aligned}
&\mathbb{E}[(\delta_t\lambda^* - r_t)\mathbb{I}(I_t = i)] \\
&\quad = \mathbb{E}[(\delta_{i,t}(x_t)\lambda^* - r_{i,t}(x_t))\mathbb{I}(I_t = i)] \\
&\quad = \mathbb{E}[\mathbb{E}[(\delta_{i,t}(x_t)\lambda^* - r_{i,t}(x_t))\mathbb{I}(I_t = i)\,|\mathcal{F}_{t-1}, x_t]] \\
&\quad = \mathbb{E}[\mathbb{I}(I_t = i)\,\mathbb{E}[(\delta_{i,t}(x_t)\lambda^* - r_{i,t}(x_t))|\mathcal{F}_{t-1}, x_t]] \\
&\quad = \mathbb{E}[\mathbb{I}(I_t = i)\,\mathbb{E}[(\delta_i(x_t)\lambda^* - r_i(x_t))|\mathcal{F}_{t-1}, x_t]] \\
&\quad = \mathbb{E}[\mathbb{I}(I_t = i)\,(\delta_i(x_t)\lambda^* - r_i(x_t))]\,.
\end{aligned}$$

Now, using again that u_t does not depend on x_t, we get

$$\begin{aligned}
&\mathbb{E}[\delta_t\lambda^* - r_t] \\
&\quad = \sum_{i\in\mathcal{A}} \mathbb{E}[\mathbb{I}(u_t(x_t) = i)\,(\delta_i(x_t)\lambda^* - r_i(x_t))] \\
&\quad = -\sum_{i\in\mathcal{A}}\sum_{x\in\mathcal{X}} p(x)q_i^*(x)\mathbb{E}[\mathbb{I}(u_t(x) = i)\,|x_t = x] \\
&\quad = -\sum_{i\in\mathcal{A}}\sum_{x\in\mathcal{X}} p(x)q_i^*(x)\mathbb{E}[\mathbb{I}(u_t(x) = i)]\,.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[\delta_t\lambda^* - r_t] \;=\; & -\sum_{x\in\mathcal{X}} p(x) \sum_{i\in\mathcal{U}^*(x)} q_i^*(x)\mathbb{E}[\mathbb{I}(u_t(x) = i)] \\
& -\sum_{x\in\mathcal{X}} p(x) \sum_{i\notin\mathcal{U}^*(x)} q_i^*(x)\mathbb{E}[\mathbb{I}(u_t(x) = i)]\,.
\end{aligned}$$

Let $w_t(i|x) = \mathbb{E}[\mathbb{I}(u_t(x) = i)]$ if $i \in \mathcal{U}^*(x)$ and $w_t(i|x) = 0$ otherwise, and let $\mu_t(i|x) = w_t(i|x)/\sum_{j\in\mathcal{A}} w_t(j|x)$. Then $\mu_t(i|x) \geq w_t(i|x)$ (since $\sum_{j\in\mathcal{A}} w_t(j|x) \leq 1$), the first term of the last expression can be upper bounded by

$$v_t = -\sum_{x\in\mathcal{X}} p(x) \sum_i q_i^*(x)\mu_t(i|x).$$

Since $\mu_t(i|x) = 0$ if i is not optimal, μ_t defines an optimal (stochastic) policy and hence, Bellman's equation gives $v_t = 0$. Therefore,

$$\begin{aligned}
\mathbb{E}[\delta_t\lambda^* - r_t] \;&\leq\; -\sum_{x\in\mathcal{X}} p(x) \sum_{i\notin\mathcal{U}^*(x)} q_i^*(x)\mathbb{E}[\mathbb{I}(u_t(x) = i)] \\
&\leq\; \sum_{x\in\mathcal{X}} p(x)L^*(x) \sum_{i\notin\mathcal{U}^*(x)} \mathbb{E}[\mathbb{I}(u_t(x) = i)] \\
&=\; \sum_{x\in\mathcal{X}} p(x)L^*(x)\mathbb{E}[\mathbb{I}(u_t(x) \notin \mathcal{U}^*(x))]\,.
\end{aligned}$$

Summing up this last expression over t gives the advertised bound. □

The next statements are used to prove that with high probability $\overline{\lambda}_t$ is a good estimate of λ^*. Here and in what follows u^* denotes an arbitrary (fixed) optimal policy and $\overline{\lambda}_t^* = \overline{\lambda}_t^{u^*}$.

Proposition 3 *Assume that the following conditions are satisfied:*

$$\lambda^u \;\geq\; \overline{\lambda}_t^u - c_{t,T_u(t)}, \tag{3}$$

$$\lambda^* \;\leq\; \overline{\lambda}_t^* + c_{t,T_{u^*}(t)}. \tag{4}$$

where the first condition is meant to hold for all stationary policies $u \in \mathcal{U}$. Then

$$\lambda^* \geq \overline{\lambda}_t \geq \lambda^* - 2c_{t,T_{u^*}(t)}. \tag{5}$$

Proof. Let u' be the policy that maximizes $\overline{\lambda}_t^u - c_{t,T_u(t)}$. Since (3) holds for u', we get that $\overline{\lambda}_t = \overline{\lambda}_t^u - c_{t,T_{u'(t)}} \leq \lambda^{u'} \leq \lambda^*$, proving the upper bound for $\overline{\lambda}_t$. On the other hand, because of the choice of u', $\overline{\lambda}_t \geq \overline{\lambda}_t^* - c_{t,T_{u^*(t)}}$ which can be further lower bounded by $\lambda^* - 2c_{t,T_{u^*(t)}}$ using (4), proving the lower bound for $\overline{\lambda}_t$. □

The following proposition shows that $\overline{\lambda}_t$ is indeed a lower bound for λ^* with high probability.

Proposition 4 *Let*

$$c_{t,s} = \sqrt{\frac{2c_1 \log(t\sqrt{|\mathcal{U}|+1})}{s}}$$

where

$$c_1 = 2\max\left\{\frac{(r_{\max} - r_{\min})^2}{\delta_{\min}^2}, \frac{r_{\max}^2(\delta_{\max} - \delta_{\min})^2}{\delta_{\min}^4}\right\}.$$

Then

$$\mathbb{P}\left(\overline{\lambda}_t < \lambda^* - 2c_{t,T_{u^*}(t)}\right) + \mathbb{P}\left(\lambda^* < \overline{\lambda}_t\right) \leq \frac{2}{t}.$$

Proof. According to Proposition 3, if (3) holds for all stationary policies u and if (4) holds then $\lambda^* \geq \overline{\lambda}_t \geq \lambda^* - 2c_{t,T_{u^*}(t)}$. Hence, in order $\overline{\lambda}_t < \lambda^* - 2c_{t,T_{u^*}(t)}$ or $\overline{\lambda}_t > \lambda^*$ to hold, we must have that one of the conditions in Proposition 3 is violated. Using a union bound we get

$$\begin{aligned}
&\mathbb{P}\left(\overline{\lambda}_t < \lambda^* - 2c_{t,T_{u^*}(t)}\right) + \mathbb{P}\left(\lambda^* < \overline{\lambda}_t\right) \\
&\quad \leq \sum_u \mathbb{P}\left(\lambda^u < \overline{\lambda}_t^u - c_{t,T_u(t)}\right) + \mathbb{P}\left(\lambda^* < \overline{\lambda}_t^* + c_{t,T_u(t)}\right).
\end{aligned}$$

Fix u. By the law of total probability,

$$\mathbb{P}\left(\lambda^u < \overline{\lambda}_t^u - c_{t,T_u(t)}\right) = \sum_{s=1}^t \mathbb{P}\left(\lambda^u < \overline{\lambda}_t^u - c_{ts}, T_u(t) = s\right).$$

Define

$$\hat{r}_t^u \;=\; \sum_{s=1}^t \mathbb{I}(I_s = u(x_s))\,r_s, \quad r^u = \sum_{x\in\mathcal{X}} p(x)r_{u(x)}(x)$$

$$\hat{\delta}_t^u \;=\; \sum_{s=1}^t \mathbb{I}(I_s = u(x_s))\,\delta_s, \quad \delta^u = \sum_{x\in\mathcal{X}} p(x)\delta_{u(x)}(x).$$

Using elementary algebra, we get that

$$\begin{aligned}
&\mathbb{P}\left(\lambda^u < \overline{\lambda}_t^u - c_{ts}, T_u(t) = s\right) \\
&\quad \leq\; \mathbb{P}\left(c_{ts}\delta_{\min}/2 \leq \hat{r}_t^u/s - r^u, T_u(t) = s\right) \\
&\qquad + \mathbb{P}\left(c_{ts}\delta_{\min}^2/r_{\max} \leq \delta^u - \hat{\delta}_t^u/s, T_u(t) = s\right).
\end{aligned}$$

Exploiting that $\hat{r}_t^u$ and $\hat{\delta}_t^u$ are martingale sequences and resorting to a slight variant of the Hoeffding-Azuma bound (see, e.g. [Devroye *et al.*, 1996]), we get the bound $2/(|\mathcal{U}|+1)t^{-2}$. Summing over s and u and by an analogous argument for $\mathbb{P}\left(\lambda^* < \overline{\lambda}_t^* + c_{t,T_u(t)}\right)$, we get the desired bound. □

Now we are ready to prove the main theorem. In the proof we put a superscript '$*$' to any quantity that refers to the optimal policy u^*. For example, $\overline{r}_t^*(x) = \overline{r}_{u^*(x),t}(x)$, $\delta_t^*(x) = \overline{\delta}_{u^*(x),t}(x)$, $T^*(x,t) = T_{u^*(x)}(x,t)$, etc.

Proof of Theorem 1. Proposition 2 applied to u^{UCB} shows that it suffices if for any fixed $x \in \mathcal{X}$ and suboptimal choice $i \notin \mathcal{U}^*(x)$ we derive an $O(\log n)$ upper bound on the expected number of times choice i would be selected by u^{UCB} when the side information is x. That is, we need to show

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i, x_t = x\right)\right] \le O(\log n). \qquad (6)$$

Let $q_{it}(x) = \overline{r}_{it}(x) - \overline{\delta}_{it}(x)\overline{\lambda}_t$. Using the definition of u_t^{UCB}, if $u_t^{\mathrm{UCB}}(x) = i$ holds then $q_{it}(x) + \hat{c}_{t,T_i(x,t)} > q_t^*(x) + \hat{c}_{t,T^*(x,t)}$. Hence, for any integer $A(n,x)$,

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i\right) \le A(n,x) \\
&\quad + \sum_{t=1}^{n} \mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i, T_i(x,t-1) \ge A(n,x), x_t = x\right) \\
&\le A(n,x) + \sum_{t=1}^{n} \mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i, T_i(x,t-1) \ge A(n,x)\right).
\end{aligned}$$

We write the t-th term in the last sum as follows:

$$\begin{aligned}
&\mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i, T_i(x,t-1) \ge A(n)\right) \\
&= \mathbb{I}\Big(q_{i,t-1}(x) + \hat{c}_{t,T_i(x,t-1)} > q_{t-1}^*(x) + \hat{c}_{t-1,T^*(x,t-1)}, \\
&\qquad\qquad T_i(x,t-1) \ge A(n)\Big) \\
&= \sum_{(s,s')\in H(t)} \mathbb{I}\left(q_{i,t-1}(x) + \hat{c}_{t-1,s'} > q_{t-1}^*(x) + \hat{c}_{t-1,s}\right) Z_t(s,s'),
\end{aligned}$$

where

$$\begin{aligned}
H(t) &= \{(s,s') | 1 \le s \le t-1, A(t) \le s' \le t-1\}, \\
Z_t(s,s') &= \mathbb{I}\left(T_i(x,t-1) = s', T^*(x,t-1) = s\right).
\end{aligned}$$

Fix any $s, s' \in H(t)$. Using the definition of $q_{it}(x)$,

$$\begin{aligned}
&\mathbb{I}\left(q_{i,t-1}(x) + \hat{c}_{t-1,s'} > q_{t-1}^*(x) + \hat{c}_{t-1,s}\right) \\
&\le \mathbb{I}\Big(\overline{r}_{i,t-1}(x) - \overline{\delta}_{i,t-1}(x)\overline{\lambda}_{t-1} + \hat{c}_{t-1,s'} \\
&\qquad > \overline{r}_{t-1}^*(x) - \overline{\delta}_{t-1}^*(x)\overline{\lambda}_{t-1} + \hat{c}_{t-1,s}, \\
&\qquad \lambda^* \ge \overline{\lambda}_{t-1} \ge \lambda^* - 2c_{t-1,T_{u^*}(t-1)}\Big) \\
&\quad + \mathbb{I}\left(\overline{\lambda}_{t-1} < \lambda^* - 2c_{t-1,T_{u^*}(t-1)}\right) + \mathbb{I}\left(\lambda^* < \overline{\lambda}_{t-1}\right).
\end{aligned}$$

The expectations of the second two terms will be bounded by Proposition 4. The first term, multiplied by $Z_t(s,s')$ is bounded by

$$\begin{aligned}
&Z_t(s,s')\mathbb{I}\Big(\overline{r}_{i,t-1}(x) - \overline{\delta}_{i,t-1}(x)\lambda^* + \hat{c}_{t-1,s'} \\
&\qquad > \overline{r}_{t-1}^*(x) - \overline{\delta}_{t-1}^*(x)(\lambda^* - 2c_{t,s}) + \hat{c}_{t-1,s}\Big).
\end{aligned}$$

When this expression equals one then at least one of the following events hold:

$$\begin{aligned}
&A_{t,s,s'} = \\
&\{\overline{r}_{t-1}^*(x) - \overline{\delta}_{t-1}^*(x)\lambda^* \le r^*(x) - \delta^*(x)\lambda^* - c'_{t-1,s}, Z_t(s,s') = 1\}, \\
&B_{t,s,s'} = \\
&\{\overline{r}_{i,t-1}(x) - \overline{\delta}_{i,t-1}(x)\lambda^* \ge r_i(x) - \delta_i(x)\lambda^* + \hat{c}_{t-1,s'}, Z_t(s,s') = 1\}, \\
&C_{t,s,s'} = \{r^*(x) - \delta^*(x)\lambda^* < r_i(x) - \delta_i(x)\lambda^* + 2\hat{c}_{t,s'}\}.
\end{aligned}$$

Here $c'_{t-1,s} = \hat{c}_{t-1,s} - 2\overline{\delta}_t^* c_{t-1,s}$. Now let us give the choices for the confidence intervals. Define

$$u_{ts} = \sqrt{\log\left(t\sqrt{|\mathcal{U}|+1}\right)/s}.$$

We have already defined c_{ts} in Proposition 4: $c_{ts} = \sqrt{2c_1}u_{ts}$, where c_1 was defined there, too. We define $\hat{c}_{ts}$ implicitly, through a definition of c'_{ts} which is defined so as to keep the probability of $A_{t,s,s'}$ small: Let

$$a_0 = \sqrt{8\max\{(r_{\max} - r_{\min})^2, r_{\max}^2(\delta_{\max} - \delta_{\min})/\delta_{\min}^2\}},$$

$c'_{ts} = a_0 u_{ts}$. and $a_1 = \sqrt{2\delta_{\max}^2 c_1}$. Define

$$a = (a_0 + a_1)^2, \qquad (7)$$

and $\hat{c}_{ts} = (a_0 + a_1)u_{ts}$. Using these definitions we bound the probabilities of the above three events. We start with $A_{t,s,s'}$:

$$\begin{aligned}
\mathbb{P}\left(A_{t,s,s'}\right) &\le \mathbb{P}\left(c'_{ts}/2 \le r^*(x) - \overline{r}_t^*(x), Z_t(s,s') = 1\right) \\
&\quad + \mathbb{P}\left(c'_{ts}/(2\lambda^*) \le \overline{\delta}_t^*(x) - \delta^*(x), Z_t(s,s') = 1\right) \\
&\le \exp\left(-c'^2_{ts}\, s/(2(r_{\max} - r_{\min})^2)\right) \\
&\quad + \exp\left(-c'^2_{ts}\, s/(2(\lambda^*)^2(\delta_{\max} - \delta_{\min})^2)\right)
\end{aligned}$$

Here we used that $\sum_{s=1}^{t} \mathbb{I}\left(I_t = i, x_t = x\right) r_t$, $\sum_{s=1}^{t} \mathbb{I}\left(I_t = i, x_t = x\right) \delta_t$ are martingales for any x, i, and the above-mentioned variant of the Hoeffding-Azuma inequality. Plugging in the definition of c'_{ts} we get that the probability of event $A_{t,s,s'}$ is bounded by $2t^{-4}(|\mathcal{U}|+1)^{-2}$. The probability of $B_{t,s,s'}$ can be bounded in the same way and by the same expression since $\hat{c}_{ts} > c'_{ts}$. Therefore

$$\begin{aligned}
&\sum_{t=1}^{n} \sum_{(s,s')\in H(t)} \left[\mathbb{P}\left(A_{t,s,s'}\right) + \mathbb{P}\left(B_{t,s,s'}\right)\right] \\
&\quad \le \sum_{t=1}^{n} \sum_{(s,s')\in H(t)} \frac{4}{t^4(|\mathcal{U}|+1)^2} \le \frac{2\pi^2}{3(|\mathcal{U}|+1)^2}.
\end{aligned}$$

Moreover, define $A(t,x) = a\log(t(\sqrt{|\mathcal{U}|+1}))/\Delta_i(x)^2$. Now, if $C_{t,s,s'}$ holds then one must have $\Delta_i(x) > 2\hat{c}_{t,s'}$,

where $s' \geq A(t,x)$. The above choice makes s' large enough so that $\Delta_i(x) > 2\hat{c}_{t,s'}$ cannot hold. Hence $P(C_{t,s,s'}) = 0$. Gathering all the terms, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\left(u_t^{\mathrm{UCB}}(x) = i, x_t = x\right)\right] \\
&\leq \sum_{t=1}^{n} \mathbb{P}\left(\overline{\lambda}_{t-1} < \lambda^* - 2c_{t-1,T_{u^*}(t-1)}\right) + \mathbb{P}\left(\lambda^* < \overline{\lambda}_{t-1}\right) \\
&\quad + \sum_{t=1}^{n} \sum_{(s,s')\in H(t)} \left[\mathbb{P}\left(A_{t,s,s'}\right) + \mathbb{P}\left(B_{t,s,s'}\right)\right] + A(n,x) \\
&\leq 2(\log(n)+1) + \frac{2\pi^2}{3(|\mathcal{U}|+1)^2} + \frac{a\log\left(n(\sqrt{|\mathcal{U}|+1})\right)}{\Delta_i(x)^2}.
\end{aligned}$$

This finishes the proof of (6) and hence, by Proposition 2 we get the desired bound, (2). □

4 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 $O(\log n)$ less than that of the best policy in hindsight. The algorithm is based on the upper confidence bound idea of Auer et al. [2002]. Our algorithm, however, extends their UCB1 algorithm proposed for the multi-armed bandit problem in two ways. First of all, it estimates the long term maximum reward per unit time. For this we proposed to adopt a maximin approach: The estimate was chosen to be the optimal gain that can be guaranteed in the worst-case, with high probability, given all the data seen so far. Moreover, utilizing the structure of the problem the algorithm chooses its actions based on the sufficient statistics of the problem instead of considering each policy separately. Note that doing so would lead to a constant factor in the regret bound that grows linearly with the number of possible policies, i.e., exponentially in the size of the problem. On the other hand, because of the specialized form of our algorithm, the constants in our bound depend only polynomially on these parameters. However, we expect that the explicit dependence of the bound on the number of possible side information values can be relaxed. Note however, that we have not attempted any optimization of the actual constants that appear in our bounds. Therefore, we expect that our constants can be improved easily.

One 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. However, we would like to note that the problem of computing this quantity is very similar to computing the value of minimax Markov games. In fact, the actual definition of $\overline{\delta}_t$ is not that important: Any estimate that satisfies the conclusion of Proposition 4 would do. We speculate that since efficient methods are available for certain minimax Markov games (cf. [Szepesvári and Littman, 1999]), game theoretic techniques might yield an algorithm that not only utilizes the available information effectively, but is also computationally efficient.

In the present work we restricted ourselves to the case when the side information is allowed to take values only in a finite set. Assuming appropriate smoothness conditions on the reward and delay functions, it seems possible to extend the algorithm to the case of continuous valued side information. 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. *Adv. in Appl. 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.

[Devroye *et al.*, 1996] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, 1996.

[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. *Ann. of Appl. 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.