Leveraging Observational Learning for Exploration in Bandits

Extended Abstract

Andrei Lupu
McGill University
Montreal, Canada

Audrey Durand
McGill University
Montreal, Canada

Doina Precup
McGill University
Montreal, Canada

KEYWORDS

Observational learning; imitation learning; bandits

ACM Reference Format:
Andrei Lupu, Audrey Durand, and Doina Precup. 2018. Leveraging Observational Learning for Exploration in Bandits. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

Learning from a target has been tackled in the reinforcement learning (RL) setting [1, 7] as *imitation learning*, either through behaviour cloning or inverse RL. In the former, the agent regresses directly onto the policy of a target [5], while in the latter, the agent infers a reward function from the behaviour of other agents and optimizes this function [6]. Extending upon these notions, *observational learning* was recently introduced in RL as the ability for an agent to modify its behavior or to acquire information as an effect of observing another agent sharing its environment [3]. In this work, we study the observational learning problem under the bandit setting. More specifically, we consider a learner (agent) that has access to actions performed by a target policy in the same environment. The agent only observes the target's actions, but not their associated rewards. Note that the target actions can in fact be performed by several other agents. This should not be confused with cooperative bandits [4], where several agents share knowledge with each other regarding the actions and obtained rewards.

For this purpose, we introduce an algorithm based on the vanilla Upper Confidence Bound (UCB) algorithm [2], which we call Target-UCB. The core idea involves an action selection process influenced by the popularity of each action according to the target. We provide a theoretical bound on the performance of Target-UCB given the *quality* of the target (in terms of convergence rates and probability of selecting the optimal action). The obtained results in several bandit problems suggest that using this data can lead to much faster learning. More specifically, we show that unless the target is 100% wrong, Target-UCB will manage to cumulate logarithmic regret. They also point to some interesting behaviors in settings in which the target comes form multiple agents.

2 PROBLEM SETTING

We consider a bandit problem where $\mathcal{A}$ denotes the set of possible actions and $A := |\mathcal{A}|$ is the number of actions. Each action $a \in \mathcal{A}$ is associated with an unknown expected payoff μ_a. On each episode $t \geqslant 1$, the agent selects an action $a_t \in \mathcal{A}$ and observes reward $r_t \sim \nu(\mu_{a_t})$, where $\nu(\mu)$ denotes a probability distribution of mean μ. Let $\star := \text{argmax}_{a \in \mathcal{A}} \mu_a$ denote the *optimal action*. The goal of the agent is to minimize the cumulative pseudo-regret after T episodes:

$$\mathfrak{R}(T) := \sum_{t=1}^{T-1} (\mu_\star - \mu_{a_t}). \tag{1}$$

From now on, the term "regret" will refer to "pseudo-regret".

In observational learning bandits, the agent has access to the actions performed by an unknown *target* policy, but does not observe the associated rewards. Since the target is not aware that it is watched by the learner and is not meant to teach, it does not need to be a single entity. The so-called target can correspond to a policy describing the general behaviour of several other agents, or *neighbours*.

3 ALGORITHM

Let $N_{a,t}$ and $\tilde{N}_{a,t}$ denote number of times that action a was played up to time t (exclusively) by the player and by the target policy, respectively. Also let $m_{a,t}$ denote the empirical average given rewards obtained by playing action a up to time t (exclusively). Note that $m_{a,t}$ is computed on the rewards obtained by the player, *not by the target policy*. Formally,

$$N_{a,t} := \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} \quad \text{and} \quad m_{a,t} := \frac{1}{N_{a,t}} \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} r_s.$$

We introduce Target-UCB, a UCB-like algorithm that adjusts its optimism with respect to a specific action given how much attention this action has received from the target policy. The idea is to be optimistic for actions that the agent running Target-UCB has played less than the target policy. Algorithm 1 provides the Target-UCB routine for rewards in $[0, 1]$ (e.g., Bernoulli rewards). Under the following assumption, Theorem 3.1 provides a bound on the expected cumulative pseudo-regret given the performance of the target policy.

Assumption 1 (Optimal plays by the target policy.). *The target policy plays such that there exists some constants $\alpha \in (0, 1]$ and c_Δ for which, $\forall a \in \mathcal{A}, a \neq \star, \forall t \geqslant c_\Delta$,*

$$\tilde{N}_{\star,t} \geqslant \left(\frac{C}{C - 3/2}\right) \frac{6 \ln t}{\Delta_a^2} \quad \text{and} \quad \tilde{N}_{\star,t} \geqslant \frac{\alpha}{1 - \alpha} \tilde{N}_{a,t}.$$

Remark 1. *The constant c_Δ depends on the sub-optimality gap and the target policy, but not on t.*

Theorem 3.1. *Consider $\mathcal{A} = \{\star, a\}$ and rewards in $[0, 1]$, and assume that the target policy satisfies Assumption 1. Then, for $\alpha \in$*

[1]Recall that $a \vee b$ and $a \wedge b$ respectively denote taking the maximum and minimum value between a and b.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

Algorithm 1 Target-UCB for rewards in $[0, 1]$.

Parameters: constant $C > 3/2$.
Initialization: play each action once, s.t. $N_{a,A} = 1\ \forall a \in \mathcal{A}$.
for all $t \geqslant A + 1$ **do**
play action defined as[1]:

$$a_t = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, m_{a,t} + \underbrace{\sqrt{\frac{C \ln t}{N_{a,t}}}}_{\text{estimation optimism}} \underbrace{\sqrt{\frac{\tilde{N}_{a,t} - N_{a,t}}{\tilde{N}_{a,t}} \vee 0}}_{\text{target optimism}}$$

obtain reward r_t
update empirical mean $m_{a_t,t}$ and count $N_{a_t,t}$
update count $\tilde{N}_{a_t,t} \forall a \in \mathcal{A}$ based on target plays
end for

$(0, 1]$, *the expected cumulative regret (Eq. 1) of Target-UCB (Alg. 1) with* $C > 3/2$ *is bounded by*

$$\mathbb{E}[\Re(T)] \leqslant \Delta_a(c_\Delta + \pi^2/3) + \left(\frac{C}{2}\right)\frac{4 \ln T}{\Delta_a}$$

if $\tilde{N}_{a,T} < \left(\frac{C}{2}\right)\frac{4 \ln T}{\Delta_a^2}$*; it is bounded by*

$$\mathbb{E}[\Re(T)] \leqslant \Delta_a \mathbb{E}[\tilde{N}_{a,T}] \wedge \Delta_a(c_\Delta + \pi^2/3) + \left(\frac{C}{2}\right)\frac{12 \ln T}{\Delta_a}$$

if $\tilde{N}_{a,T} \leqslant \left(\frac{C}{2}\right)\frac{12 \ln T}{\Delta_a^2}$ *or* $\alpha \geqslant \frac{1}{2}$*; otherwise it is bounded by*

$$\mathbb{E}[\Re(T)] \leqslant \Delta_a \mathbb{E}[\tilde{N}_{a,T}] \wedge \Delta_a(c_\Delta + \pi^2/3) + \left(\frac{C}{2}\right)\frac{\left(1 + \sqrt{2} + \sqrt{\frac{1-\alpha}{\alpha}}\right)^2 \ln T}{\Delta_a}.$$

This result is comparable to the cumulative regret upper-bound of UCB [2]. More specifically, UCB has the term $8 \ln T/\Delta_a$. Therefore, we would expect a Target-UCB to outperform UCB when the target policy is *good enough*. This intuition is supported by empirical results.

4 NUMERICAL EXPERIMENTS

The following experiments evaluate the potential of Target-UCB ($C = 2$) in various settings. Bernoulli reward distributions are used in all experiments. All the results are obtained by averaging over 2000 independent runs.

Figure 1 shows the cumulative regret for an α-optimal target which plays the optimal action with probability α, a greedy follower which always selects the action chosen most often so far by the target, and Target-UCB, for different values of α. We observe that the convergence of Target-UCB is influenced by the quality of the target policy – it converges much faster for a larger α. However, note that Target-UCB still converges even for a *bad* target (low α), which is not the case for the greedy follower that blindly follows the target. This is due to the properties of Target-UCB, according to which the influence of the target's optimism necessarily decreases as more actions are played by the learner. As long as the target is not 100% wrong ($\alpha = 0$), Target-UCB is able to learn something. This is important as we may not be able to guarantee a learning rate for every agent encompassed under the target function, for example in a multi-agent setting.

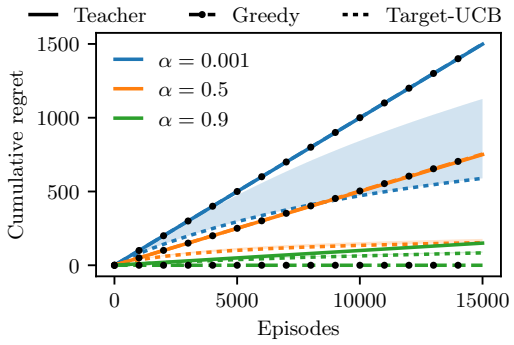

Figure 1: Target-UCB vs Greedy with an α-optimal target on a 2-actions setting ($\mu_\star = 0.9$, $\Delta_a = 0.1$). Standard deviation of UCB and greedy are omitted for clarity.

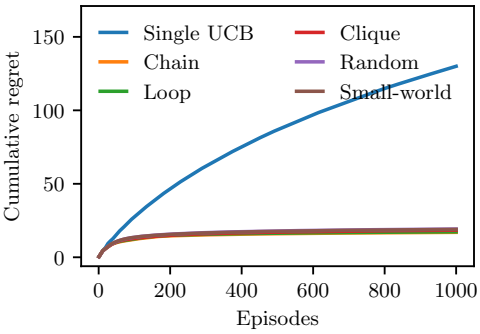

Figure 2: Single UCB vs Target-UCB graphs of 20 agents on randomly generated 10-actions settings.

We then evaluate the potential of improvement in multi-agent settings, where all agents in a graph follow the Target-UCB policy and use the empirical average of the actions taken by their neighbours as the target policy. Note that the greedy follower baseline is not available anymore, as it requires its own target. Figure 2 shows that Target-UCB graphs consistently achieve a much lower regret than a single UCB agent. Recall that there is no explicit information sharing between the Target-UCB agents. These results thus show the potential of a fully decentralized multi-agent system.

5 FUTURE WORKS

This work studies the benefits and tradeoffs of using observational data in the exploration-exploitation dilemma highlighted by the bandit setting. It is especially interesting from the perspective of considering humans as targets in a human-robot interaction setting, where it is not easy to precisely quantify the human behaviour in terms of regret convergence. An important point that has not been addressed here is the explicit ability to detect when following the target is not efficient. Indeed, learning from a bad target can lead to larger regret (even though still logarithmic) than using a vanilla UCB. Being able to characterize the quality of the target as a target could help in avoiding this situation.

REFERENCES

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

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

[3] D. Borsa, B. Piot, R. Munos, and O. Pietquin. 2017. Observational Learning by Reinforcement Learning. *arXiv preprint arXiv:1706.06617* (2017).

[4] P. Landgren, V. Srivastava, and N. E. Leonard. 2016. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*. 167–172.

[5] N. Ratliff, J. A. Bagnell, and S. S. Srinivasa. 2007. Imitation learning for locomotion and manipulation. In *Proceedings of the 7th IEEE-RAS International Conference on Humanoid Robots*. 392–397.

[6] S. Russell. 1998. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*. 101–103.

[7] S. Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences* 3, 6 (1999), 233–242.