Leveraging Observations in Bandits: Between Risks and Benefits

Andrei Lupu
School of Computer Science
McGill University
andrei.lupu@mail.mcgill.ca

Audrey Durand
School of Computer Science
McGill University
audrey.durand@mcgill.ca

Doina Precup
School of Computer Science
McGill University
dprecup@cs.mcgill.ca

Abstract

Imitation learning has been widely used to speed up learning in novice agents, by allowing them to leverage existing data from experts. Allowing an agent to be influenced by external observations can benefit to the learning process, but it also puts the agent at risk of following sub-optimal behaviours. In this paper, we study this problem in the context of bandits. More specifically, we consider that an agent (learner) is interacting with a bandit-style decision task, but can also observe a target policy interacting with the same environment. The learner observes only the target's actions, not the rewards obtained. We introduce a new bandit optimism modifier that uses conditional optimism contingent on the actions of the target in order to guide the agent's exploration. We analyze the effect of this modification on the well-known Upper Confidence Bound algorithm by proving that it preserves a regret upper-bound of order $\mathcal{O}(\ln T)$, even in the presence of a very poor target, and we derive the dependency of the expected regret on the general target policy. We provide empirical results showing both great benefits as well as certain limitations inherent to observational learning in the multi-armed bandit setting. Experiments are conducted using targets satisfying theoretical assumptions with high probability, thus narrowing the gap between theory and application.

1 Introduction

The imitating behaviour of human beings has been studied for a long time in psychology and cognitive sciences (Miller and Dollard 1941; Bandura and Walters 1963). Our societies have exploited the human capability to learn from a teacher in order to increase learning speed. Imitation behaviour has also been studied in economics (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1998) in order to explain phenomena such as *herding*, in which individuals forget about their own experiences, relying only on others' actions to guide their own behavior. Learning from a teacher has been tackled in reinforcement learning (RL) (Schaal 1999; Argall et al. 2009) through *imitation learning* algorithms, such as behaviour cloning or inverse RL. In the former, the agent trains by using regression of target actions provided by a teacher policy (Ratliff, Bagnell, and Srinivasa 2007), while in the latter, the agent infers a reward function from the behaviour of other agents, then optimizes it (Russell 1998).

Observational learning, also known as *social learning*, has recently been introduced in RL to model the ability of an agent to modify its behavior or to acquire information by observing another agent sharing its environment (Borsa et al. 2017). Unlike typical imitation learning, observational learning does not strictly lead to a duplication of the behavior exhibited by the teacher (Bandura and Walters 1963; Bandura 1977). More precisely, observational learning is characterized by the following principles (Borsa et al. 2017): the agent observes the teacher through its own perception of the environment; the agent is only rewarded for performing the task and doesn't receive any incentive to imitate the teacher; the teacher is not aware that it is watched by the learner and does not intentionally teach or provide extra information to the learner. We highlight these differences with imitation learning by rather referring to the teacher as a *target*.

In this work, we study the observational learning problem in the context of bandits, the simplest setting for studying the explore-exploit trade-off faced by an agent in an unknown environment. We consider a learner (agent) that observes actions performed by a target policy in the same environment, but not their associated rewards. Note that the target actions can in fact be performed by several other agents. When collected from a good target, this data can potentially improve the behaviour of the learner, specifically, by speeding up the learning process. Consequently, we would like an agent equipped with the ability to leverage it whenever available. This should not be confused with cooperative bandits (Landgren, Srivastava, and Leonard 2016), where several agents share knowledge with each other regarding the actions and obtained rewards.

Human imitative behaviour in social learning experiments has been studied extensively in the bandits setting, when a learner can observe both the actions and rewards of other agents, e.g. (Schlag 1998; Rendell et al. 2010; Toyokawa, Kim, and Kameda 2014). The observational learning bandits setting provides a framework for extending social learning experiments to situations where the reward of peers is not available to agents. This setting arises naturally in human interactions. Specifically, our work was motivated by a real psychology dataset in which teens can observe the behaviour of peers (such as consuming certain soft drinks) but cannot observe the actual reward (enjoyment) directly. This

is a more realistic scenario, since explaining internal rewards to another human can be both daunting and imprecise.

The bandits setting is also extensively used in marketing, for example for optimizing ads placement (Schwartz, Bradlow, and Fader 2017). The problem we study arises naturally in the context of "ad transparency" centres, such as those established by Facebook and Twitter, which require advertisers to reveal all ads run by their page, along with information regarding the targeted demographic. This means that a competitor can observe the actions (ads placed) but not the rewards (number of clicks on these ads). A startup making sportsware, for instance, could use information about the ad campaign of a famous sport company in order to shape its own advertisement.

In this work, we tackle observational learning in bandits by introducing a new factor, which we denominate *target optimism*. The idea is to leverage the popularity of each action according to the target during the action selection process. To study whether using this additional information can potentially accelerate learning while being robust against *poor* targets, we consider a variant of the well-known Upper Confidence Bound (UCB) algorithm (Auer, Cesa-Bianchi, and Fischer 2002), which enjoys a well-understood analysis. This yields an altered algorithm, Target-UCB, for which we provide theoretical guarantees on the performance given the target *quality* (in terms of convergence rates and probability of selecting the optimal action). We also provide some theoretical insights regarding what makes a good target. More precisely, we show that unless the target is always wrong, Target-UCB is considered robust in that it will manage to maintain logarithmic regret. Moreover, if the target performs below a given threshold, Target-UCB will necessarily outperform its target. Our empirical results suggest that using target data can yield vast learning improvements over the unaltered algorithm, and also show that we can successfully leverage target actions from multiple other agents.

2 Problem setting

We consider the stochastic 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 $t \geq 1$, the agent selects an action $a_t \in \mathcal{A}$ and observes reward $r_t \sim \nu(\mu_{a_t})$, where $\nu(\mu)$ is a probability distribution of mean μ. Let $\star := \operatorname{argmax}_{a \in \mathcal{A}} \mu_a$ 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*. The goal of the learner is still to minimize the cumulative regret (Eq. 1).

Related work The closest setting is probably the one where a predictor (akin to the target in the observational setting) initially provides a probability distribution over actions (Rosin 2011). This distribution is provided once, in the beginning, and is then fixed for the whole horizon of the game. This could be seen as an instance of the observational setting with a stationary target policy, with the additional difficulty that the target distribution would need to be empirically estimated.

3 Target optimism

Let $N_{a,t}$ and $\tilde{N}_{a,t}$ denote the number of times that action a was played up to time t (exclusively) by the player and by the target policy, respectively. Formally,

$$N_{a,t} := \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}.$$

To leverage the observational data available in this setting, we introduce the target optimism[1]:

$$\sqrt{\frac{\tilde{N}_{a,t} - N_{a,t}}{\tilde{N}_{a,t}} \vee 0}. \tag{2}$$

This quantity, which is bounded in $[0, 1]$, measures the discrepancy between the actions taken by the agent and the ones taken by the target. Target optimism for a given action a increases as the target selects action a and the learning agent selects other actions $a' \neq a$. Note that the target optimism is only greater than 0 for actions that are less played by the learning agent than by the target. Incorporating this quantity into an agent behaviour, the idea would be to bias the agent toward selecting actions which were previously chosen more often by the target policy. In this work, we study the potential gains and risk of target optimism as part of the seminal UCB algorithm.

4 Target-UCB

Let $m_{a,t}$ denote the empirical average of the rewards obtained by the learning agent while playing action a up to time t (exclusively). Note that $m_{a,t}$ does not contain any information about the rewards obtained by the target policy. The standard UCB (Auer, Cesa-Bianchi, and Fischer 2002) policy for reward distributions with support in $[0, 1]$ plays each action once and then selects action

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}} m_{a,t} + \sqrt{\frac{2 \ln t}{N_{a,t}}} \tag{3}$$

for $t > A$. Thanks to its well-understood theoretical analysis, UCB has been well studied in the literature, where

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

it has been extended to linear (Li et al. 2010), switching (Garivier and Moulines 2011), and combinatorial bandit settings (Chen, Wang, and Yuan 2013), to name a few. By adding target optimism (Equation 2) to the UCB algorithm, we obtain a new variant: Target-UCB. This one allows to incorporate observational information in order to adjust the typical UCB 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 learning agent (running Target-UCB) has played less than the target policy. Algorithm 1 outlines Target-UCB for reward distributions with support $[0, 1]$ (e.g., Bernoulli rewards).

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

Parameters: action set $\mathcal{A}$ and constant $C > 3/2$.
Initialization: play each action once.
for all $t \geq A + 1$ **do**
  Play action defined as:

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

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

Note that there are two estimation *modes* for a given action a. If $N_{a,t} \geq \tilde{N}_{a,t}$, meaning that a has been played at least as much by Target-UCB as by the target, Target-UCB evaluates a based only on its empirical average. We say that Target-UCB is being *realistic* with regards to a. On the other hand, if $N_{a,t} < \tilde{N}_{a,t}$, Target-UCB evaluates a based on its inflated empirical average. In this case, we say that Target-UCB is being *optimistic*.

Intuition

We distinguish two parts in the optimism term of Target-UCB: the *target optimism* introduced previously and the *estimation optimism*. One can see that the seminal UCB (Equation 3) algorithm is a special case of Target-UCB where $C = 2$ and $\tilde{N}_{a,t} = \infty$ for all a, t, giving full focus on estimation optimism. Rather than using optimism solely to overcome uncertainty in reward estimation (through empirical means), Target-UCB relies on optimism to compensate for uncertainty in its own policy, compared with the target policy. This optimism pushes Target-UCB to explore actions that might be good given the additional attention that they received from the target policy. One might see Target-UCB, when realistic, as a greedy policy that chooses to explore when the target policy makes it doubt its own choices (becoming optimistic).

Making Target-UCB optimistic for a given action requires both estimation optimism and target optimism at the same time for this action. For low values of $N_{a,t}$, target optimism rapidly tends to one, such that Target-UCB falls back to a UCB behaviour for action a, fully using estimation optimism to compensate for a possible under-estimation of $m_{a,t}$. However, as $N_{a,t}$ grows closer to $\tilde{N}_{a,t}$, Target-UCB is allowed to be less optimistic than UCB would be (with respect to action a). Receiving such guidance from a target policy may allow Target-UCB to reduce its optimism compared with UCB. This is shown by regret upper-bounds, which we present next.

Regret upper-bound

We consider bandit settings with A actions and reward distributions with support in $[0, 1]$. Let $\Delta_a := (\mu_\star - \mu_a)$ denote the gap between action a and optimal action $\star$ and let $\Delta = \min_{a \in \mathcal{A}} \Delta_a$. Under the following assumption, Theorem 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 $\alpha_a \in (0, 1]$ and c_Δ for which, $\forall a \in \mathcal{A}, a \neq \star, \forall t \geq c_\Delta$,*

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

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

Remark 2. *If the 1st part of Assumption 1 is satisfied for action a, then there must exist an α_a satisfying the 2nd part.*

Remark 3. *If the 1st part of Assumption 1 is satisfied for the smallest gap Δ, then it is also satisfied for all gaps Δ_a.*

Theorem 1. *Consider rewards in $[0, 1]$ and assume that the target policy satisfies Assumption 1. Then, for $\alpha \in (0, 1]$, the expected cumulative regret (Eq. 1) of Target-UCB (Alg. 1) with $C > 3/2$ is bounded as follows:*
If $\tilde{N}_{a,T} < \frac{6 \ln T}{\Delta_a^2}$ for all $a \in \mathcal{A}, a \neq \star$,

$$\mathbb{E}[\mathfrak{R}(T)] \leq \sum_{\substack{a \in \mathcal{A} \\ a \neq \star}} \left[\frac{6 \ln T}{\Delta_a} + \Delta_a \Big(c_\Delta + \frac{\pi^2}{3} \Big) \right];$$

if $\tilde{N}_{a,T} \leq \frac{(\sqrt{6} + \sqrt{C})^2 \ln T}{\Delta_a^2}$ or $\alpha_a \geq \frac{1}{2}$ for all $a \in \mathcal{A}, a \neq \star$,

$$\mathbb{E}[\mathfrak{R}(T)] \leq \sum_{\substack{a \in \mathcal{A} \\ a \neq \star}} \min \begin{cases} \Delta_a \mathbb{E}[\tilde{N}_{a,T}] \\ \Delta_a \big(c_\Delta + \frac{\pi^2}{3} \big) + \frac{(\sqrt{6} + \sqrt{C})^2 \ln T}{\Delta_a} \end{cases};$$

otherwise it is bounded by

$$\mathbb{E}[\mathfrak{R}(T)] \leq \sum_{\substack{a \in \mathcal{A} \\ a \neq \star}} \min \begin{cases} \Delta_a \mathbb{E}[\tilde{N}_{a,T}] \\ \Delta_a \big(c_\Delta + \frac{\pi^2}{3} \big) + \\ \qquad \frac{(\sqrt{\frac{3}{2}} + \sqrt{C} + \sqrt{\frac{3}{2}} \sqrt{\frac{1 - \alpha}{\alpha}})^2 \ln T}{\Delta_a} \end{cases}.$$

This result shows that, when following a UCB target, the proposed approach cannot do worse than UCB – it can only improve upon it. More specifically, UCB has the term $8 \ln T / \Delta_a$. Therefore, Target-UCB necessarily outperforms UCB when the target policy is *good*. When following *bad* targets, the theorem does indicate that Target-UCB could

achieve a potentially worse bound than UCB. However, as long as the target policy satisfies Assumption 1, Target-UCB maintains regret of order $\mathcal{O}(\ln T)$. Consequently, leveraging observations through target optimism is not meant as a hard guarantee to outperform classical bandit algorithms for all targets. Instead, it is a valuable tool to be used when the available data is expected to be of some minimal quality, while also providing robustness in case of uncertainty about the performance of the target policy. This intuition is supported by empirical results (see Section 6).

Here follows a proof sketch of Theorem 1. The complete proof can be found in the supplementary material[2].

Proof outline

We can express the cumulative regret (Equation 1) as $\mathfrak{R}(T) = \sum_{a \in \mathcal{A}} \Delta_a N_{a,T}$. This quantity can be bounded by controlling the number of sub-optimal plays $N_{a,T}$. Let us introduce the following events to characterize the concentration of the empirical means.

Definition 1. *Let E_t^a and $E_t^\star$ respectively denote the events in which $m_{a,s} - \mu_a \leq \sqrt{\frac{3 \ln t}{2 N_{a,s}}}$ and $\mu_\star - m_{\star,s} \leq \sqrt{\frac{3 \ln t}{2 N_{\star,s}}}$ simultaneously $\forall s \leq t$.*

The idea is to decompose

$$N_{a,T} \leq \ell + \sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a, E_t^a, E_t^\star, N_{a,t} \geq \ell\} + \sum_{t=A+1}^{T-1} \mathbb{I}\{\bar{E}_t^a \cup \bar{E}_t^\star\}$$

and control the two sums separately. Focusing on the first sum, cumulating sub-optimal plays under the occurrence of events E_t^a and $E_t^\star$, we consider two situations:

- Target-UCB being *better* than the target policy with respect to action a ($N_{a,t}/\tilde{N}_{a,t} < 1$) and
- Target-UCB being *worse* than the target policy with respect to action a ($N_{a,t}/\tilde{N}_{a,t} \geq 1$).

Also recall that sub-optimal action a is played if

$$m_{a,t} + \sqrt{\frac{C \ln t}{N_{a,t}}} \sqrt{1 - \left(\frac{N_{a,t}}{\tilde{N}_{a,t}} \wedge 1\right)} \geq m_{\star,t} + \sqrt{\frac{C \ln t}{N_{\star,t}}} \sqrt{1 - \left(\frac{N_{\star,t}}{\tilde{N}_{\star,t}} \wedge 1\right)}.$$

We deduce that, under events E_t^a and $E_t^\star$, selecting action a at episode t requires that

$$\Delta_a \leq \sqrt{\frac{3 \ln t}{2 N_{a,t}}} + \sqrt{\frac{C \ln t}{N_{a,t}}} \sqrt{1 - \left(\frac{N_{a,t}}{\tilde{N}_{a,t}} \wedge 1\right)} \quad (4)$$

$$+ \sqrt{\frac{3 \ln t}{2 N_{\star,t}}} - \sqrt{\frac{C \ln t}{N_{\star,t}}} \sqrt{1 - \left(\frac{N_{\star,t}}{\tilde{N}_{\star,t}} \wedge 1\right)}. \quad (5)$$

[2] https://github.com/lupuandr/Target-UCB/blob/master/supplemental.pdf

Now, the count of optimal plays by Target-UCB could be in one of two cases: $N_{\star,t}/\tilde{N}_{\star,t} \leq 1 - \frac{3}{2C}$ and $N_{\star,t}/\tilde{N}_{\star,t} \geq 1 - \frac{3}{2C}$. Combining this with Target-UCB being better or worse than the target (with respect to action a) results in four situations. Using elementary algebra, we can derive upper bounds on $N_{a,t}$ and $N_{\star,t}$ that are required in order to satisfy Equation 4 for each of these four cases. Then, we can use use Assumption 1 and pick ℓ such that Equation 4 cannot be satisfied anymore. Finally, by showing that the probability of nonoccurrence of events is controlled by the definition of events, we obtain Theorem 1.

5 Target policy

The only requirement for the target policy is summarized by Assumption 1. Note that this assumption can be satisfied by any target that plays action $\star$ *once in a while*, at the price of a larger c_Δ. Let us look at some candidate policies that could constitute valid targets. Complete proofs for each of these candidates are provided in the supplementary material.

UCB

Since UCB begins by selecting each action at least once, the condition $(1 - \alpha_a)\tilde{N}_{\star,t} \geq \alpha_a \tilde{N}_{a,t}$ is necessarily satisfied. Also, since UCB is known to enjoy sub-linear regret, there must exists a time c_Δ such that $c_\Delta - \sum_{a \in \mathcal{A}} \tilde{N}_{a,c_\Delta} \geq \left(\frac{C}{C - 3/2}\right) \frac{6 \ln t}{\Delta_a^2}$. As a concrete illustration, the expected number of sub-optimal plays of action a after t episodes using target policy UCB is upper bounded by:

$$\mathbb{E}[\tilde{N}_{a,t}] \leq \frac{8 \ln t}{\Delta_a^2} + 1 + \frac{\pi^2}{3}.$$

By an application of Azuma-Hœffding's inequality for martingales, we obtain that

$$\tilde{N}_{a,t} \leq \frac{8 \ln t}{\Delta_a^2} + 1 + \frac{\pi^2}{3} + \sqrt{2t \ln(1/\delta)}$$

with probability higher than $1 - \delta$. By considering that $\tilde{N}_{\star,c_\Delta} = c_\Delta - \tilde{N}_{a,c_\Delta}$ in the 2-actions setting, we can show that there exists some time c_Δ that allows to satisfy Assumption 1 with high probability. Section 6 provides results showing that UCB policy is a good target and also that Target-UCB outperforms this target.

α-optimal

Now consider a basic family of policies that plays the optimal action with probability higher than $\alpha \in (0, 1]$. The expected number of optimal plays after t episodes using such a target policy is lower-bounded by:

$$\mathbb{E}[\tilde{N}_{\star,t}] \geq \alpha t.$$

Per definition, this satisfies the second condition of Assumption 1. By an application of Azuma-Hœffding's inequality for martingales, we obtain that

$$\tilde{N}_{\star,t} > \alpha t - \alpha \sqrt{2t \ln(1/\delta)} \qquad \text{for } \frac{1}{2} \leq \alpha \leq 1$$

$$\tilde{N}_{\star,t} > \alpha t - (1 - \alpha) \sqrt{2t \ln(1/\delta)} \qquad \text{for } 0 < \alpha < \frac{1}{2}$$

with probability higher than $1 - \delta$. By considering that $\tilde{N}_{\star,c_\Delta} = c_\Delta - \tilde{N}_{a,c_\Delta}$ in the 2-actions setting, we can find the value of c_Δ that satisfies the first condition of Assumption 1. Note that this holds with high probability. In other cases c_Δ may be higher. Section 6 provides results showing that the α-optimal policy is indeed a good target. More specifically, results show a slower convergence of Target-UCB for a lower α, but it converges nonetheless.

Remark 4. *The constant c_Δ in Assumption 1 will be larger for a lower α. Also, this requires $\alpha > 0$.*

Average of neighbours

Consider a Target-UCB agent who can observe the actions of several other agents, which we consider its "neighbours". Target-UCB can then consider a target policy that encompasses all neighbours, with $\tilde{N}_{a,t}$ corresponding to the *average* number of action plays among neighbours. Averaging neighbours is especially useful in the case in which, if taken independently, they would not all satisfy Assumption 1, but their average satisfies this assumption with high probability. More specifically, the average of neighbours can be seen as an α-optimal policy, with α depending on the number of neighbours that select the optimal action $\star$.

The power of neighbours First we consider the setting in which two Target-UCB agents are using each other as target policies. Recall that Target-UCB is optimistic for some action a only if this action has been played less by Target-UCB than by the target policy. Therefore, if two Target-UCB agents display the same, equally bad, behaviour, they will both be realistic (and stay bad). This situation happens if both algorithms are unlucky and obtain a low sample reward from the optimal action during initialization.

Now consider J algorithms (neighbours) taken together. Recall that we are assuming Bernoulli reward distributions. Hence we want to bound the probability of all J agents being unlucky on their first sample from the optimal action:

$$(1 - \mu_\star)^J \leq \delta \tag{6}$$

for $J \geq \ln(\delta)/\ln(1 - \mu_\star)$, with $\delta \in (0, 1)$. Averaging over at least J neighbours, as specified per Equation 6, ensures the convergence of Target-UCB with high probability δ.

6 Experiments

The following experiments evaluate the potential of Target-UCB (C=2) in various settings. Bernoulli reward distributions are used in all experiments. Unless indicated otherwise, all results are obtained by averaging over 2000 independent runs. In all figures, shaded areas indicate one standard deviation above the mean. We also provide an implementation of the Target-UCB algorithm: https://github.com/lupuandr/Target-UCB.

Learning better than the target

We first consider a 2-actions problem, using UCB as the target (see Section 5) in order to assess the benefits of observing a target that is also a learning agent. We therefore have

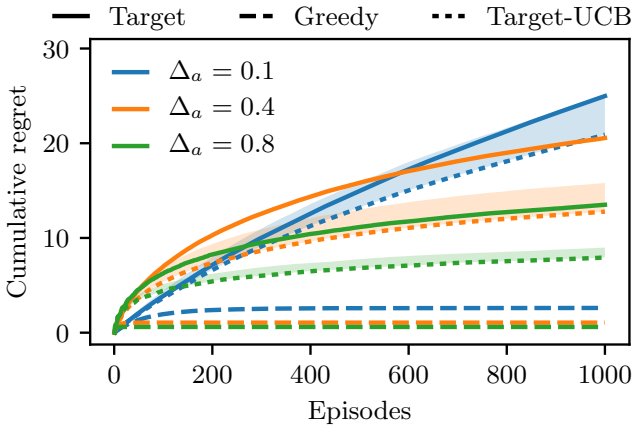

Figure 1: Target-UCB vs greedy with a UCB target on a 2-actions setting ($\mu_\star = 0.9$). Std. dev. of UCB and greedy omitted for clarity.

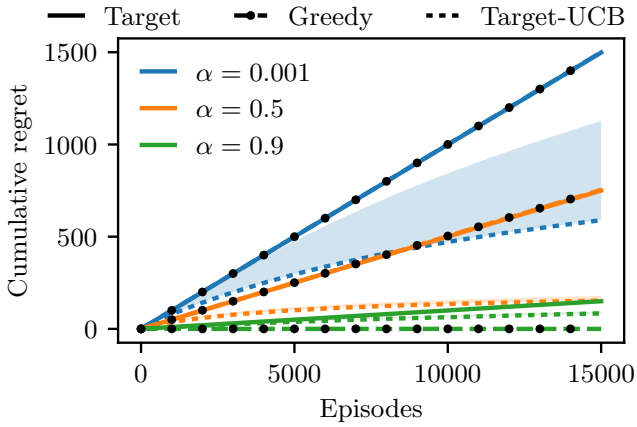

Figure 2: Target-UCB vs greedy with an α-optimal target on a 2-actions setting ($\mu_\star = 0.9$, $\Delta_a = 0.1$). Std. dev. of UCB and greedy omitted for clarity.

two agents: one UCB, who is learning solely from the environment, and one learner, who is using observational data from its target (the UCB) as well as its own rewards to shape its policy. We evaluate Target-UCB as a learner by comparison to a greedy follower, which always selects the action chosen most often so far by the target:

$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, \tilde{N}_{a,t}. \tag{7}$$

Figure 1 shows the cumulative regret for the target (UCB), the greedy follower, and Target-UCB, for different configurations of reward expectations. We observe that Target-UCB is able to outperform its target in all scenarios. We also notice that the greedy follower baseline performs even better. This is not surprising as the considered target (UCB) is good (and improving). The next experiments analyze settings where the target is less proficient.

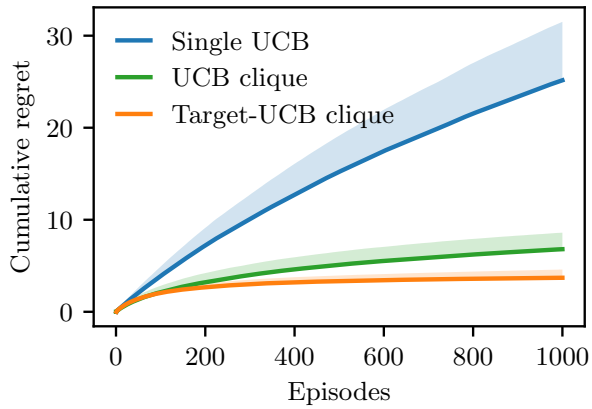

Figure 3: Single UCB and UCB clique of 11 agents vs Target-UCB clique of 11 agents on a 2-actions setting ($\mu_\star = 0.5, \Delta_a = 0.1$).

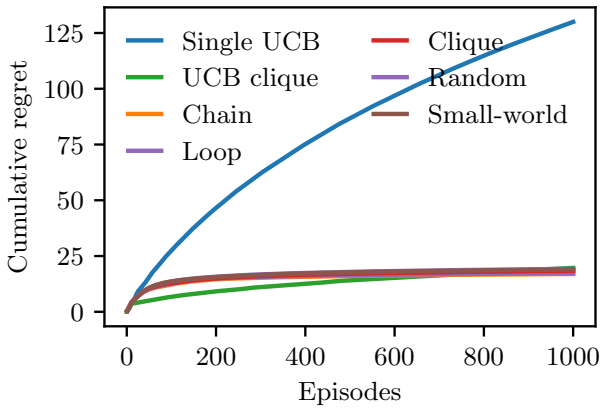

Figure 4: Single UCB and UCB clique of 20 agents vs five Target-UCB graphs of 20 agents on randomly generated 10-actions settings.

Learning from a non-learner

In the same 2-actions setting, we now consider an α-optimal policy (see Section 5) as target. Note that the α-optimal agent is not *learning* – it just plays the optimal action with probability α. As in the previous experiment, we have two agents: one target (α-optimal) and one learner. We evaluate Target-UCB as a learner, in comparison to the greedy follower (see Equation (7)).

Figure 2 shows the cumulative regret for the target (the α-optimal), the greedy follower, 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 (see Section 4), 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 and maintain logarithmic regret. 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.

Learning from Target-UCB

We now 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. We compare the cumulative regret averaged over all nodes of the graph with the cumulative regret of a single UCB agent in the same bandit problem, as well as with a clique of UCB agents sharing full information (both actions and rewards) at all steps. Since UCB is deterministic, a UCB clique is equivalent to a single UCB agent receiving n samples per arm pull, where n is the size of the clique. Note that the greedy follower baseline is not available anymore, as it requires its own target.

The first experiment considers a clique (fully connected) graph structures on 2-actions bandits with $\mu_\star = 0.5$. The clique size is selected using Equation (6) for $\delta = 0.001$. Figure 3 shows that a Target-UCB clique achieve a much lower regret than a single UCB agent. In this setting, the Target-UCB clique also outperforms the UCB one. This shows that target optimism is a very effective way of leveraging observational information to improve learning when a good target is available.

We also carried out the experiments with a variety of 20-agents graph structures, in which the agents have to solve 10-actions problems. Recall that each agent has two neighbours in loops. In chains, agents at the end have one neighbour and others have two. The small-world graph follows the *Barabási-Albert* model with at least one neighbour per agent, whereas the edges in the random graph are sampled with probability $p = 0.5$ according to the *Erdős-Rényi* model (Albert and Barabási 2002). Figure 4 shows that Target-UCB graphs consistently achieve a much lower regret than a single UCB agent. Furthermore, all structures achieve a similar average performance. Recall that there is no explicit information sharing between the Target-UCB agents. The UCB clique cumulates less regret in the beginning (< 500 episodes), but ends with a similar performance to the Target-UCB graphs. This is most likely due to optimism falling to 0 for some arms when using Target-UCB, thus reducing the regret cumulated as a result of exploration, as compared to UCB.

These results thus show the potential of a fully decentralized multi-agent system, where the simple integration of observational data in the policy greatly benefits all learning agents.

Playing with humans

Finally, we designed a real experiment involving human subjects playing a 2-actions bandit problem over 100 episodes[3]. In a first version, humans were playing alone. In a second version, four humans were play-

[3]The complete methodology is described in the supplementary material.

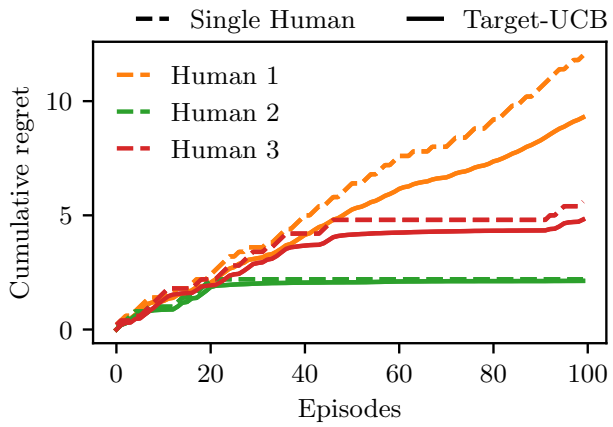

Figure 5: Target-UCB with human targets on a 2-actions setting ($\mu_\star = 0.6$, $\Delta_a = 0.2$).

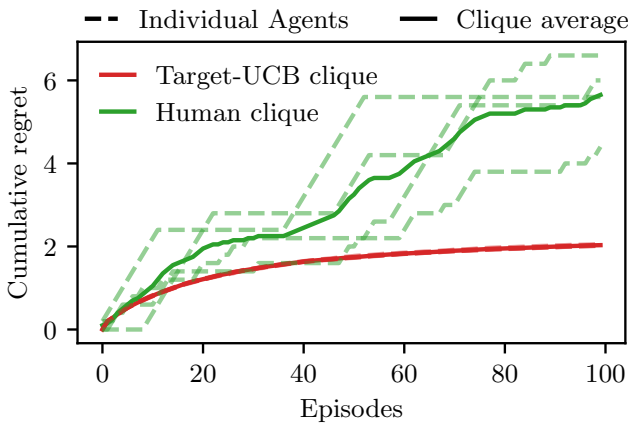

Figure 6: Cliques of humans vs Target-UCB (4 agents) on a 2-actions setting ($\mu_\star = 0.6$, $\Delta_a = 0.2$).

ing simultaneously (as a clique) and had access to each others' previous actions. The resulting dataset can be found at: https://github.com/lupuandr/Target-UCB/tree/master/Human%20bandit%20dataset.

We first evaluate the performance of Target-UCB (averaged over 200 runs) when learning from a single human target. One of the potential values of the Target-UCB algorithm is its potential to learn from human-generated data. Figure 5 shows that Target-UCB learns to become better than its target[4]. This is despite the fact that we cannot guarantee that human players satisfy Assumption 1.

We also compare the performance of a clique of four Target-UCB agents (averaged over 200 runs) against a clique of four human players. Figure 6 shows that Target-UCB agents seem more efficient than humans at leveraging observational data from their peers[5]. The Target-UCB clique rapidly converges towards the optimal action, with minimal variance in chosen actions between agents.

7 Conclusion and future work

This work studies the benefits and trade-offs of using observational data in bandit problems. We proposed a mathematical term (target optimism) that takes into account the actions selected by a target policy without observing the rewards obtained by that target. To better understand and illustrate the effect of observational learning on a bandit agent, we incorporated target optimism into UCB to obtain the Target-UCB algorithm. We provided regret upper-bounds for this algorithm that depend on the quality of the target and we considered various possible policies that could serve as valid targets to Target-UCB. Unsurprisingly, we have found that learning from a relatively good target leads to better performance (faster convergence). However, when learning from a bad target, Target-UCB can still converge and outperform its target. This is interesting especially 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.

In sum, target optimism can be a powerful tool to accelerate benefit by drawing guidance from external observational data, all while managing risks by still preserving logarithmic regret even when the target performs worse than expected.

The proposed approach could be used to study phenomena which appear in online social networks, such as the emergence of online influencers, the prediction of viral trends, or the modeling of other social behaviours relying heavily on imitation with limited communication between individuals. We also note that the proposed approach could be used in future work in order to build safe learning agents, which can temper their optimism based on information provided by the target.

An important point that has not been addressed in this paper is the explicit ability to detect when following the target is not efficient. Indeed, as observed in the numerical experiments results, learning from a bad target can lead to larger regret than using a single UCB. Being able to characterize the quality of the target could help in avoiding this situation.

[4]One pair (Target-UCB and Human) omitted for clarity; see the supplementary material for complete plot.

[5]See the supplementary material for additional results using a second human clique.

References

Albert, R., and Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of modern physics* 74(1):47.

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

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

Bandura, A., and Walters, R. H. 1963. Social learning and personality development.

Bandura, A. 1977. Social learning theory.

Banerjee, A. V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics* 107(3):797–817.

Bikhchandani, S.; Hirshleifer, D.; and Welch, I. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives* 12(3):151–170.

Borsa, D.; Piot, B.; Munos, R.; and Pietquin, O. 2017. Observational learning by reinforcement learning. *arXiv preprint arXiv:1706.06617*.

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning (ICML)*, 151–159.

Garivier, A., and Moulines, E. 2011. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory (ALT)*, 174–188. Springer.

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2016. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *IEEE Conference on Decision and Control (CDC)*, 167–172.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *International conference on World Wide Web (WWW)*, 661–670. ACM.

Miller, N. E., and Dollard, J. 1941. Social learning and imitation.

Ratliff, N.; Bagnell, J. A.; and Srinivasa, S. S. 2007. Imitation learning for locomotion and manipulation. In *IEEE-RAS International Conference on Humanoid Robots*, 392–397.

Rendell, L.; Boyd, R.; Cownden, D.; Enquist, M.; Eriksson, K.; Feldman, M. W.; Fogarty, L.; Ghirlanda, S.; Lillicrap, T.; and Laland, K. N. 2010. Why copy others? insights from the social learning strategies tournament. *Science* 328(5975):208–213.

Rosin, C. D. 2011. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence* 61(3):203–230.

Russell, S. 1998. Learning agents for uncertain environments. In *Computational Learning Theory (COLT)*, 101–103.

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

Schlag, K. H. 1998. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of economic theory* 78(1):130–156.

Schwartz, E. M.; Bradlow, E. T.; and Fader, P. S. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* 36(4):500–522.

Toyokawa, W.; Kim, H.-R.; and Kameda, T. 2014. Human collective intelligence under dual exploration-exploitation dilemmas. *PloS one* 9(4):e95789.