European Workshop on Reinforcement Learning 14 (2018) October 2018, Lille, France.

Leveraging Observational Learning for Exploration in Bandits

Audrey Durand AUDREY.DURAND@MCGILL.CA
Andrei Lupu ANDREI.LUPU@MAIL.MCGILL.CA
Doina Precup DPRECUP@CS.MCGILL.CA
School of Computer Science, McGill University, Montreal, Canada

Abstract

Imitation learning has been widely used to speed up learning in novice agents, by allowing them to leverage existing data from experts. 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. Our goal is to leverage the target data in order to guide the agent's exploration. We propose a method that builds on the Upper Confidence Bound algorithm by using conditional optimism contingent on the actions of the target. We provide a regret upper-bound of order $\mathcal{O}(\ln T)$ in two-actions settings and 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 of this type of imitation learning in the multi-armed bandit setting.

1. Introduction

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 et al., 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, 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. We would like to leverage this data to improve the behaviour of the learner, specifically, to speed up the learning process. This should not be confused with cooperative bandits (Landgren et al., 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 et al., 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.

For this purpose, we introduce an algorithm based on the vanilla Upper Confidence Bound (UCB) algorithm Auer et al. (2002), which we call Target-UCB. The core idea involves an action selection process

in uenced 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 speci cally, we show that unless the target is 100% wrong, Target-UCB will manage to cumulate logarithmic regret. The results 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 A denotes the set of possible actions. Each action $a \in A$ is associated with an unknown expected payoff $_a$. On each episode $t = 1$, the agent selects an action $a_t \in A$ and observes reward $r_t \sim (\ _{a_t})$, where $(\ )$ is a probability distribution of mean. Let $? := \operatorname{argmax}_{a \in A}\ _a$ denote the optimal action. The goal of the agent is to minimize the cumulative pseudo-regret after T episodes:

$$R(T) := \sum_{t=1}^{T} (\ ? \quad _{a_t}): \tag{1}$$

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 in the neighbourhood. The goal of the learner is still to minimize the cumulative regret in Eq. 1.

3. Algorithm

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

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

We introduce Target-UCB, a UCB-style algorithm that adjusts its optimism with respect to a speci c 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 outlines Target-UCB for reward distributions with support $[0; 1]$ (e.g., Bernoulli rewards).[1]

3.1 Intuition

We distinguish two parts in the optimism term of Target-UCB: *estimation optimism* and *target optimism*. One can see that the seminal UCB (Auer et al., 2002) algorithm is a special case of Target-UCB where $N^{*}_{a;t} = 1$ for all $a; t$, giving full focus on estimation optimism. But, rather than using optimism solely to overcome uncertainty in reward estimation (through empirical means), Target-UCB relies on optimism to compensate

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

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

Parameters: constant $C > 3/2$.
Initialization: play each action once.
for all $t > A + 1$ do
  Play action defined as:

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

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

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 μ_a. However, as $N_{a,t}$ grows closer to $N_{a,t}$, Target-UCB is allowed to be less optimistic than UCB would be (with respect to action a). Similarly, Target-UCB is quickly forced to become realistic regarding actions that are rarely played by the target ($N_{a,t}$ low). Receiving such guidance from a target policy allows Target-UCB to reduce its optimism drastically compared with UCB. This is shown by regret upper-bounds, which we present next.

3.2 Regret upper-bound

We consider a bandit with two-actions $A = \{a^\star, a\}$, and reward distributions with support in $[0, 1]$. Let $\Delta_a := (\mu^\star - \mu_a)$ denote the gap between action a and optimal action $a^\star$. Under the following assumption, Theorem 2 provides a bound on the expected cumulative pseudo-regret given the performance of the target.

Assumption 1 (Optimal plays by the target policy.) The target policy plays such that there exists some constants $\alpha \in (0, 1]$ and c_α for which, $\forall a \in A, a \neq a^\star, \forall t > c_\alpha$,

$$N_{a^\star,t} > \frac{C}{C - 3/2} \frac{6 \ln t}{\alpha \Delta_a^2} \quad \text{and} \quad N_{a^\star,t} > \frac{1}{1 - \alpha} N_{a,t}.$$

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

The following result shows that Target-UCB maintains a logarithmic regret even when taking inspiration from a poorly behaving target. It also shows that, when following a UCB target, the proposed approach cannot do worse than UCB (Auer et al., 2002) – it can only improve upon it. Moreover, since UCB has the term $8 \ln T / \Delta_a$, we would expect Target-UCB to outperform UCB when the target policy is good. These intuitions are supported by empirical results (see Sec. 5).

Theorem 2 Consider $A = \{?, a\}$ and rewards in $[0, 1]$, and assume that the target policy satisfies Assumption 1. Then, for $\in (0, 1]$, the expected cumulative regret (Eq. 1) of Target-UCB (Alg. 1) with $= 3/2$ is bounded as follows:

If $N_{a,T} < \frac{6 \ln T}{\Delta_a^2}$, it is bounded by

$$E[R(T)] \leq \Delta_a \left(c + 2^{-} \right) + \frac{6 \ln T}{\Delta_a},$$

if $N_{a,T} \leq \frac{(\sqrt{6} + \sqrt{C})^2 \ln T}{\Delta_a^2}$ or $> \frac{1}{2}$, it is bounded by

$$E[R(T)] \leq \min \begin{cases} \Delta_a E[N_{a,T}] \\ \Delta_a \left(c + 2^{-} \right) + \frac{(\sqrt{6} + \sqrt{C})^2 \ln T}{\Delta_a}, \end{cases}$$

otherwise it is bounded by

$$E[R(T)] \leq \min \begin{cases} \Delta_a E[N_{a,T}] \\ \Delta_a \left(c + 2^{-} \right) + \frac{\left(\sqrt{\frac{3}{2}} + \sqrt{C} + \sqrt{\frac{3}{2} \frac{1}{}} \right)^2 \ln T}{\Delta_a}. \end{cases}$$

3.3 Proof outline

Here follows a proof sketch of Theorem 2. The complete proof can be found in Appendix A. It essentially borrows ideas from Auer et al. (2002) and Baransi et al. (2014), where we begin by expressing the cumulative regret (Eq. 1) as $R(T) = \sum_a \Delta_a N_{a,T}$. 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 3 Let E_t^a and $E_t^?$ respectively denote the events in which

$$m_{a,s} + \sqrt{\frac{3 \ln t}{2 N_{a,s}}} \leq \quad \text{and} \quad m_{?,s} \leq \sqrt{\frac{3 \ln t}{2 N_{?,s}}} \quad \text{simultaneously for } s \leq t.$$

The idea is to decompose

$$N_{a,T} \leq \ell + \sum_{t=\ell}^{T} \mathbb{1}\{a_t = a, E_t^a, E_t^?, N_{a,t} > \ell\} + \sum_{t=A+1}^{T} \mathbb{1}\{E_t^a \cup E_t^?\}$$

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

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

Also recall that sub-optimal action a is played if

$$m_{a,t} + \sqrt{\frac{C \ln t}{N_{a,t}} \left(1 - \frac{N_{a,t}}{N_{a,t}} \wedge 1 \right)} > m_{?,t} + \sqrt{\frac{C \ln t}{N_{?,t}} \left(1 - \frac{N_{?,t}}{N_{?,t}} \wedge 1 \right)}.$$

We deduce that, under events E_t^3 and $E_t^?$, selecting action a at episode t requires that

$$a \; 6 \quad \sqrt{\frac{3 \ln t}{2N_{a;t}}} + \sqrt{\frac{C \ln t}{N_{a;t}}} \quad 1 \quad \frac{N_{a;t}}{N_{?;t}} \wedge 1 \quad + \quad \sqrt{\frac{3 \ln t}{2N_{?;t}}} \quad \sqrt{\frac{C \ln t}{N_{?;t}}} \quad 1 \quad \frac{N_{?;t}}{N_{?;t}} \wedge 1 \; : \qquad (2)$$

Now, the count of optimal plays by Target-UCB could be in one of two cases: $N_{?;t}=N_{?;t} \; 6 \; 1 \quad \frac{3}{2C}$ and $N_{?;t}=N_{?;t} > 1 \quad \frac{3}{2C}$. 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 $N_{?;t}$ that are required in order to satisfy Eq. 2 for each of these four cases. Then, we can use Assumption 1 and pick $`$ such that Eq. 2 cannot be satisfied anymore. Finally, by showing that the probability of nonoccurrence of events is controlled by the definition of events, we obtain Theorem 2.

4. 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 $?$ once in a while, at the price of a larger c . Let us look at some candidate policies that could constitute valid targets.

4.1 UCB

Since UCB begins by selecting each action at least once, the condition $N_{?;t} > N_{a;t}$ is necessarily satisfied. Also, since UCB is known to enjoy sub-linear regret, there must exists a time c such that $c \quad \sum_{a2A} N_{a;c} > \frac{C}{C \; 3=2} \frac{6 \ln t}{\frac{2}{a}}$. 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:

$$E[N_{a;t}] \; 6 \; \frac{8 \ln t}{\frac{2}{a}} + 1 + \frac{2}{3}:$$

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

$$N_{a;t} \; 6 \; \frac{8 \ln t}{\frac{2}{a}} + 1 + \frac{2}{3} + \sqrt{2t \ln(1=)}:$$

By considering that $N_{?;c} \; = \; c \quad N_{a;c}$ in the two-actions setting, we can show that there exists some time c s.t. Assumption 1 holds with high probability (e.g. see Appendix B). Section 5.1 provides results showing that UCB policy is indeed an appropriate target and also that Target-UCB outperforms this target.

4.2 -optimal

Now consider a basic family of policies that plays the optimal action with probability higher than $2 (0; 1]$. The expected number of optimal plays after t episodes using such a target policy is lower-bounded by:

$$E[N_{?;t}] > \; t:$$

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

$$N_{?;t} > \; t \quad \sqrt{2t \ln(1=)} \quad \text{for } \frac{1}{2} 6 \quad 6 \; 1 \quad \text{and} \quad N_{?;t} > \; t \quad (1 \quad) \sqrt{2t \ln(1=)} \quad \text{for } 0 < \; < \; \frac{1}{2}:$$

By considering that $N_{?;c} = c \cdot N_{a;c}$ in the two-actions setting, we can find the value of c that satisfies the first condition of Assumption 1 with high probability. In other cases, c may be higher. Section 5.2 provides results showing that the optimal policy is indeed an appropriate target. More specifically, results show a slower convergence of Target-UCB for a lower α, but it converges nonetheless.

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

4.3 Average of neighbours

Consider a Target-UCB agent who can observe the actions of several other agents, denoted as its neighbours. Target-UCB can then consider a target policy that encompasses all neighbours, with $N_{a,t}$ 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 a near-optimal policy, with α depending on the number of neighbours that select the optimal action a^*.

5. Experiments

The following experiments evaluate the potential of Target-UCBα ($\alpha = 2$) 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://goo.gl/vYD1gY .

5.1 Learning better than the target

We first consider a 2-actions problem, using UCB (Auer et al., 2002) as target, in order to assess the benefits of observing a target that is also a learning agent. We therefore have 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 greedy follower, which always selects the action chosen most often so far by the target. Fig. 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.

5.2 Learning from a non-learner

In the same 2-actions setting, we now consider an optimal policy, that plays optimal action with probability ρ, as target. Note that the optimal agent is not learning. As in the previous experiment, we evaluate Target-UCB as a learner, in comparison to the greedy follower. Fig. 2 shows the cumulative regret for the target, the greedy follower, and Target-UCB, for different values of ρ. We observe that the convergence of Target-UCB is influenced by the quality of the target policy – it converges much faster for a larger ρ. However, note that Target-UCB still converges even for a bad target (low ρ), which is not the case for the greedy follower that blindly follows the target. As long as the target is not 100% wrong ($\rho > 0$), 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.

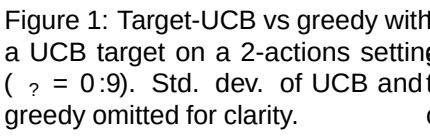

Figure 1: Target-UCB vs greedy with a UCB target on a 2-actions setting ($_?$ = 0:9). Std. dev. of UCB and greedy omitted for clarity.

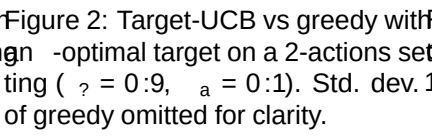

Figure 2: Target-UCB vs greedy with an -optimal target on a 2-actions setting ($_?$ = 0:9, $_a$ = 0:1). Std. dev. of greedy omitted for clarity.

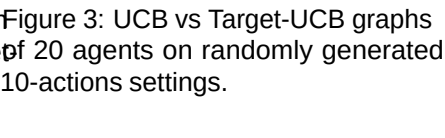

Figure 3: UCB vs Target-UCB graphs of 20 agents on randomly generated 10-actions settings.

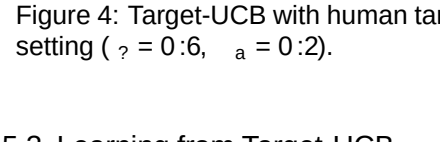

Figure 4: Target-UCB with human targets on a 2-actions setting ($_?$ = 0:6, $_a$ = 0:2).

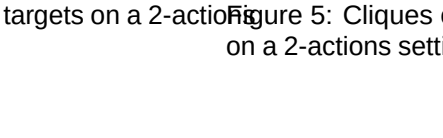

Figure 5: Cliques of humans vs Target-UCB (4 agents) on a 2-actions setting ($_?$ = 0:6, $_a$ = 0:2).

5.3 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. Note that the greedy follower baseline is not available anymore, as it requires its own target. The experiments is carried out 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 Erdos-Rényi model (Albert and Barabási, 2002). Fig. 3 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.

5.4 Playing with humans

We nally designed a real experiment involving human subjects playing a 2-actions bandit problem over 100 episodes. In a rst version, humans were playing alone. In a second version, four humans were playing simultaneously (as a clique) and had access to each others' previous actions. The resulting dataset can be found at:https://goo.gl/qoZBRo . We rst evaluate the performance of Target-UCB (averaged over

200 runs) when learning from a single human target. Fig. 4 shows that Target-UCB learns to become better than its target. 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. Fig. 5 shows that Target-UCB agents seem more efficient than humans at leveraging observational data from their peers. The Target-UCB clique rapidly converges towards the optimal action, with minimal variance in chosen actions between agents. Additional results and complete methodology are provided in Appendix C.

6. Conclusion and future work

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. 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.

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 vanilla UCB. Being able to characterize the quality of the target as a target could help in avoiding this situation.

References

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

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

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

A. Bandura. Social learning theory, 1977.

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

A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML)*, pages 115–131, 2014.

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

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

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

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

S. Russell. Learning agents for uncertain environments. In Proceedings of the eleventh annual conference on Computational learning theory, pages 101–103, 1998.

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

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

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

Appendix A. Regret upper-bound

This proof essentially follows the ideas from Auer et al. (2002) and Baransi et al. (2014). We can express the cumulative regret (Equation 1) as

$$R(T) = \sum_a \Delta_a N_{a;T}.$$

This quantity can be bounded by controlling sub-optimal plays:

$$N_{a;T} = 1 + \sum_{t=A+1}^{T} \mathbb{I}f\, a_t = a\}.$$

Also let us introduce the following events to characterize the concentration of the empirical means.

Definition 5 (Events E_t^a and $E_t^?$) Let these events respectively denote the situations where

$$m_{a;s} < \mu_a + \sqrt{\frac{3 \ln t}{2N_{a;s}}} \quad \text{and} \quad \mu_? < m_{?;s} + \sqrt{\frac{3 \ln t}{2N_{?;s}}}$$

simultaneously for all $s \leq t$.

The idea is to decompose

$$\begin{aligned} N_{a;T} &\leq 1 + \sum_{t=A+1}^{T} \mathbb{I}f\, a_t = a, E_t^a, E_t^?\} + \sum_{t=A+1}^{T} \mathbb{I}f\, E_t^a \cup E_t^?\} \\ &\leq \ell + \sum_{t=\ell}^{T} \mathbb{I}f\, a_t = a, E_t^a, E_t^?, N_{a;t} > \ell\} + \sum_{t=A+1}^{T} \mathbb{I}f\, E_t^a \cup E_t^?\}. \end{aligned} \tag{3}$$

and control the two sums separately. Then, the general idea of the proof will be to pick ℓ such that the first sum is controlled and to show that the second sum is controlled by definition of events. Let us begin by bounding the first sum, that is the sum of sub-optimal plays under the occurrence of events E_t^a and $E_t^?$. We can decompose

$$\begin{aligned} \sum_{t=\ell}^{T} \mathbb{I}f\, a_t = a, E_t^a, E_t^?, N_{a;t} > \ell\} \leq & \sum_{t=\ell}^{T} \mathbb{I}\left\{ \frac{N_{a_t}}{N_{a_t}} < 1 \right\} \mathbb{I}\{ a_t = a, E_t^a, E_t^?, N_{a;t} > \ell, \frac{N_{a_t}}{N_{a_t}} < 1 \} \\ & + \sum_{t=\ell}^{T} \mathbb{I}\left\{ \frac{N_{a_t}}{N_{a_t}} > 1 \right\} \mathbb{I}\{ a_t = a, E_t^a, E_t^?, N_{a;t} > \ell, \frac{N_{a_t}}{N_{a_t}} > 1 \} \end{aligned}$$

The first part refers to Target-UCB being better than the target policy with respect to action a, while the second part refers to Target-UCB being worse. Also recall that playing sub-optimal action a requires Equation 2 to be satisfied. This will be useful in the following developments.

A.1 Sub-optimal plays when worse

If $N_{a;t}/N^T_{a;t} > 1$ for sub-optimal action a, then we can say that Target-UCB has played worse than (or equal to) its target up to time t, with respect to action a. We want to bound the selection of action a in this situation, where Equation 2 simplifies to

$$a \leq \sqrt{\frac{3 \ln t}{2N_{a;t}}} + \sqrt{\frac{3 \ln t}{2N_{?;t}}} \sqrt{\frac{C \ln t}{N_{?;t}}} \left(1 - \frac{N_{?;t}}{N^T_{?;t}}\right)^{\wedge} 1 \quad : \tag{4}$$

We consider two situations A) $N_{?;t}/N^T_{?;t} \leq 1 - \frac{3}{2C}$ or B) $N_{?;t}/N^T_{?;t} > 1 - \frac{3}{2C}$.

A.1.1 Case A)

When $N_{?;t}/N^T_{?;t} \leq 1 - \frac{3}{2C}$, Equation 4 further simplifies to

$$a \leq \sqrt{\frac{3 \ln t}{2N_{a;t}}}:$$

Therefore, in order for

$$If\ a_t = a; E^a_t; E^?_t; N_{a;t} > `; N_{a;t}/N^T_{a;t} > 1g$$

to happen when $N_{?;t}/N^T_{?;t} \leq 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a;t} \leq \frac{3 \ln t}{2 \frac{2}{a}}: \tag{5}$$

A.1.2 Case B)

When $N_{?;t}/N^T_{?;t} > 1 - \frac{3}{2C}$, Equation 4 further simplifies to

$$a \leq \sqrt{\frac{3 \ln t}{2N_{a;t}}} + \sqrt{\frac{3 \ln t}{2N_{?;t}}}:$$

Now, either $N_{a;t} > N_{?;t}$ or $N_{a;t} \leq N_{?;t}$. The first situation requires that

$$N_{?;t} \leq \frac{6 \ln t}{\frac{2}{a}}$$

to hold. Thanks to Assumption 1, this is not possible for $N_{?;t}/N^T_{?;t} > 1 - \frac{3}{2C}$. Therefore, we must be in the situation where $N_{a;t} \leq N_{?;t}$, such that, in order for

$$If\ a_t = a; E^a_t; E^?_t; N_{a;t} > `; N_{a;t}/N^T_{a;t} > 1g$$

to happen when $N_{?;t}/N^T_{?;t} > 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a;t} \leq \frac{6 \ln t}{\frac{2}{a}}: \tag{6}$$

A.1.3 Summary

Recall that being worse than (or equal to) the target implicitly requires that $N_{a;t} > N^T_{a;t}$. Therefore, Target-UCB cannot be worse than its target if $N^T_{a;t} > \frac{6 \ln t}{\frac{2}{a}}$.

A.2 Sub-optimal plays when better

If $N_{a;t}/\tilde{N}_{a;t} < 1$ for sub-optimal action a, then we can say that Target-UCB has played better than its target up to time t, with respect to action a. We want to bound the selection of action a in this situation. Again, we consider that either A) $N_{?;t}/\tilde{N}_{?;t} \leq 1 - \frac{3}{2C}$ or B) $N_{?;t}/\tilde{N}_{?;t} > 1 - \frac{3}{2C}$.

A.2.1 Case A)

When $N_{?;t}/\tilde{N}_{?;t} \leq 1 - \frac{3}{2C}$, Equation 2 simplifies to

$$\Delta_a \leq \sqrt{\frac{3 \ln t}{2N_{a;t}}} + \sqrt{\frac{C \ln t}{N_{a;t}}}.$$

Therefore, in order for

$$If\ a_t = a; E_t^a; E_t^?; N_{a;t} > \ell; N_{a;t}/\tilde{N}_{a;t} < 1g$$

to happen when $N_{?;t}/\tilde{N}_{?;t} \leq 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a;t} \leq \frac{(\sqrt{3/2} + \sqrt{C})^2 \ln t}{\Delta_a^2}. \tag{7}$$

A.2.2 Case B)

When $N_{?;t}/\tilde{N}_{?;t} > 1 - \frac{3}{2C}$, either 1) $\frac{N_{a;t}}{\tilde{N}_{a;t}} \leq \frac{N_{?;t}}{\tilde{N}_{?;t}}$ or 2) $\frac{N_{a;t}}{\tilde{N}_{a;t}} > \frac{N_{?;t}}{\tilde{N}_{?;t}}$.

Case B.1) When $\frac{N_{a;t}}{\tilde{N}_{a;t}} \leq \frac{N_{?;t}}{\tilde{N}_{?;t}}$, the condition given by Equation 2 simplifies to

$$\Delta_a \leq \sqrt{\frac{3 \ln t}{2N_{a;t}}} + \sqrt{\frac{C \ln t}{N_{a;t}}} + \sqrt{\frac{3 \ln t}{2N_{?;t}}}.$$

If $N_{a;t} \leq N_{?;t}$, we have that in order for

$$If\ a_t = a; E_t^a; E_t^?; N_{a;t} > \ell; N_{a;t}/\tilde{N}_{a;t} < 1g$$

to happen when $N_{?;t}/\tilde{N}_{?;t} > 1 - \frac{3}{2C}$, the following condition must hold:

$$N_{a;t} \leq \frac{(\sqrt{6} + \sqrt{C})^2 \ln t}{\Delta_a^2}. \tag{8}$$

As for the situation where $N_{a;t} > N_{?;t}$, we have per Assumption 1 that

$$N_{?;t} > \frac{1}{\gamma} N_{a;t}$$

such that

$$N_{a;t} \leq \frac{(\sqrt{3/2} + \sqrt{C} + \sqrt{3/2}\sqrt{(1-\gamma)/\gamma})^2 \ln t}{\Delta_a^2}. \tag{9}$$

More precisely, this leads $N_{a;t} \leq \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ for $\gamma > \frac{1}{2}$. By definition of being better than the target, we therefore have $N_{a;t} \leq \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2} \wedge \tilde{N}_{a;t}$ for $N_{a;t} \leq \frac{(\sqrt{6}+\sqrt{C})^2 \ln t}{\Delta_a^2}$ or $\frac{1}{2} \leq \gamma \leq 1$, and $N_{a;t} \leq \frac{(\sqrt{3/2}+\sqrt{C}+\sqrt{3/2}\sqrt{(1-\gamma)/\gamma})^2 \ln t}{\Delta_a^2} \wedge \tilde{N}_{a;t}$ otherwise.

Case B.2) When $\frac{N_{a;t}}{\tilde{N}_{a;t}} > \frac{N_{?;t}}{\tilde{N}_{?;t}} > 1 \quad \frac{3}{2C}$, then the condition given by Equation 2 simplifies to

$$_a 6 \quad \sqrt{\frac{3\ln t}{2N_{a;t}}} + \quad \sqrt{\frac{3\ln t}{2N_{a;t}}} + \quad \sqrt{\frac{3\ln t}{2N_{?;t}}}:$$

If $N_{a;t} 6 N_{?;t}$, we have that in order for

$$If a_t = a; E_t^a; E_t^?; N_{a;t} > `; N_{a;t}=\tilde{N}_{a;t} < 1g$$

to happen, the following condition must hold:

$$N_{a;t} < \frac{14\ln t}{\frac{2}{a}}:$$

By definition of being better than the target, we have $N_{a;t} 6 \quad \frac{14\ln t}{\frac{2}{a}} \wedge \tilde{N}_{a;t}$. As for the situation where $N_{a;t} > N_{?;t}$, this leads to

$$\sqrt{\frac{C\ln t}{N_{a;t}}}\sqrt{1 \quad \frac{N_{a;t}}{\tilde{N}_{a;t}}} \wedge 1 \quad \sqrt{\frac{C\ln t}{N_{?;t}}}\sqrt{1 \quad \frac{N_{?;t}}{\tilde{N}_{?;t}}} \wedge 1 \quad 6 \; 0$$

such that

$$_a 6 \quad \sqrt{\frac{3\ln t}{2N_{a;t}}} + \quad \sqrt{\frac{3\ln t}{2N_{?;t}}} 6 \quad \sqrt{\frac{3\ln t}{2N_{?;t}}} + \quad \sqrt{\frac{3\ln t}{2N_{?;t}}}:$$

In order for this to occur, we need

$$N_{?;t} 6 \frac{6\ln t}{\frac{2}{a}}$$

to hold. Thanks to Assumption 1, this is not possible for $N_{a;t}=\tilde{N}_{?;t} > 1 \quad \frac{3}{2C}$.

A.2.3 SUMMARY

Recall that being better than the target implicitly requires that $N_{a;t} < \tilde{N}_{a;t}$. Therefore, for $\tilde{N}_{a;t} 6 \frac{(\sqrt{6}+\sqrt{C})^2\ln t}{\frac{2}{a}}$ or $> 1=2$, we have $N_{a;t} < \tilde{N}_{a;t} \wedge \frac{(\sqrt{6}+\sqrt{C})^2\ln t}{\frac{2}{a}}$; otherwise $N_{a;t} 6 \frac{(\sqrt{3=2}+\sqrt{C}+\sqrt{\frac{3=2}{(1 \quad)=}})^2\ln t}{\frac{2}{a}} \wedge \tilde{N}_{a;t}$.

A.3 Bounding the nonoccurrence of events

We can decompose

$$\sum_{t=A+1}^{T} \frac{1}{If E_t^a [\; E_t^? g} 6 \sum_{t=1}^{T} If E_t^a g + \sum_{t=1}^{T} If E_t^? g:$$

Then, using Chernoff-Hoeffding and a simple union bound, we have that for $a^0 2 A$,

$$P \; 9s 2 f 1; : : : ; tg : m_{a^0;t} \quad _{a^0} > \sqrt{\frac{c\ln t}{2N_{a^0;t}}} \quad 6 \; \frac{1}{t^{c \; 1}} \quad 8c > 1:$$

Therefore, we can find that for $c = 3$,

$$E \sum_{t=A+1}^{T} \frac{1}{If E_t^a [\; E_t^? g} \quad 6 \sum_{t=1}^{T} P \; E_t^a + \sum_{t=1}^{T} P \; E_t^? \quad 6 \; 2 \sum_{t=1}^{T} t^{\; 2} = \frac{2}{3}: \qquad (10)$$

A.4 Putting everything together

Taking the expectation of Equation 3 and using Equation 10, we have

$$E[N_{a;T}] \leq \ell + E\left[\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a; E_t^a; E_t^?; N_{a;t} > \ell\}\right] + E\left[\sum_{t=A+1}^{T-1} \mathbb{I}\{E_t^a [\ E_t^?\}\right]$$
$$\leq \ell + E\left[\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a; E_t^a; E_t^?; N_{a;t} > \ell\}\right] + \frac{2}{3}.$$

For $N_{a;t} \leq \frac{4 \ln t}{\Delta_a^2}$ and $\ell = \frac{4 \ln T}{\Delta_a^2} + 1$ or for $N_{a;t} > \frac{12 \ln t}{\Delta_a^2}$ and $\ell = \frac{12 \ln T}{\Delta_a^2} + 1$, we have

$$\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a; E_t^a; E_t^?; N_{a;t} > \ell\} = 0.$$

For $\frac{4 \ln t}{\Delta_a^2} \leq N_{a;t} \leq \frac{12 \ln t}{\Delta_a^2}$ and $\ell = 1$ we have

$$\sum_{t=\ell}^{T-1} \mathbb{I}\{a_t = a; E_t^a; E_t^?; N_{a;t} > \ell\} < N_{a;T}.$$

Putting all this together concludes the proof of Theorem 2.

Appendix B. UCB target policy

The expected number of sub-optimal plays of action a after t episodes using target policy UCB is upper bounded by

$$E[N_{a;t}] \leq \frac{8 \ln t}{\Delta_a^2} + 1 + \frac{2}{3}.$$

Let us introduce the following random variable

$$X_t = N_{a;t} - E[N_{a;t}].$$

By construction $|X_t - X_{t-1}| \leq 1$. Thus, by an application of Azuma-Hoeffding's inequality for martingales, we obtain that for all $\epsilon \in (0, 1)$, with probability higher than $1 - \epsilon$,

$$N_{a;t} - E[N_{a;t}] < \sqrt{2t \ln(1/\epsilon)}.$$

Therefore we have that

$$N_{a;t} \leq \frac{8 \ln t}{\Delta_a^2} + 1 + \frac{2}{3} + \sqrt{2t \ln(1/\epsilon)},$$

which allows to satisfy Assumption 1 for some c.

Example 1 Consider a two-actions setting with $\Delta_a = 0.3$. With probability higher than 0.9, using $C = 3$, we have $N_{?;t} > \frac{12 \ln t}{\Delta_a^2}$ for

$$t > \frac{20 \ln t}{\Delta_a^2} + 1 + \frac{2}{3} + \sqrt{2t \ln(10)}.$$

For example, this happens for $t > 1800$, which also leads to $N_{?;t} > 0.5 N_{a;t}$. Note that this estimate is higly conservative and that experiments suggest that the assumption is respected when following a UCB target with $C = 2$.

Appendix C. Playing with humans

C.1 Methodology of experiments with human subjects

In this section we describe the general process of gathering the human player data on the 2-actions setting considered in Section 5.4. Two versions were considered: 1) humans playing along and 2) a clique of four humans having access to each others' actions. Note that the human players used for the individual runs depicted in Figure A7 are the same as those used for the second human clique in Figure A8.

Bandit setting Bernoulli reward distributions with $\mu_? = 0.6$ and $\mu_a = 0.4$ were used in all experiments. These distributions were identical and static for all players, but each reward was randomly and independently sampled for each action selection. This means that two players selecting the same action on the same episode could potentially obtain different results (i.e. one getting a win whereas the other gets a loss).

For each episode, each player clicked on their desired action (A or B), after which the buttons became deactivated and greyed-out. After a few seconds, the buttons were reactivated, and the reward for the selected action was displayed: either "Win!" in green or "Loss" in red. In the clique experiment, once all players had selected their action for an episode, a new row in the spreadsheet was displayed, showing the action played by each user. This marked the start of the next episode. Both the single player and the clique experiments were conducted over 100 episodes.

Game interface A player interface was created using *Google Apps Script*[2] and the *Google Sheets API*[3] in order to simplify the process of interacting with the bandit setting and accessing the actions played by other players. Figure A6 shows a screen capture of the player interface used for the human clique experiment. The interface for the single player experiment is almost identical, except there is a unique "Player" column (each player plays in its own sheet).

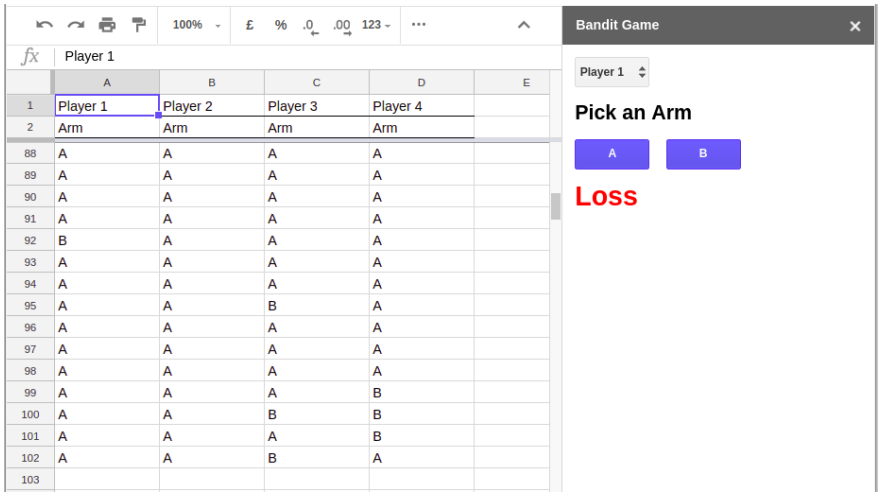

Figure A6: Screen capture of the player interface used for gathering the human clique data presented in Section 5.4.

Single player experiment The players were gathered in a room and each was playing on a separate computer, on their dedicated interface (sheet), with no communication between them.

2. https://developers.google.com/apps-script/
3. https://developers.google.com/sheets/api/

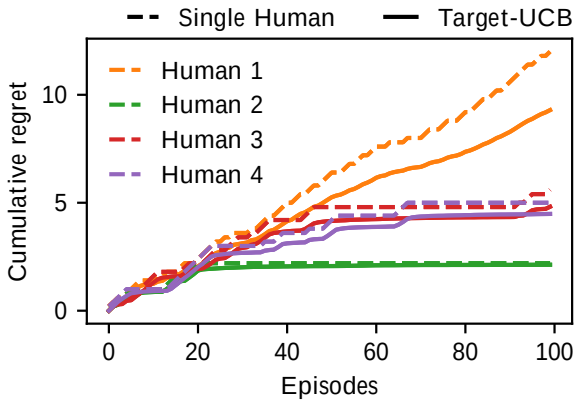

Figure A7: Target-UCB with human targets on a two-actions setting ($\mu_? = 0.6$, $\Delta_a = 0.2$). The fourth human target and Target-UCB pair (purple) shown here was omitted in Figure 4.

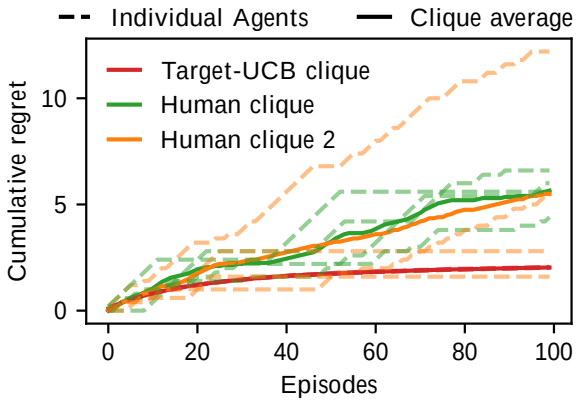

Figure A8: Cliques of humans vs Target-UCB (4 agents) on a two-actions setting ($\mu_? = 0.6$, $\Delta_a = 0.2$). The second human clique shown here was omitted in Figure 5.

Clique player experiment The players were gathered in a room and each was playing on a separate computer, with no communication between them. Player IDs (i.e 1 to 4) were assigned randomly and silently, such that players did not know which person was associated to which player ID.

C.2 More results

Figure A7 is simply a second version of Figure 4, but with no pair (Target-UCB and human target) omitted. Figure A8 shows the performance of a clique of 4 Target-UCB agents against two cliques of 4 human agents, with the first (green) one also shown in Figure 5. The second human clique shown here has much higher variance between players, but both human cliques seem to accumulate regret at a very similar rate, on average.