Leveraging observations in bandits: Between risks and benefits

Andrei Lupu, Audrey Durand, Doina Precup

McGill University
Montreal, Canada

AAAI-19

Outline

Observational Bandits – Problem Definition

- Bandit problem with $\mathcal{A}$ being the set of possible actions.
- Each action $a \in \mathcal{A}$ is associated with an unknown expected payoff μ_a.
- $\star := \operatorname{argmax}_{a \in \mathcal{A}} \mu_a$ is the *optimal action*.
- Goal of agent is to minimize the cumulative (pseudo-)regret after T rounds:

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

- Agent has access to the actions performed by an unknown *target* policy, but does not observe the associated rewards.

Motivation

- **Expand bandit theory** to the multi-agent setting;
- Observational learning occurs naturally in humans;
- Can be used for designing **marketing strategies**.

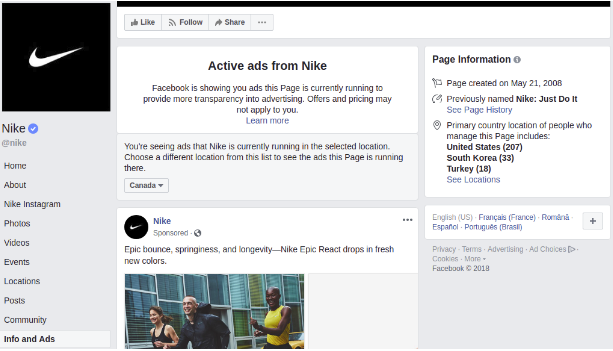


Figure 1: Example of a Facebook Ad information page

Target-UCB Algorithm

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:

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

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

end for

Assumption (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, $\forall a \in \mathcal{A}, a \neq \star, \forall t \geqslant c_\Delta$,

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

Regret Bound

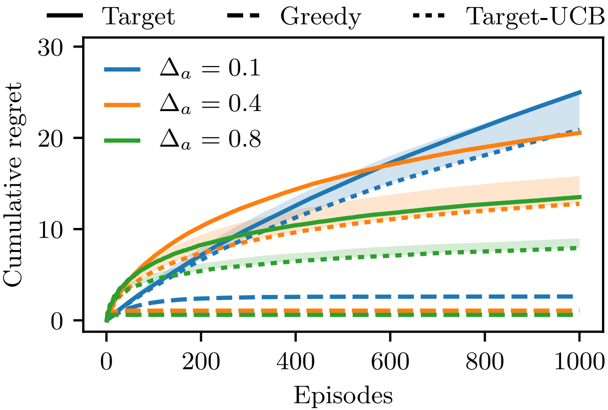
Theorem Summary

Given the assumption holds, Target-UCB :

- Achieves logarithmic regret;
- Outperforms all (known) targets;
- Outperforms UCB given a good enough target.

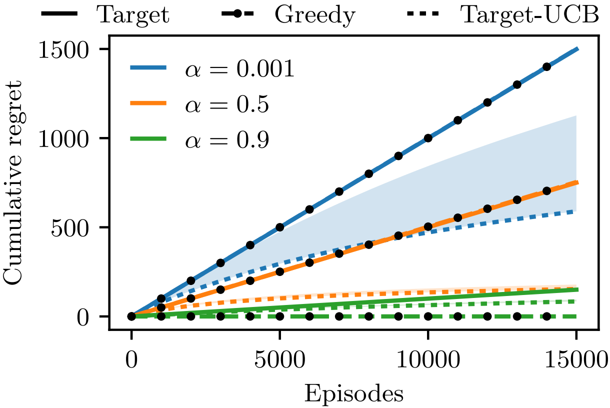
Learning from a learning target

- **UCB**: $a_t = \text{argmax}_{a \in \mathcal{A}} \, m_{a,t} + \sqrt{\frac{2 \ln t}{N_{a,t}}}$
- $\mu_\star = 0.9$, $\mu_a \in \{0.1, 0.5, 0.8\}$



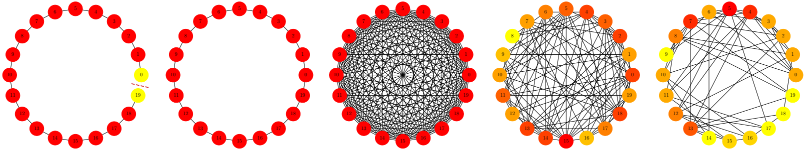
Learning from a non-learner

- α-optimal: $a_t = \star$ with probability α
- $\mu_\star = 0.9$, $\mu_a = 0.8$

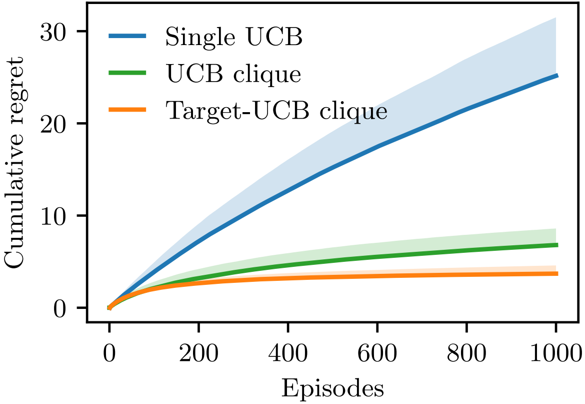


Agent Graphs

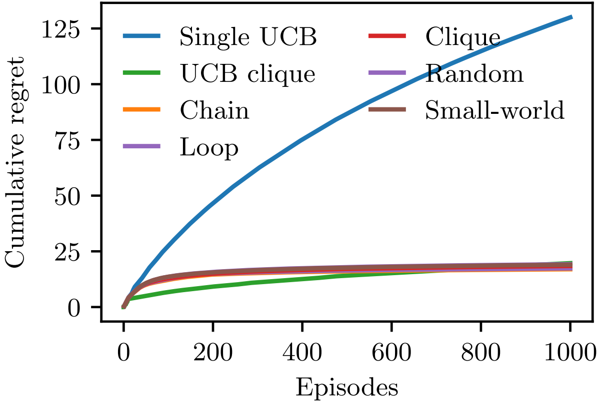
- Chain
- Loop
- Clique: All agents connected
- Random: *Erdős-Rényi* model
- Small-world: *Barabási-Albert* model



The Power of Neighbours

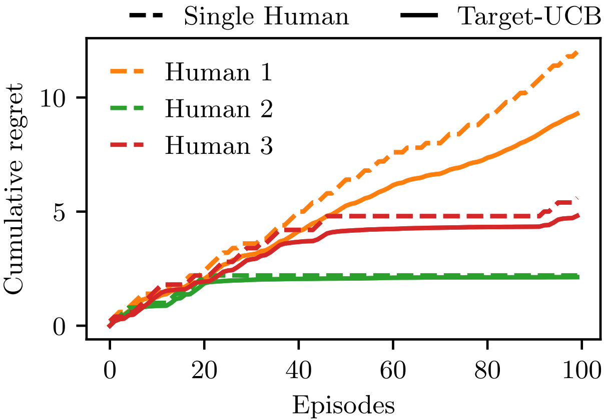


Single UCB and UCB clique of 11 agents vs Target-UCB clique of 11 agents on a 2-actions setting ($\mu_\star = 0.5, \Delta_a = 0.1$).



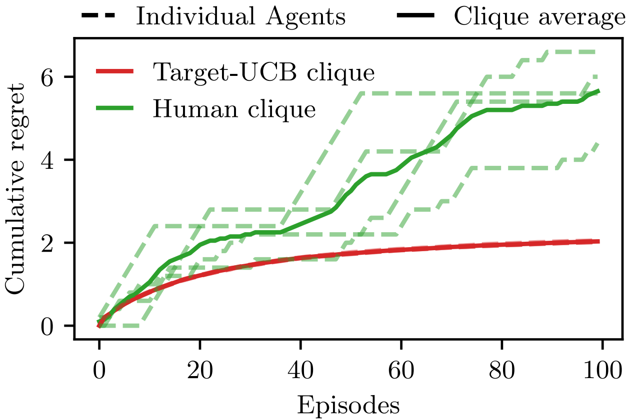
Single UCB and UCB clique of 20 agents vs five Target-UCB graphs of 20 agents on randomly generated 10-actions settings.

Human Target



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

Human Clique Comparison



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

Wrap-up

- Learning from a **good target** leads to **better performance** (faster convergence).
- When learning from a **bad target**, Target-UCB can still converge and **outperform its target**, including humans.
- Future work:
  - Determine efficiency of following target
  - Extend to full information setting

Thank you!

Selected References

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

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

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.