The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Imitation Upper Confidence Bound for Bandits on a Graph

Andrei Lupu, Doina Precup
Reasoning and Learning Lab
McGill University, 3480 University St.
Montreal, Quebec H3A 0E9
Phone: (514) 398-6443
E-mail: andrei.lupu@mail.mcgill.ca, dprecup@cs.mcgill.ca

Introduction

Bandit problems are of high theoretical interest, as they represent a simple setup for the study of the explore-exploit trade-off faced by an agent in an unknown environment. Despite the large corpus on the topic, most prior work has focused on a single-agent setup (Bubeck and Cesa-Bianchi 2012).

Here, we consider the scenario of a graph of multiple interconnected agents implementing a common policy and each playing a bandit problem with identical reward distributions. We impose a restriction on the information propagated in the graph such that "neighbouring" agents can only observe each other's actions, but not the corresponding payoffs.

Not only does the resulting problem expand bandit theory, but it can also be of particular applicability to online influencers, the prediction of viral trends, or the modeling of other social behaviours relying heavily on imitation with limited communication between individuals.

Our approach extends the vanilla Upper Confidence Bound (UCB) algorithm (Auer, Cesa-Bianchi, and Fischer 2002) to the imitating multi-agent bandit problem described above. The resulting Imitation Upper Confidence Bound (IUCB) algorithm involves an action selection process in two parts according to which an agent either acts individually using UCB or imitates the most popular action among all its neighbours.

IUCB Algorithm

Let k be an agent in the graph and let B^k be the set of all agents connected to k with a single edge (i.e. the neighbours of k). For each of its own plays of arm i, the agent updates a running average $X_i^k(t)$ of rewards and a play count $n_i^k(t)$ for that arm. At each timestep t, the agent also keeps track of the number $P_i^k(t)$ of times that the arm was selected by one of its neighbours in the last 10 time steps. Formally, this measure of popularity is defined to be

$$P_i^k(t) = \sum_{j \in B^k} \sum_{s=t-10}^{t-1} \{A^j(s) = i\}, \tag{1}$$

where $A^j(t)$ is the arm selected by agent j at time t and $\{I(t)\}$ is simply the indicator function of event $I(t)$. Also, let

$$C_i^k(t) := X_i^k(t) + c\sqrt{\frac{\ln t}{n_i^k(t)}} \tag{2}$$

be the original UCB confidence bound, with $c > 0$ a tunable parameter.

The action selection proceeds as such: every agent initializes the IUCB algorithm by selecting each bandit arm exactly once. Then, for all the following timesteps, a random variable $v^k(t) \in [0, 1]$ is sampled from a uniform distribution for each agent. The action is subsequently selected according to:

$$A^k(t) = \begin{cases} \arg\max_i \left[C_i^k(t)\right], & v^k(t) \geq \alpha \\ \arg\max_i \left[P_i^k(t)\right], & v^k(t) < \alpha, \end{cases} \tag{3}$$

where $\alpha \in [0, 1]$ is the imitation frequency.

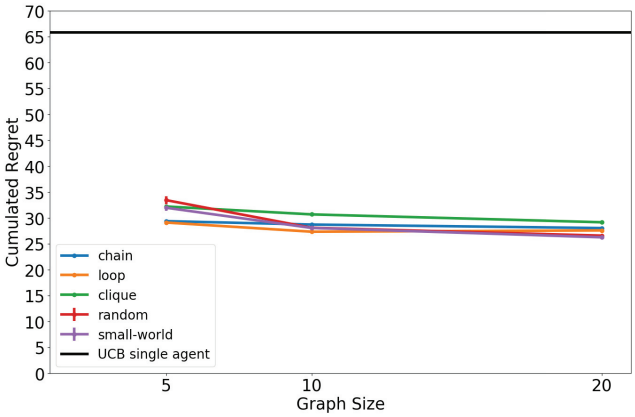
Results

We implemented the IUCB algorithm on graphs with different structures and sizes, and empirically demonstrate the improved performance over vanilla UCB on a large set of 10-armed bandits with Gaussian reward distributions. Full details of the experimental methodology can be found in the supplemental material.

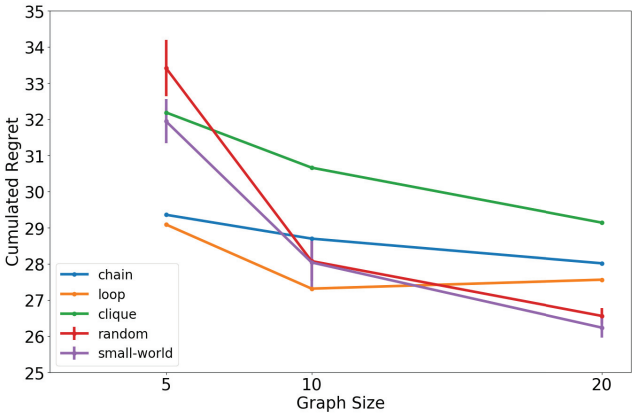
Fig. 1 shows the average cumulated regret $\bar{R}$ for all structures tested as a function of graph size. As can be seen in panel (a), the IUCB algorithm cumulates only approximately half the regret of vanilla UCB, thus achieving a significantly better performance. This is a direct effect of the actions done through imitation, as will become apparent from Fig. 2

From the detailed view in panel (b), we observe that regret draws a non-linear benefit from an increase in graph size. This benefit is especially seen in the case of the Erdős-Rényi random and Barabási-Albert small-world graphs (Albert and Barabási 2002).

In the case of the BA small-world graphs, the sharp decrease in regret is particularly noteworthy, since they were all generated with the same new node degree parameter ($m = 3$). Consequently, the larger graphs are effectively *sparser* than smaller ones.



(a)



(b)

Figure 1: Average cumulated regret per agent over 1000 time steps. (a) Comparison of various graphs to vanilla UCB (black line). (b) Detail showing average cumulated regret per agent as a function of graph size for different structures.

This apparent tendency for increased performance of the IUCB algorithm in sparser graph structures is accentuated by the fact that the clique (i.e. fully-connected graph) is on average the worse performing structure among all those tested. Our intuition for this observation is that, within a clique, all agents share an identical neighbourhood. Thus, when imitating, each agent observes very similar values of $P_i^k(t)$, which leads to a less efficient exploration during the earlier time steps.

Fig. 2 shows how the *average imitation benefit* $\bar{b}(t)$ evolves in time. This metric is formally defined in the supplemental material, but it essentially measures the true gap between the action $A^k(t)$ selected by an agent *when imitating* and the action maximizing the UCB bound $C_i^k(t)$.

Thus, the curves demonstrate that imitation allows agents to substitute a fraction of early actions that would have been used through UCB with more optimal ones. We can interpret imitation as a form of informed exploration, partially compensating for inaccurate early estimates resulting from the low number of samples taken by individual agents.

Joint analysis of Fig. 1b and Fig. 2 seems to indicate that

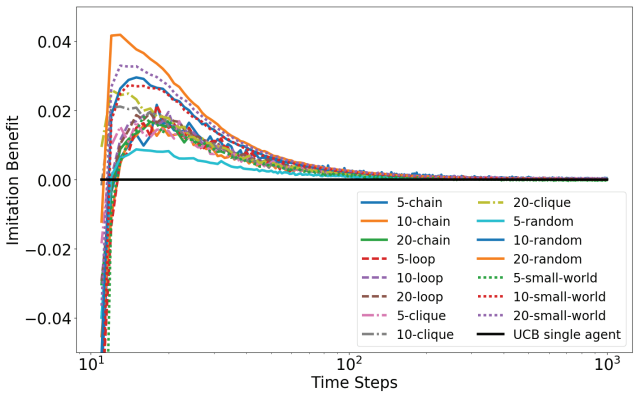


Figure 2: Average imitation benefit per agent for different graph sizes and structures. Time scale is logarithmic.

the best performing graphs are also the ones with the highest imitation benefit, which is to be expected.

Furthermore, $\bar{b}(t)$ is also an indirect metric of convergence between agents of a given graph. Indeed, since the imitation frequency α is constant, a decrease in imitation benefit is only attributable to a decrease in the average gap between the perceived optimal action and the action taken.

Looking at Fig. 2, we remark that all graphs seem to converge at approximately the same rate, despite differences in benefit during early time steps. We can therefore conclude that the IUCB algorithm asymptotically falls back upon vanilla UCB as imitation and individual actions become nearly equivalent.

Conclusion

We developed a new algorithm, the Imitation Upper Confidence Bound (IUCB), which empirically achieves better regret than vanilla UCB in the setting of imitating multi-agent bandit problems. By measuring the imitation benefit, we also provided insight on the effect of imitation on individual decision-making and on group consensus.

However, theoretical analysis of IUCB and the derivation of a regret bound remains an open problem and is thus the target of future work.

Acknowledgements

Particular thanks to the members of the RL Lab for being eager to answer my many questions, and especially to Jad Kabbara, whose support and suggestions helped in writing this paper. Finally, we thank IVADO for its funding.

References

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

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

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.