Collaboratively Learning the Best Option on Graphs, Using Bounded Local Memory

Lili Su
Massachusetts Institute of Technology, EECS
lilisu@mit.edu

Martin Zubeldia
Massachusetts Institute of Technology, EECS

Nancy Lynch
Massachusetts Institute of Technology, EECS

ABSTRACT

We consider multi-armed bandit problems in social groups wherein each individual has bounded memory and shares the common goal of learning the best arm/option. We say an individual learns the best option if eventually (as $t \to \infty$) it pulls only the arm with the highest expected reward. While this goal is provably impossible for an isolated individual due to bounded memory, we show that, in social groups, this goal can be achieved easily with the aid of social persuasion (i.e., communication) as long as the communication networks/graphs satisfy some mild conditions. In this work, we model and analyze a type of learning dynamics which are well-observed in social groups. Specifically, under the learning dynamics of interest, an individual sequentially decides on which arm to pull next based on not only its private reward feedback but also the suggestion provided by a randomly chosen neighbor. To deal with the interplay between the randomness in the rewards and in the social interaction, we employ the *mean-field approximation* method. Considering the possibility that the individuals in the networks may not be exchangeable when the communication networks are not cliques, we go beyond the classic mean-field techniques and apply a refined version of mean-field approximation. Notably, our results hold even if the communication graphs are highly sparse.

KEYWORDS

Distributed systems; bio-inspired distributed algorithms; mean-field approximation

ACM Reference Format:
Lili Su, Martin Zubeldia, and Nancy Lynch. 2019. Collaboratively Learning the Best Option on Graphs, Using Bounded Local Memory. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3309697.3331498

1 INTRODUCTION

Individuals often need to make a sequence of decisions among a fixed finite set of options (alternatives), whose rewards/payoffs can be regarded as stochastic, for example:

- Human society: In many economic situations, individuals need to make a sequence of decisions among multiple options, such as when purchasing perishable products [3] and when designing financial portfolios [14]. In the former case, the options can be the product of the same kind from different sellers. In the latter, the options are different possible portfolios.
- Social insect colonies: Foraging and house-hunting are two fundamental problems in social insect colonies, and both of them have inspired counterpart algorithms in robotics [10]. During foraging, each ant/bee repeatedly refines its foraging areas to improve harvesting efficiency. House-hunting refers to the collective decision process in which the entire social group collectively identifies a high-quality site to immigrate to. For the success of house-hunting, individuals repeatedly scout and evaluate multiple candidate sites, and exchange information with each other to reach a collective decision.

Many of these sequential decision problems can be cast as *multi-armed bandit problems* [1, 4, 7]. These have been studied intensively in the centralized setting, where there is only one player in the system, under different notions of performance metrics such as pseudo-regret, expected regret, simple regret, etc. [1, 4, 4, 7, 8, 12]. Specifically, a K-armed bandit problem is defined by the reward processes of individual arms/options $\left(R_{k,i} : i \in \mathbb{Z}_+\right)$ for $k = 1, \cdots, K$, where $R_{k,i}$ is the reward of the i–th pull of arm k. At each stage, a player chooses one arm to pull and obtains some observable payoff/reward generated by the chosen arm. In the most basic formulation the reward process $\left(R_{k,i} : i \in \mathbb{Z}_+\right)$ of each option is stochastic and successive pulls of arm k yield $i.i.d.$ rewards $R_{k,1}, R_{k,2}, \cdots$. Both asymptotically optimal algorithms and efficient finite-time order optimal algorithms have been proposed [1, 4, 12, 13]. These algorithms typically have some non-trivial requirements on individuals' memorization capabilities. For example, upper confidence bound (UCB) algorithm requires an individual to memorize the cumulative rewards of each arm he has obtained so far, the number of pulls of each arm, and the total number of pulls [1, 12]. Although this is not a memory-demanding requirement, nevertheless, this requirement cannot be perfectly fulfilled even by humans, let alone by social insects, due to bounded rationality of humans, and limited memory and inaccurate computation of social insects. In human society, when a customer is making a purchase decision of perishable products, he may recall only the brand of product that he is satisfied with in his most recent purchase. Similarly, in ant colonies, during house-hunting, an ant can memorize only a few recently visited sites.

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6678-6/19/06.
https://doi.org/10.1145/3309697.3331498

In this paper, we capture the above memory constraints by assuming an individual has only bounded/finite memory. The problem of multi-armed bandits with *finite memory constraint* has been proposed by Robbins [12] and attracted some research attention [5, 6, 15]. The subtleties and pitfalls in making a good definition of memory were not identified until Cover's work [5, 6]. We use the memory assumptions specified in [5], which require that an individual's history be summarized by a finite-valued memory. We say an individual learns the best option if eventually (as $t \to \infty$) it pulls only the arm with the highest expected reward.

For an isolated individual, learning the best option is provably impossible [5].[1] Nevertheless, successful learning is still often observed in social groups such as human society [3], social insect colonies [9] and swarm robotics [10]. This may be because in social groups individuals inevitably interact with others. In particular, in social groups individuals are able to, and tend to, take advantage of others' experience through observing their neighbors [2, 11]. Intuitively, it appears that as a result of this social interaction, the memory of each individual is "amplified", and this *amplified shared memory* is sufficient for the entire social group to collaboratively learn the best option.

Approach and key contributions: In this paper, we rigorously show that the above intuition is correct with a focus on the impact of the graph structures on the performance of collaboratively learning. Concretely, we assume time is continuous and each individual has an independent Poisson clock with common rate. When an individual's local clock ticks, it attempts to perform an update immediately via two steps:

(1) **Sampling**: **If** the individual does not have any preference over the K arms yet, **then**:
   (a) With probability $\mu \in [0, 1]$, the individual pulls one of the K arms uniformly at random (uniform sampling).
   (b) With probability $1-\mu$, the individual chooses one neighbor uniformly at random, and pulls the arm suggested by the chosen neighbor (peer recommendation); pulls no arm if the chosen neighbor does not have any preference over the K arms yet.

   else The individual chooses one neighbor uniformly at random, and pulls the arm suggested by the chosen neighbor (peer recommendation); pulls no arm if the chosen neighbor does not have any preference over the K arms yet.

(2) **Adopting**: **If** a reward is obtained by pulling the chosen arm, **then** the individual updates its preference to this arm.

Note that if the awake individual pulls no arm, it will not get a reward; thus, its preference is unchanged.

A key analytical challenge of our learning dynamics is to deal with the interplay of the randomness in the rewards and that in the social interaction. Comparing to the case when the communication graphs are cliques, this interplay is significantly complicated by the lack of exchangeability among the individuals on general communication graphs. Observing this, we go beyond the classic mean-field techniques and apply a refined version of mean-field approximation:

- Using coupling we show that, if the communication graph is connected and is either regular or has doubly-stochastic degree-weighted adjacency matrix, with probability $\to 1$ as the social group size $N \to \infty$, every individual in the social group learns the best option.
- If the minimum degree of the graph diverges as $N \to \infty$, over an arbitrary but given finite time horizon, the sample paths describing the opinion evolutions of the individuals are asymptotically independent. In addition, the proportions of the population with different opinions converge to the unique solution of a system of ODEs. Interestingly, the obtained system of ODEs are invariant to the structures of the communication graphs. In the solution of the obtained ODEs, the proportion of the population holding the correct opinion converges to 1 exponentially fast in time.

Notably, our results hold even if the communication graphs are highly sparse.

[1] A less restricted memory constraint – stochastic fading memory – is considered in [16], wherein similar negative results when memory decays fast are obtained.

ACKNOWLEDGMENTS

L. Su was supported in part by the NSF Grant CCF-1461559 and CCF-0939370.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.
[2] Albert Bandura. 1969. Social-learning theory of identificatory processes. *Handbook of socialization theory and research* 213 (1969), 262.
[3] Dirk Bergemann and Juuso Vlimki. 1996. Learning and Strategic Pricing. *Econometrica* 64, 5 (1996), 1125–1149. http://www.jstor.org/stable/2171959
[4] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.
[5] T. Cover and M. Hellman. 1970. The two-armed-bandit problem with time-invariant finite memory. *IEEE Transactions on Information Theory* 16, 2 (March 1970), 185–195. https://doi.org/10.1109/TIT.1970.1054427
[6] Thomas M Cover. 1968. A note on the two-armed bandit problem with finite memory. *Information and Control* 12, 5 (1968), 371–377.
[7] Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.
[8] Shie Mannor and John N Tsitsiklis. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5, Jun (2004), 623–648.
[9] Kazuaki Nakayama, Masato Hisakado, and Shintaro Mori. 2017. Nash Equilibrium of Social-Learning Agents in a Restless Multiarmed Bandit Game. *Scientific Reports* 7 (2017).
[10] Giovanni Pini, Arne Brutschy, Gianpiero Francesca, Marco Dorigo, and Mauro Birattari. 2012. Multi-armed Bandit Formulation of the Task Partitioning Problem in Swarm Robotics.. In *ANTS*. Springer, 109–120.
[11] Luke Rendell, Robert Boyd, Daniel Cownden, Marquist Enquist, Kimmo Eriksson, Marc W Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N Laland. 2010. Why copy others? Insights from the social learning strategies tournament. *Science* 328, 5975 (2010), 208–213.
[12] Herbert Robbins. 1956. A sequential decision problem with a finite memory. *Proceedings of the National Academy of Sciences* 42, 12 (1956), 920–923.
[13] Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh. 2017. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing* 65, 16 (2017), 4281–4292.
[14] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. 2015. Portfolio Choices with Orthogonal Bandit Learning.. In *IJCAI*. 974.
[15] Carter Vincent Smith and Ronald Pyke. 1965. The Robbins-Isbell two-armed-bandit problem with finite memory. *The Annals of Mathematical Statistics* (1965), 1375–1386.
[16] Kuang Xu and Se-Young Yun. 2018. Reinforcement with Fading Memories. In *Sigmetrics*.