Multi-agent Dynamics in Multi-armed Bandit Problem with Heterogeneous Stochastic Interactions

Udari Madhushani

Analysis of stochastic processes on graphs has received significant attention from a wide range of research fields including controls, theoretical ecology, mathematical physics and discrete probability theory. Different frameworks such as game theoretic models, simple and complex contagion processes, coloring models and particle interaction models, have been developed in this context to model natural and artificial phenomenas of interest. These models can be effectively utilized to study key features of collective dynamics in multi agent networks. Competitive and cooperative populations exhibit a rich set of behaviors consisting dominance, consensus and optimum trade-offs.

In decision theory, Multi-armed Bandit problems serve as a model that captures the salient features of the trade-off between exploring and exploiting. [1]. In this problem, an agent is repeatedly faced with the task of choosing an option from a set of options. After every execution of action, the agent receives a numerical reward drawn from a specific unknown probability distribution. The goal of this exercise is to maximize the cumulative reward in the long run. This is equivalent to minimizing the cumulative regret. If all the reward probability distributions are known, the agent achieves maximum expected cumulative reward by consistently sampling from the option with maximum expected reward. In a realistic scenario where reward distributions are unknown, the agent is required to perform sufficient exploring to estimate the expected reward values of options and identify the best option while exploiting the options with high estimated expected reward values to maximize instantaneous rewards.

Existing literature has extended this problem to multi-agent setting with deterministic interactions [2], [3]. Each agent observes either estimates or instantaneous rewards and actions of his neighbors according to a directed or undirected static network graph. As a part of my current research, I have extended this problem to a multi-agent setting with heterogeneous stochastic interactions. A set of agents are simultaneously choosing options and trying to maximize individual cumulative rewards. I considered the case where each agent can observe actions and rewards of his neighbours (1-hop neighbors) through stochastic interactions. At any given time step the agent k observes each of his neighbors with probability p_k. Since agent interactions are probabilistic and probability values are agent based, observations are made according to a dynamic directed network graph. In this work, I interpret the observation probability values of agents as their *sociability*. According to the interpretation, high (low) sociability values imply that agents pay more (less) attention to their neighbors.

In a heterogeneous sociability distribution with a homogeneous degree distribution, it is apparent that high sociability values correspond to obtaining more observations, hence better performance. With this intuition, one would expect that relative performance ranks will be ordered according to individual sociability values. This is in fact true in a well mixed multi-agent system where all agents observe each others instantaneous actions and rewards through stochastic interactions. For this case I proposed a relative performance measure, based on individual sociability values, to predict ranks of agents according to their performance. The proposed measure agrees with the analytical and computational expected cumulative regret bounds.

However, appealing contradictions to this supposition occur, in cyclic network graphs, when agents with equal or fairly different sociability values, have neighbors who fall into two extremes of sociability spectrum. Having neighbors with less sociability can be equivalently interpreted as having neighbors who explore more. This allows an agent to exploit more to increase instantaneous rewards while gathering sufficient information to identify the best option through observations. As a result, an agent with a reasonable sociability value and neighbors who are less sociable,

Department of Mechanical and Aerospace Engineering, Princeton University, NJ 08544, USA. {udarim}@princeton.edu

can outperform an agent who has a higher sociability value and neighbors who are more sociable. Based on this, I proposed a relative performance measure, that agrees with analytical and computational expected cumulative regret bounds, as a function of individual sociability values and sociability values of neighbors.

A cascade phenomena emerges in more complex settings when performance of an agent is significantly affected by sociability values of n-hop neighbors with $n \geq 2$. Better performance can be obtained by having n-hop neighbors who are less sociable with odd n values and n-hop neighbors who are more sociable with even n values. This opens up the possibility of a natural extension to incorporate adaptive sociability dynamics with positive and negative feedback. Often in realistic scenarios observing neighbors has a cost associated with it. This motivates agents to increase or decrease their sociability values according to cost and sociability values of neighbors. For instance, in a high observation cost setting, if an agent has a few 1-hop neighbors with high sociability values and more 2-hop neighbors with low sociability values, it is more beneficial to have a low sociability value. I plan to explore this direction by defining dynamics with positive 1-hop neighbor feedback and negative 2-hop neighbor feedback.

This problem can be reformulated by adapting a game theoretic model [4] in a case where agents chose sociability values from a discrete set. Simplest case is agents choose between two probability values, defined as high (H) sociability and low (L) sociability. Maximum benefit $b_{H\to L}$ can be obtained by being more sociable and observing a less sociable neighbor. In a cost free setting, following the similar convention, relative benefits can be given as $b_{H\to L} > b_{H\to H} > b_{L\to H} > b_{L\to L}$. Introducing a observation cost can change the ordering of benefits. (i. e. It is possible to obtain more benefit by being less sociable when observing a more sociable neighbor. $b_{L\to H} > b_{H\to H}$). In contrast to deterministic payoff matrices in conventional game theory, I intend to utilize a probabilistic payoff matrix. This is because, at any given time step, having a low sociability neighbor does not guarantee that the agent observes a reward from a less rewarding option, but it increases the chance of observing a sample from a suboptimal option.

Up to present I have considered the degree distribution to be homogeneous. I plan to further this work by analyzing the effect of graph topology by incorporating random graphs to capture heterogeneous degree distributions. Simplest case in this setting is, analyzing relative performance ranks for a set of homogeneous agents. (All agents have equal sociability values.) It is natural to assume that relative performance ranks of agents will correspond to the relative degree distribution of agents. However, in more complex settings two agents with same number of degrees can have distinct performance ranks due to their centrality. I plan to extend this to analyze relative performance ranks in a network of heterogeneous agents with heterogeneous degree distributions.

Another direction I am interested in pursuing is, analyzing performance of agents in dynamic environments. In foraging, animal groups exploit feeding grounds known for better resources. Due to continuous consumption eventually resources grow thinner and animals start exploring for better feeding grounds. After a certain period of time, resources grow back and abandoned feeding grounds become rewarding. This can be captured by a dynamic MAB model where expected reward value of an option decreases when it is chosen and increases when it is not chosen. In a multi-agent setting it is less complicated to assume that increasing and decreasing rates are known to the agents. Since interactions among agents are stochastic, agents do not know the total number of times each option has been sampled. In a setting where agents observe instantaneous actions and rewards of neighbors, they are required to utilize observed reward values to estimate the total number of times each option has been sampled. This allows agents to predict the expected reward values of options based on current estimates. I plan to explore effectiveness of this approach analytically and computationally.

References

[1] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards,” *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, November 1987.

[3] P. Landgren, V. Srivastava, and N. E. Leonard, “Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms,” in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 167–172.

[4] A. Traulsen, N. Shoresh, and M. A. Nowak, “Analytical results for individual and group selection of any intensity,” *Bulletin of Mathematical Biology*, vol. 70, no. 5, p. 1410, Apr 2008. [Online]. Available: https://doi.org/10.1007/s11538-008-9305-6