– Master Project Proposal–

STRUCTURED SET OF BANDITS

O-A. MAILLARD
SequeL, Inria Lille – Nord de France
E-MAIL: odalric.maillard@inria.fr

Keywords: Multi-armed bandits, Structure, Lower bounds

Introduction Reinforcement Learning is a field of research that models the problem of a learning agent interacting with a partially known dynamical system, that plays actions based on past observations and accumulates rewards that the agent tries to maximize. The Stochastic Multi-armed bandits problem is a building block model that is used in order to formalize the fundamental trade-off that arises in sequential decision making with partial observation called the *exploration-exploitation* trade-off.

While the standard stochastic multi-armed bandit setting is fairly well understood from a theoretically perspective, in many situations (typically, recommender systems) the decision making problem can be modeled as a learner facing a *set* of multi-armed bandits problems instead of a single one. Further, this set of bandit problems is structured and it makes sense to try to take advantage of this structure. In this context, lower bounds on the achievable performance of a learner must be revisited, and suggest novel algorithms should be investigated.

Answering these questions may trigger significant progress in the theoretical and practical understanding of reinforcement learning, and lead to a major progress in the field much beyond multi-armed bandits.

Goal A little more precisely, in this project, we want to better understand the following situation. We consider a *structured* bandit problem specified by a set of real-valued distributions $\nu = (\nu_{a,b})_{a\in A, b\in B}$, where A is a finite set of arms and B is a finite set. Such a ν is called a (bandit) configuration and each $\nu_b = (\nu_{a,b})_{a\in A}$ a bandit problem. At time t, an index $b_t \in B$ is given. Then the learner must choose $a_t \in A$ and receives a reward from ν_{a_t,b_t}. We assume that ν is unknown but belongs to a known set of configurations C. This set modifies the achievable performance for this problem, and it is natural to ask how to best make use of this knowledge in order to provide improved strategies.

Main tasks The goal of this Master project is first to study the lower bounds for this problem, that suggests a natural `KL-UCB` style algorithm for this setting. As the generic lower bound may not be easily interpretable, you will investigate typical examples of configuration set C and see how this affects the lower bound. If you have time, you will also study the case when the configuration set C is unknown and what can be done in this case (lower bounds, algorithm). Further details will be provided on request.

Other information The student should be mathematically strong and interested in solving theoretical problems using probability, statistics and optimization. A prior knowledge of the multi-armed bandit literature would be a great addition. While the main goal of this internship is to solve a theoretical problem, the student should be able to run some simple numerical experiments to assess the practical performance of the algorithms (in the programming language of his/her choice).

SequeL is an Inria research team based in Lille and specialized in all aspects of sequential decision making, with a rich scientific activity. This research internship proposal is part of a national research project funded by the ANR that focuses on handling non-stationarity and structure in multi-armed bandits.

References

[1] Agrawal, Rajeev, Teneketzis, Demosthenis, and Anantharam, Venkatachalam. Asymptotically Efficient Adaptive Allocation Schemes for Controlled I.I.D. processes. In *IEEE Transactions on Automatic Control*, 34(3):258–267, 1989.

[2] Maillard, Odalric-Ambrym and Mannor, Shie. Latent Bandits, Extended version of the paper accepted to ICML 2014 (paper and supplementary material) In *International Conference on Machine Learning (ICML)*, 2014.

[3] Garivier, A., Ménard, P. and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint* , arXiv:1602.07182, 2016.

[4] Cesa-Bianchi, Nicolo, Claudio Gentile, and Giovanni Zappella. A gang of bandits. In *Advances in Neural Information Processing Systems*, 2013.