Internship proposal: “Sequential Learning in Path Planning under Delayed and Anonymous Feedback"

Keywords : Multi-armed Bandit, Combinatorial Bandits, Path Planning, Exponential Average Weights algorithm, Delayed Feedback.

Supervisors :

- Alonso SILVA (Nokia Bell Labs France)
- Dong Quan VU (Nokia Bell Labs France & EDITE UPMC)

Background and topic description

The multi-armed bandit is an elegant model to solve sequential prediction problems where a learner needs to choose at each time step $t = 1, \dots, T$ an action (also called arm) in a given set, then he suffers a loss and observes a feedback corresponding to that chosen action. The objective of the learner is to guarantee that the accumulated loss is not much larger than that of the best fixed action in hindsight (that is, to minimize the regret). In this work, we focus on a bandit problem with combinatorial structure, namely Path Planning; where actions correspond to paths on acyclic directed graphs, and the loss on a chosen path is the summation of losses on edges contained in that path.

The classical algorithms for K-arms bandits problems in the literature (see e.g., EXP3 in [1]) offer an upper bound of the expected regret by a (sub-linearly) polynomial term on the cardinality of the action set. These bounds are inefficient in the case of Path Planning due to the fact that the considering graph typically contains an exponential number of paths. However, by exploiting the specific structure of the graph, we can use the exponential average weights algorithms (e.g., EXP2 in [4, 5]) and bound the regret in terms of the dimension of the representative space (typically a smaller number, e.g., the number of edges). The computational issues of these algorithms remain as an open question that will be attentively studied in this work.

Another aspect that will be considered is the feedback that the learner observes at every time step. Classically, the learner can receive either bandit feedback (he only observes the incurred loss by the chosen action) or either full-information feedback (he observes losses of all actions). Recently, several new settings of feedback are proposed by [2, 3] that model realistic problems:

- Delayed feedback: The loss of time t will only be observed at time $t + d$ for $d > 0$ fixed.
- Anonymous composite feedback: At time t, the learner only observes the total sum of the loss from time steps $t - d + 1, t - d, \dots, t$ for $d > 0$ fixed.

Several techniques based on EXP3 algorithm are presented to solve these two new settings. The natural arising question is whether these techniques can be extended into combinatorial bandits, particularly the Path Planning problem, with EXP2 algorithm of [4, 5]. The implementation of these algorithms to conduct computational experiments is another important objective.

Some potential application of this work include advertising on social networks, scheduling data transmission on a network, decision making in recommendation systems, etc.

Further information and application procedure Candidates should have a strong background in mathematics (preferentially in either sequential learning or game theory/optimization or both), be competent in scientific communication in English and having basic programming experience (Python, R or C++, etc.). Interested candidates are invited to send the following documents to *alonso.silva@nokia-bell-labs.com* and *quan_ dong.vu@nokia.com*:

- A detailed CV
- The most recent academic grades and any other useful information for the application
- The name of 1-2 references willing to provide a recommendation letter

The internship is fully funded and will be mainly based in Nokia Bell Labs France research center (Nozay - Paris region) and LINCS (Paris 13ème).

References

[1] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, *The nonstochastic multiarmed bandit problem*, SIAM journal on computing **32** (2002), no. 1, 48–77.

[2] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour, *Nonstochastic bandits with composite anonymous feedback*.

[3] Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora, *Delay and cooperation in nonstochastic bandits*, Journal of machine learning research **49** (2016), 605–622.

[4] Nicolo Cesa-Bianchi and Gábor Lugosi, *Combinatorial bandits*, Journal of Computer and System Sciences **78** (2012), no. 5, 1404–1422.

[5] Varsha Dani, Sham M Kakade, and Thomas P Hayes, *The price of bandit information for online optimization*, Advances in Neural Information Processing Systems, 2008, pp. 345–352.