# Internship proposal (research): "Sequential Learning in Combinatorial Bandits under Delayed and Anonymous Feedback"

**Keywords:** Sequential learning, multi-armed bandits, combinatorial bandits, delayed feedback.

**Lab:** Laboratoire d'Informatique de Grenoble (LIG), Grenoble, France (head: Eric Gaussier)
**Team:** POLARIS (head: Arnaud Legrand)

**Supervisors:**
Patrick Loiseau (Inria/Univ. Grenoble Alpes, LIG) – patrick.loiseau@univ-grenoble-alpes.fr
Dong Quan VU (Nokia Bell Labs France & EDITE UPMC) – quan_dong.vu@nokia.com

### Background

The multi-armed bandit (MAB) problem is an elegant model to solve sequential prediction problems. A learner needs to choose at each time step 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 action-in-hindsight (that is, to minimize the regret).

In this work, we focus on bandit problems with *combinatorial structure* and linear loss function. The classical algorithms for K-arms bandits problems such as EXP3 [1] offer a regret upper-bound that is inefficient in this case due to the fact that the considered action set typically contains an exponential number of actions. However, exploiting the combinatorial structure of the action set, we can use the exponential average weights algorithms (e.g., EXP2 [4]) to bound the regret in terms of the dimension of the representative space (typically a smaller number). This algorithms work, however, only under the classical feedback models: 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).

### Work description

In this internship, the goal is to extend combinatorial bandits algorithms and results to more complex feedback structure that would enable realistically modeling applications such as advertising on social networks, scheduling data transmission on a network, or decision making in recommendation systems.

To that end, we will leverage several new settings of feedback that were recently proposed by [2, 3]: $(i)$ Delayed feedback: The loss of time $t$ will only be observed at time $t+d$ for $d > 0$ fixed. $(ii)$ Anonymous composite feedback: At time $t$, the learner only observes the total sum of the loss from time steps $t-d+1, t-d, \ldots, t$ for $d > 0$ fixed. For these settings, techniques based on classical EXP3 algorithm for MAB are presented in [2, 3]. The student will investigate theoretically whether these techniques can be extended to combinatorial bandits in order to obtain algorithms with good regret bounds for these cases. Then, he/she will propose an efficient implementation of the new extended algorithms and perform simulations to validate the results and to evaluate the actual performance of the algorithms.

### Expected abilities of the student

Strong background in mathematics (in particular probability and statistics); if possible initiation to sequential learning; and basic programming experience (e.g., Python, R or C++).

## 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.