# Internship Proposal (Research): "Fair Online Classification with Partial Feedback"

**Keywords:** Online Learning, Fairness, Machine learning

**Host lab & team:** Laboratoire d'Informatique de Grenoble (LIG), Grenoble, France; POLARIS team

**Supervisors:**
Nicolas Gast (Inria) – `http://polaris.imag.fr/nicolas.gast/` – `nicolas.gast@inria.fr`
Patrick Loiseau (Inria & Ecole Polytechnique) – `http://lig-membres.imag.fr/loiseapa/` – `patrick.loiseau@inria.fr`

### Background

Many real world problems can be viewed as classification tasks. In lending, for instance, banks study the applicants' data to predict whether they will repay their loans or default. In online advertising, an algorithm decides who to target by predicting whether the advertised item will be interesting to the user or not. In most of these cases, however, the classification algorithm receives a feedback for its decisions only from one side: whether the prediction is correct becomes known only for "selected" applicants, e.g., the information about whether an applicant repays or defaults becomes known only for those who actually were accepted to get loans. This one-sided (or partial) observation structure can naturally lead to fairness issues (discriminations) where the algorithm shows degradation in performance on some groups of candidates, usually minorities. Such discrimination is at best unethical and in many cases prohibited (e.g., when groups are defined by race, gender, or sexual orientation).

### Goal of the internship

In this internship, we propose to study the problem of *fair online classification with partial feedback*. We will use the paper [1] as a starting point. In this paper, the authors study the problem of fair binary classification where fairness is defined as equalizing false positive rates among the groups at each period of time from 1 to $T$. They reduce the problem of classification with partial observation to a contextual bandit problem and implement their solution using the algorithm of [2] and a cost-sensitive classification approach from [3]. They show that the resulting algorithm reaches an optimal regret of $O(\sqrt{T})$ and is memory efficient.

The algorithm developed in [1], however, has several issues. The authors impose that the false positive rates are approximately equal at every stage of decision. This is quite restrictive and leads to very suboptimal decisions at the beginning of learning where the algorithm has to accept everyone at the first several rounds. In addition, the developed algorithm, while being efficient, is complex to implement.

In this work, our aim is to find a simple and efficient solution to the problem of classification with partial feedback. To this end, the approach that we propose is to implement a resampling technique which can be used on top of existing bandit algorithms. The basic idea is that at each period after the algorithm takes its decision by assigning a label, we will resample the label such that the equality of opportunity condition is (asymptotically) satisfied. This approach is similar in spirit to the approach used in [4] for offline supervised learning. The intern will jointly propose an appropriate resampling technique and perform a theoretical analysis of the regret of such approach for classical bandit algorithms such as EXP4 and/or Thompson Sampling; and he/she will perform numerical simulations to investigate the method's performance in practical cases.

### Expected ability of the student and additional information

A strong background in probability is necessary; knowledge of classification and of online learning is a strong plus. The internship is part of the Explainable and Responsible AI chair of the MIAI@Grenoble Alpes institute and will be hosted in the POLARIS team, a joint team between Inria and LIG (Grenoble CS lab). Besides the two supervisors, the intern will also work with Vitalii Emelianov, a PhD student in the team. The internship may be continued as a PhD. For more information, please contact `patrick.loiseau@inria.fr` and `nicolas.gast@inria.fr`.

## References

[1] Y. Bechavod, K. Ligett, A. Roth, B. Waggoner and Z. S. Wu. Equal Opportunity in Online Classification with Partial Feedback. In the Proceedings of NeurIPS (2019).

[2] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li and R. E. Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In the Proceedings of ICML (2014).

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford and H. Wallach. A Reductions Approach to Fair Classification. In the Proceedings of ICML (2018).

[4] M. Hardt, E. Price and N. Srebro. Equality of Opportunity in Supervised Learning In the Proceedings of NIPS (2016).