Internship subject: Online Boosting

Supervisors: Pierre Gaillard (Inria Grenoble) and Olivier Wintenberger (Sorbonne Université)

December 8, 2021

Today, the world is facing an unprecedented increase in the volume and speed of available data streams. Many applications must move from offline methods to sequential methods that can acquire, adapt to, and process data on the fly. At the same time, the data is becoming increasingly sophisticated. Traditional statistical assumptions such as stationarity (or i.i.d. data) are no longer satisfied. Designing efficient algorithms that can learn from data as it comes in with as few assumptions as possible is a major challenge in today's machine learning. Harnessing the potential of these real-time data streams is the goal of online learning, the domain of this internship project.

A popular approach in classical machine learning is boosting, which builds and combines a set of weak learners (such as simple decision trees) to produce a strong learner with higher predictive ability. Methods based on boosting (such as XGboost or LightGBM, Chen et al. [2015]) achieve state-of-the-art performance for many machine learning problems. Recent work has considered adapting these types of methods to the online learning paradigm: training and optimizing new weak learners on the fly to be combined based on sequentially collected data (see Chapter 12 of Hazan [2019] and references therein). Yet many interesting questions remain. For example, what are the precise convergence rates that can be obtained from these methods in a non-parametric setting? Can optimal rates of convergence be recovered? Moreover, the theoretical adaptation of boosting to the online setting still suffers from many flaws that create a huge gap between theory and practical applications that do not meet the assumptions. The objective of this internship is to find a good practical application for online boosting, to model it in a clean theoretical framework and to provide mathematical guarantees. One possible application is online density estimation by a mixture of Gaussians, the latter being considered as weak learners to be estimated online [Li and Barron, 2000]. Optimistic algorithms [Rakhlin and Sridharan, 2013] could also be a promising technical tool that we would like to investigate and try to leverage in an online boosting algorithm.

This internship is a continuation of another internship Darolles [2021] on the subject done last year and from which the student may start.

References

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.

Aymeric Darolles. Study of online boosting methods. master thesis, 2021.

Elad Hazan. Introduction to online convex optimization. arXiv preprint arXiv:1909.05207, 2019.

Jonathan Q Li and Andrew R Barron. Mixture density estimation. In Advances in neural information processing systems, pages 279–285, 2000.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. arXiv preprint arXiv:1311.1869, 2013.