ONLINE GRADIENT BOOSTING

Context:

Today, the world faces an unprecedented increase in the volume and speed of available data streams. Many applications must move 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 significant challenge in today's machine learning. Harnessing the potential of these real-time data streams is the goal of online (streaming, recursive) learning.

A popular approach in classical machine learning is the gradient boosting introduced by Friedman [2002], which builds and combines a set of weak learners (such as simple decision trees) to produce a strong learner with higher predictive ability. Offline (batch, non-recursive) methods based on gradient 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 adopting 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). Nevertheless, the recent Master thesis of Darolles [2021] shows several issues of this recent approach, the main one being that the notion of weak learner excludes simple binary trees. Moreover, several questions remain: Do optimal guarantees from the batch setting extend to the online setting? Could the considerable gap between the good theoretical results of online boosting be implemented and observed in practice?

The objective of the doctoral project: provide new online Gradient Boosting algorithms

This project aims at building online gradient boosting methods that learn from the data step by step, improving when observing more information. In other words, the goal is to adapt successful Boosting algorithms to the sequential paradigm. The project mixes theoretical research with applications on real data through interdisciplinary applications. It plans to design algorithms with robust theoretical guarantees and good practical performance.

Goal 1: Provide new algorithms for online learning inspired by the Boosting literature.
Our idea is to combine weak learners using aggregation techniques (see the monograph of Cesa-Bianchi and Lugosi [2006]) with this focus in mind. The key idea is to use both technical tools from the online learning community and Boosting used by the batch learning community [Freund et al., 1999]. A research direction we have in mind is the setting of sequential forecasting. A learner is asked in a series of rounds to predict the following observation based on past and contextual information. For example, a meteorologist could be asked to forecast the next day's temperature each day. Actual boosting algorithms would most likely provide accurate forecasts in the offline setting (i.e., non-sequential). In the sequential setting, however, since the number of data increases over time, the boosting algorithm should be tuned to be increasingly refined (by considering more weak learners) to balance the bias-variance trade-off. A solution that we would like to investigate in this project is learning from expert advice techniques that sequentially combine a set of experts to perform and the best in hindsight. The experts would be weak learners (simple trees, Gaussian kernels, or shallow neural networks). Following the spirit of Boosting, we will regularly enrich the set of weak learners seen as experts by adding new ones to the online combination that tries to correct the errors of previous ones.

Goal 2: Provide solid theoretical results for tree-based online boosting.
We seek to provide robust theoretical guarantees on the performance of our new procedures. The regret classically measures the performance of online algorithms. The regret is the difference between the cumulative past losses suffered by the algorithm and the best-fixed function in a reference class F. The larger the reference class of function F is, the harder it is for the learner to be competitive with it, and the more significant is the regret. In online learning, various algorithms aggregating simple trees achieve a small regret for complex nonparametric classes of functions such as Sobolev spaces [see Gaillard and Gerchinovitz, 2015, Rakhlin and Sridharan, 2014, 2015, Beygelzimer et al., 2015]. However, these algorithms all suffer from severe drawbacks such as suboptimality of the performance guarantee or computational inefficiency for most functional spaces F.
Moreover, the class F needs to be known and fixed in advance. We want to design an online boosting algorithm with a sharp regret bound for specific functional spaces (Sobolev, Besov). In particular, we seek to compare the theoretical performance with the existing ones. Our goal is to achieve this by designing a combination of an increasing number of experts (well-designed weak-learners), each expert trying to learn how to correct the error of the other ones. [Mourtada and Maillard, 2017] studied aggregation with an increasing number of experts in online learning. Hopefully, our approach will provide new tools and algorithms for online nonparametric regression.

Goal 3: Extension to other weak learners.
Simple trees are the typical weak learners in batch algorithms for computational considerations because they are piecewise constant and straightforward to fit. However, other weak learners have considerable approximation power when combined. The recursive step for tuning consists of a simple gradient step that is adaptable to any weak learners. Online gradient boosting should extend to weak learners that are not trees. For instance, we could use Gaussian kernels as weak learners. Li and Barron [2000] consider an i.i.d. setting where a distribution with density f generates the data. They aim to estimate this density by a mixture of Gaussian kernels. They proposed an iterative estimation based on the maximum likelihood principle. Their Boosting method is batched and must go through all the data at each step. Another promising class of weak learners is shallow neural networks that, when combined, have excellent approximation guarantees in large dimensions, as shown by Barron [1994].
Moreover, the optimal number of elements |F| explicitly depends on the number of data and thus should increase over time in an online approach. We want to enrich the mixture sequentially and optimize its size simultaneously. We will explore and compare several schemes to add a new element to the mixture at each round. Inspired by Boosting, we would add weak learners that correct the cumulative error of the previous element on past data thanks to a gradient step.

Goal 4: Developing software and applying the method to real-world problems: The final objective of this doctoral project is to develop a good practical application for online Boosting, developing codes in packages on R or Python software. Based on the experience of the supervisors, we would like to highlight the practical results by participating in competitions on electricity load forecasting.

References:

Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. Machine learning 14.1, 1994.
Alina Beygelzimer, Elad Hazan, Satyen Kale, and Haipeng Luo. Online gradient boosting. In Advances in neural information processing systems, 2015.
Nicolò Cesa-Bianchi and Gabor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015.

Aymeric Darolles. Study of online boosting methods. Master Thesis, MVA and Centrale Paris, 2021 (on demand)
Jerome H. Friedman. Stochastic gradient boosting. Computational statistics & data analysis 38.4, 2002.
Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. Journal-Japanese Society For Articial Intelligence, 14, 1999.
Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.
Jonathan Q Li and Andrew R Barron. Mixture density estimation. In Advances in neural information processing systems, 2000.
Jaouad Mourtada and Odalric-Ambrym Maillard. Efficient tracking of a growing number of experts. arXiv preprint arXiv:1708.09811, 2017.
Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In Conference on Learning Theory, 2014.
Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. arXiv preprint arXiv:1501.06598, 2015.

References and roles of the supervisors:

The experience of Pierre Gaillard (50%) in computer science complements the ones in statistics of Olivier Wintenberger (50%). We have already collaborated on different subjects in online learning and co-supervised the Master thesis Darolles [2021] on the subject.

Eric Adjakossa, Yannig Goude and Olivier Wintenberger. Kalman Recursions Aggregated Online. Arxiv preprint arXiv:2002.12173
Pierre Gaillard and Olivier Wintenberger Efficient online algorithms for fast-rate regret bounds under sparsity, NIPS, 2018.
Pierre Gaillard and Sebastien Gerchinovitz. A chaining algorithm for online nonparametric regression. COLT, 2015.
Pierre Gaillard and Olivier Wintenberger Sparse Accelerated Exponential Weights, AISTAT, JMLR, 2017
Olivier Wintenberger Optimal learning with Bernstein Online Aggregation, Machine Learning, 106, 2017.
Olivier Wintenberger Stochastic Online Convex Optimization; Application to probabilistic time series forecasting. ArXiv preprint arXiv:2102.00729, 2021.
Oleksandr Zadorozhnyi, Pierre Gaillard, Sebastien Gerschinovitz, Alessandro Rudi. Online nonparametric regression with Sobolev kernels. ArXiv preprint arXiv:2102.03594, 2021
Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, Sébastien Gerchinovitz.Online nonparametric learning, chaining, and the role of partial feedback. COLT, 2017.

Suitability for the institute offering a doctoral research grant:

The study of online algorithms such as online Gradient Boosting is the subject of much attention in the learning community. This research axis belongs naturally within the Sorbonne Center for Artificial Intelligence (SCAI) mathematical challenges as a rigorous study of complex algorithms.

Profile of the candidate:
Student with a Master's degree in statistics, statistical learning or an engineering degree (with a specialization in applied mathematics and/or statistics and/or statistical learning statistics) with a strong background in theoretical and numerical statistics numerical (programming in R or Python).