The LASSO and some Far-West hostilities

Keywords: federated machine learning, distributed optimization, missing values, sparse regression

In a high-dimensional sparse regression framework, the LASSO estimator has been introduced as an efficient variable selector. The latter has been extensively studied, see for instance the book [1].
Much work remains to be done regarding the application of such statistical methods in realistic contexts, i.e. in presence of missing data or in a distributed learning setting. Indeed, on the one hand, today's large-scale data makes unavoidable the problem of missing values [2] (due for instance to "forgot to fill in the form" entry, failure of the measuring device, no time to measure in an emergency situation, aggregating data sets from multiple sources), which represent a real obstacle to sparse regression.
On the other hand, for privacy reasons, learning must be often done in a federated way [3,4]. In such a case, a central server updates the global model of sparse regression, given the non-exhaustive information sent by the local agents. The fact that the whole dataset is not anymore at our disposal makes the sparse regression with missing covariates even more arduous. During this internship, we will study sparse linear regression with missing covariates in the framework of distributed learning.

The subject combines two aspects of a scientific work: on the one hand, a more methodological development could lead to efficient algorithms; on the other hand, a more thorough theoretical study of this issue will allow to establish nice statistical and optimization results. Both directions are important, and can be modulated according to the candidate's affinities.

Supervisors: Claire Boyer (Sorbonne Université), Aymeric Dieuleveut (Ecole Polytechnique), Erwan Scornet (Ecole Polytechnique)

Required skills: M2 level trainee in statistics/machine learning/optimization. Motivation for pursuing a PhD. thesis in machine learning is a real plus.
Applicants should send CV, transcripts of the last two years and the name of a referee to
claire.boyer@sorbonne-universite.fr, aymeric.dieuleveut@polytechnique.edu
erwan.scornet@polytechnique.edu

[1] Statistical Learning with Sparsity: The Lasso and Generalizations, Hastie, Tibshirani and Wainwright, 2015
[2] On the consistency of supervised learning with missing values. Josse, Prost,, Scornet & Varoquaux (2019).
[3] Federated learning: Challenges, methods, and future directions. Li, Sahu, Talwalkar & Smith (2020). *IEEE Signal Processing Magazine*
[4] Advances and Open Problems in Federated Learning, Kairouz et al., 2019