1 year ML research engineer position in Inria Lyon

Context

Generalized Linear Models are massively used in large scale applications such as genomics, geophysics or signal processing. Their considerable impact has only been made possible by high quality implementations in open source packages. The reference Python package for general Machine Learning, scikit-learn, implements many GLMs in an API which has become the gold standard for practitioners. However, it can be slow to fit models on very large data and the core codebase, written in Cython for efficiency, prevents the handling of user-defined losses and penalties. In [1], Celer, a python library implementing a fast solver for the Lasso and other sparse models was proposed, with accelerations up to two orders of magnitude over scikit-learn[1]. This package follows the sklearn API and thus provides a drop-in replacement in existing codebases. It also implements models missing in sklearn, such as Group Lasso. It has been downloaded more than 20,000 times and has been used successfully in genomics, medicine or geophysics research works. Recently, a generalization of the Celer algorithm was proposed [2], which would allow it to handle more complex and powerful penalties (mixed penalties, non convex penalties). Instead of Cython, a proof-of-concept package has been implemented in Numba, a novel approach that makes the packaging and the interfacing with user defined code much easier. Our POC allows for straightforward user definition of any penalty and losses, opening the way to scale up more complex models[2]

The goal of this project is to bring our proof-of-concept project andersonCD to maturity by including it in Celer in order to ensure its diffusion and its use in practical applications.

Mission

The recruited person will lead the extension of Celer to a much wider class of Generalized Linear Models, following the technique proposed in [2]. In particular, he/she will handle the extension to group penalties, and write default estimators for popular estimators (Logistic

[1] https://github.com/mathurinm/celer
[2] https://github.com/mathurinm/andersoncd

regression, Poisson regression, etc.) and penalties (Huber, MCP, SCAD, Berhu, etc.) These will require the design of automated cross-validation procedures and the compliance of estimators into sklearn Pipeline objects. The candidate will write documentation and examples to ensure the dissemination of the project. For benchmarking, he/she will use the Benchopt library, and make contributions to it[3]. Benchopt is an open source benchmarking toolbox developed collaboratively, which aims at making easy and reproducible comparison of optimization algorithms. In particular, the candidate will develop new benchmarking tools for non convex optimization algorithms, assessing not only the convergence speed but the quality of the solution in terms of metrics such as generalization error and sparsity. This will ensure that other researchers can easily evaluate the proposed approach and potential subsequent improvements.

Requirements

- Proficiency in Python, experience with numpy/scipy stack
- Background in optimization (first order methods)

The following would be a plus:

- Knowledge of scikit-learn, cython and numba
- Experience with continuous integration tools, unit testing
- Experience in composite non smooth/sparse optimization

Contract details
Starting date: December 2021 or January 2022.
Duration: 1 year
Location: Inria Lyon
Salary: according to Inria salary grid

Contact mathurin.massias@gmail.com

References

[1] M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. Dual extrapolation for sparse Generalized Linear Models. *Journal of Machine Learning Research*, 21(234):1–33, 2020.

[2] Q. Bertrand and M. Massias. Anderson acceleration of coordinate descent. In *AISTATS*, 2021.

[3] https://benchopt.github.io/