Methods and Algorithms for Approximate Machine Unlearning

Internship proposal for academic year 2024-2025

- Intended for: Students in second year of master in machine learning or related topic.
- Duration: 5 or 6 months, preferably from April 2025.
- Supervisors: Olivier Cappé (CSD, DI-ENS, Ecole Normale Supérieure—PSL, CNRS), Jamal Atif (MILES Team, LAMSADE, Université Paris Dauphine—PSL, CNRS)[*].
- Location: Institut PR[AI]RIE - Paris Santé Campus, 75015 Paris.
- Follow up: Priority will be given to candidates interested by pursuing a PhD thesis on the topic (fully funded 3 years PhD position available).

About

In recent years, the impressive development of systems based on machine learning has raised the awareness about potential misuses of personal data or more generally human-generated data that deserves to be protected. Among other concerns, this raises the question of the feasibility of *unlearning* defined as the ex post removal of some of the data points that have been used to train a machine learning model.

A framework proposed to derive principled solutions to this problem -beyond the obvious option of retraining the model from scratch with the removed data omitted- is that of *approximate unlearning* [1, 2], where the approximate unlearning guarantees are obtained using methods initially developed for analyzing differentially private data protection mechanisms. Recent works have investigated how this idea can be instantiated when using gradient-based learning in convex [3] or non-convex models [4].

The aim of the internship is to develop a deep understanding of these works, both from the theoretical and practical point of views, and to investigate research directions from this starting point.

References

[1]. A. A. Ginart, M. Y. Guan, G. Valiant, and J. Zou. Making AI forget you: Data deletion in machine learning. NeurIPS, 2019.

[2]. A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh, “Remember what you want to forget: Algorithms for machine unlearning, NeurIPS 2021.

[3]. S. Neel, A. Roth, and S. Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. ALT, 2021.

[4]. E. Chien, H. Wang, Z. Chen, and P. Li. Langevin Unlearning: A New Perspective of Noisy Gradient Descent for Machine Unlearning. NeuIPS, 2024.

[*]Contact: olivier.cappe@cnrs.fr, jamal.atif@dauphine.fr