Privacy Auditing for Machine Learning

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)[*].
- 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).

Differential Privacy (DP) is a framework that has been elaborated since 2006 so a to produce a set of results and methods that can be applied to modern data processing pipelines, such as those used in machine learning, in order to protect personal or, more generally, private or sensitive data from unwanted disclosure. Despite this significant body of works and the adoption of DP by some parties, most machine learning models are still trained nowadays using largely unknown, if any, data privacy protection measures.

Among other concerns, this raise the question of auditing -in the sense of empirically measuring and quantifying- privacy guarantees from existing machine learning pipelines. An interesting line of works is based on the strong connection between concepts used in DP and the probability of success of *membership inference attacks (MIA)* [1,2,3]. In this view, the purpose of carrying out MIA is to obtain numerical lower bounds on the privacy leakage, as defined by DP. This idea is however faced with various difficulties, some of them related to the practical feasibility of the approach in large-scale machine learning applications [4] and others to the applicability of the approach to the (challenging but more realistic) setting where retraining of the model is not an option (which is sometimes referred to as post-hoc) [5,6].

The aim of the internship is to familiarize with these works, both from the theoretical and practical points of views (experimenting in particular with the ideas exposed in [5]), and to investigate related research directions.

References

[1]. P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. ICML, 2015.

[2]. J. Dong, A. Roth, and W. J. Su, Gaussian differential privacy. JRSSB, 2022.

[3]. B. Kulynych, J. F. Gomez, G. Kaissis, F. Calmon, and C. Troncoso. Attack-Aware Noise Calibration for Differential Privacy. NeurIPS, 2024

[4]. T. Steinke, M. Nasr, and M. Jagielski. Privacy auditing with one (1) training run. NeurIPS, 2023

[5]. M. Kazmi, H. Lautraite, A. Akbari, Q. Tang, M. Soroco, T. Wang, S. Gambs, M. Lécuyer. PANORAMIA: Privacy Auditing of Machine Learning Models without Retraining. NeurIPS, 2024

[6]. J. Zhang, D. Das, G. Kamath, F. Tramèr. Membership Inference Attacks Cannot Prove that a Model Was Trained On Your Data. SaTML 2025.

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