Algorithms for Simultaneous Guarantee of Robustness and Privacy in Machine Learning

- **Keywords:** differential privacy, robustness, adversarial attacks, membership inference attacks, reconstruction attacks, certified defenses
- **Duration:** 6 months
- **Supervisors:**
  Muni Sreenivas Pydi muni.pydi@lamsade.dauphine.fr
  Jamal Atif jamal.atif@lamsade.dauphine.fr
  Olivier Cappé olivier.cappe@ens.fr
- **Place:** MILES Team, LAMSADE, Université Paris Dauphine—PSL
  DI-ENS, Ecole Normale Supérieure–PSL

Context

Recent research shows that Deep Learning (DL) models are vulnerable to a variety of privacy attacks. In a *membership inference attack*, an adversary is able to identify if a specific data point is used in training the DL model. In a *reconstruction attack*, an adversary is able to approximately reconstruct the data used for training the model. Such attacks pose a serious privacy risk to the use of DL models, especially when it comes to sensitive applications like healthcare.

In addition to privacy attacks, DL models are also vulnerable to adversarial attacks that drastically reduce the performance of the algorithm by a slight change to the data either in the training phase or inference phase. Broadly, the adversarial attacks fall into two categories: (1) In *evasion attacks*, the attacker makes an imperceptible change to a test data point in the inference phase so as to reduce the performance of the algorithm on the specific data point. (2) In *poisoning attacks*, the attacker manipulates a small fraction of the training data in the training phase, so that the overall performance of the algorithm is reduced.

While there are methods available to defend against robustness [1, 2] and privacy [3] attacks, it is common for these defenses to address one type of threat while ignoring the other. This approach can leave vulnerabilities exposed if only one aspect is emphasized and the other is not adequately considered. For instance, recent research has indicated that algorithms that are designed to be adversarially robust can be more vulnerable to privacy attacks than those that do not prioritize this aspect [4]. On the other hand, algorithms that are designed to be differentially private can be more susceptible to evasion attacks than non-private algorithms in some cases [5]. Hence, it is important for both robustness and privacy to be considered and addressed together in order to effectively protect against attacks from both the fronts.

Goals

The primary goal is to develop efficient algorithms that simultaneously guarantee robustness and privacy, while scaling well to large datasets and large DL models. A secondary goal is to derive theoretical bounds that precisely capture the trade-offs / synergies between differential privacy and the two types of robustness (i.e. robustness to evasion attacks and robustness to poisoning attacks).

Organisation

- Survey the literature on existing attack modalities for privacy and robustness
- Survey the literature for SOTA defenses (particularly, certified defenses) against attacks.
- Gain an understanding of the trade-offs / synergies between privacy and robustness constraints
- Develop and implement new algorithms for simultaneous guarantee of privacy and robustness
- Publish the results in the form of a research paper at top machine learning conferences (ICML, NeurIPS, ICLR, AISTATS, AAAI) and publish the code on GitHub.

Profile

The ideal candidate will meet the following criteria.

- Pursuing a Masters degree or equivalent in Computer Science, Mathematics, Data Science or Artificial Intelligence
- Mathematical maturity, and a strong theoretical background in probability theory, statistics and machine learning
- Experience in programming (Python) and DL frameworks (PyTorch)
- Exposure to differential privacy and robust machine learning is a plus

References

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *International Conference on Learning Representations*, 2017.

[2] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*, pp. 1310–1320, PMLR, 2019.

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[4] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 241–257, 2019.

[5] N. Tursynbek, A. Petiushko, and I. Oseledets, "Robustness threats of differential privacy," *NeurIPS Workshop on Privacy-Preserving Machine Learning*, 2020.