Theoretical Limits on Privacy in Overparametrised Machine Learning Models

- **Keywords:** Differential privacy, overparametrised models, interpolation, generalization error
- **Duration:** 6 months
- **Supervisors:**
  Muni Sreenivas Pydi `muni.pydi@lamsade.dauphine.fr`
  Jamal Atif `jamal.atif@lamsade.dauphine.fr`
  Olivier Cappé `olivier.cappe@cnrs.fr`
- **Place:** MILES Team, LAMSADE, Université Paris Dauphine—PSL
  DI-ENS, Ecole Normale Supérieure–PSL

Context

Modern machine learning models are often over-parametrized, i.e., the number of model parameters typically far exceed the number of data points by orders of magnitude. In the over-parametrized setting, learning algorithms typically achieve close to 100% training accuracy and are said to be in the "interpolation regime" where the algorithm almost perfectly fits all the training data points [1]. By nature, interpolating models memorize almost all the training samples. This makes them particularly vulnerable to membership inference attacks [2], which guess whether a particular data point was used for training the model, and reconstruction attacks, which approximately reconstruct the data points that were used for training [3].

Existing defenses against privacy attacks aim to reduce overfitting through a variety of techniques such as sub-sampling and noise injection [4]. However, defenses often come with a significant drop in accuracy, indicating the necessity of sacrificing utility to guarantee privacy in the interpolation regime. The loss in accuracy is particularly steep for high-dimensional and long-tail distributions [5]. This insight is in sharp contrast to some recent theoretical results on the generalization benefits of differential privacy [6].

The ubiquity of interpolating algorithms in modern machine learning combined with their unique drawbacks in data privacy, calls for a focused study of privacy in the overparametrised setting.

Goals

The primary goal is to reevaluate the privacy utility trade-off in the interpolation regime by deriving new theoretical bounds that precisely capture the trade-off between privacy and generalization in overparametrised models. A good starting point for this is the recent work on differentially private learning with margin guarantees [7].

Organisation

- Survey existing literature on generalization error bounds for differentially private algorithms
- Survey existing literature on generalization error bounds for overparametrised algorithms
- Develop new theoretical results on privacy in the overparametrised setting
- Publish the results in the form of a research paper at top machine learning conferences (ICML, NeurIPS, ICLR, AISTATS, AAAI).

Profile

The ideal candidate will meet the following criteria.

- Pursuing a Masters degree or equivalent in Computer Science, Mathematics, Data Science or Electrical Engineering
- A strong theoretical background in probability theory, statistics and machine learning
- Exposure to differential privacy and generalization error bounds is a plus

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, IEEE, 2022.

[3] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," *Conference on Neural Information Processing Systems*, 2022.

[4] 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.

[5] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Chasing your long tails: Differentially private prediction in health care settings," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 723–734, 2021.

[6] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

[7] R. Bassily, M. Mohri, and A. T. Suresh, "Differentially private learning with margin guarantees," *Conference on Neural Information Processing Systems*, 2022.