Privacy in Overparametrized Machine Learning Models

- **Keywords:** Differential privacy, linear regression, overparametrised models, interpolation, membership inference attacks
- **Duration:** 5 or 6 months
- **Supervision:** Muni Sreenivas Pydi, Jamal Atif (MILES Team, LAMSADE, Université Paris Dauphine—PSL), Olivier Cappé (CSD, DI-ENS, Ecole Normale Supérieure—PSL)*
- **Location:** Paris Santé Campus, 75015 Paris
- **Follow up:** Priority will be given to candidates interested by pursuing a PhD thesis on the same topic (fully funded 3 years PhD position available)

Context

Modern machine learning models (e.g. deep neural networks) are often over-parametrized, i.e., the number of model parameters p far exceed the number of data points n. In the over-parametrized setting, learning algorithms 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. This makes them particularly vulnerable to membership inference attacks (MIA) [1], wherein an adversary reveals whether a particular data point was used for training the model. For example, a simple and effective attack is to check the loss incurred by the model on a data point. If the loss is close to zero, then the data point is likely in the training set.

A good starting point to understand the interpolation regime in modern machine learning models (e.g. deep neural networks) is to analyze linear regression where the number of data points exceeds the number of parameters. Concretely, in a linear regression setup with design matrix $X \in \mathbb{R}^{n \times p}$, observation vector $y \in \mathbb{R}^p$, parameter vector $\beta \in \mathbb{R}^p$ and a Gaussian model $y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}_p, \sigma^2 I_{p \times p})$, the phenomenon of benign overfitting has been studied by analyzing the minimum norm interpolating solution $\hat{\beta}$ that minimizes $\|\hat{\beta}\|_2$ while satisfying $y = X\hat{\beta}$ in the case when $p \gg n$ [2, 3]. As $n, p \to \infty$ and $p/n \to \gamma > 1$, it is shown that the minimum norm interpolator indeed has lower risk with higher p. In the same regime, a recent work [4] shows that the model becomes more vulnerable to MIA with higher p.

Differential Privacy (DP) has become the de facto standard for enforcing user privacy on machine learning systems [5]. Recent works study the problem of differentially private linear regression [6, 9, 10, 11], derive bounds on the best attainable error rates [8] and optimal rates of convergence of parameters under the Gaussian model [7] — all in the low-dimensional setting of $p < n$. However, analogous questions in the overparametrized regime remain unanswered.

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

Goals

The primary goal of this internship is to understand the fundamental limits of privacy in the overparametrized linear regression setting with Gaussian model. The project will start with a close study of benign overfitting in linear regression under Gaussian model under the regime of $n, p \to \infty$ and $p/n \to \gamma > 1$. Then, the project will explore new techniques for guaranteeing DP in linear regression under the overparametrized setting, begining with exploring the efficacy of existing techniques such as DP-SGD, parameter-perturbation, and data-perturbation under the new regime.

Organization

- Understand the phenomenon of benign overfitting in the context of linear regression [2, 3].
- Understand existing approaches for differentially private linear regression [6, 7, 8, 10, 11].
- Develop new theory and algorithms for differentially private linear regression in the overparametrised setting.
- Evaluate the robustness of proposed approaches to MIA inspired by [4].

Profile of Candidate

- Pursuing a Master's degree in Computer Science or Mathematics
- Strong theoretical background in probability theory, statistics and machine learning
- Exposure to differential privacy in the form of a masters-level course is a plus

References

[1] 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.

[2] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30063–30070, 2020.

[3] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020.

[4] J. Tan, B. Mason, H. Javadi, and R. Baraniuk, "Parameters or privacy: A provable tradeoff between overparameterization and membership inference," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17488–17500, 2022.

[5] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[6] D. Alabi, A. McMillan, J. Sarathy, A. D. Smith, and S. P. Vadhan, "Differentially private simple linear regression," *Proceedings on Privacy Enhancing Technologies*, vol. 2022, pp. 184 – 204, 2020.

[7] T. T. Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," *The Annals of Statistics*, vol. 49, no. 5, pp. 2825–2850, 2021.

[8] P. Varshney, A. Thakurta, and P. Jain, "(nearly) optimal private linear regression for sub-gaussian data via adaptive clipping," in *Annual Conference Computational Learning Theory*, 2022.

[9] Y.-X. Wang, "Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain," in *Conference on Uncertainty in Artificial Intelligence*, 2018.

[10] O. Sheffet, "Old techniques in differentially private linear regression," in *Algorithmic Learning Theory*, pp. 789–827, PMLR, 2019.

[11] K. Amin, M. Joseph, M. Ribero, and S. Vassilvitskii, "Easy differentially private linear regression," in *International Conference on Learning Representations*, 2023.