Performative Learning for long-term Trustworthy Machine Learning

Edwige Cyffers and Yann Chevaleyre

January 2026

Research Statement

Differential privacy provides formal guarantees against information leakage in machine learning, but introduces a privacy-utility trade-off traditionally viewed as a constraint on achievable utility. However, this perspective neglects a key feedback effect: if users perceive the privacy guarantees as insufficient, they may opt out of data collection entirely or even perturb their inputs (possibly adversarially) to protect their privacy, thereby modifying the training distribution itself. The term *Performative privacy* [8] waseven coined to describe active resistance to data collection due to data misuse, for instance people wearing large hoodies to escape public surveillance, but has not been studied yet from a mathematical point of view. Both Adversarial Attacks and Performative learning offer however natural frameworks to analyze this phenomenon, as it explicitly models distribution shifts induced by a deployed model and propose mitigation.

The goal of this PhD is to mathematically encode such a setting and to prove under which conditions enforcing privacy leads to a better equilibrium.

Context

The setting of this PhD lies at the intersection of Differential Privacy, Performative Learning, and Adversarial Attacks. We briefly review these three topics below and discuss how they relate to each other.

Differential Privacy [2] mathematically quantifies the worst-case information leakage about a single entity in the dataset through its influence on the algorithm's outputs, thus providing protection against all attacks. More precisely, an algorithm $\mathcal{A}$ is differentially private if, for all pairs of datasets $D \sim D'$ differing by a single participant, and every subset $\mathcal{S} \subset Z$, the following inequality holds:

$$\mathbb{P}(\mathcal{A}(D) \in \mathcal{S}) \leq \exp(\varepsilon)\, \mathbb{P}(\mathcal{A}(D') \in \mathcal{S}) + \delta.$$

DP is used in deployment (e.g., Apple statistics on emoji, Google next-word prediction, LinkedIn audience measurement, the US Census, Wikimedia pageviews) and enables private statistics and private machine learning [3].

Performative Learning [7] addresses distribution changes due to model deployment, aiming to minimize

$$\mathrm{PR}(\theta) = \mathop{\mathbb{E}}_{Z \sim \mathcal{D}(\theta)} \ell(Z; \theta).$$

Unlike classical machine learning where $\mathcal{D}$ is fixed, here it is parameterized by the same parameter θ used in the loss function. Finding a performatively optimal solution requires accounting for the model's performative effect on the distribution. Typical examples include loan applications or hiring, where applicants may modify their features to increase their probability of being classified in the positive class, and the optimal classifier should thus take into account the extent to which a given feature may be modified. This problem has also been studied in terms of regret minimization [5], to ensure that the whole trajectory minimizes the risk.

Adversarial Attacks [9, 4] are imperceptible perturbations added to the input data that induce high loss for the learning. The adversarial risk is traditionnaly defined as:

$$\mathrm{AR}(\theta) = \mathop{\mathbb{E}}_{Z \sim \mathcal{D}} \left[\sup_{Z' : d(Z', Z) \leq \epsilon} \ell(Z'; \theta) \right].$$

Defenses include adversarial training [6], which minimizes this worst-case risk, and certified methods [1], which provide provable robustness guarantees. Adversarial attacks can be seen as a form of performative feedback, where users modify their input to maximize their loss.

References

[1] Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. "Certified Adversarial Robustness via Randomized Smoothing". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1310–1320.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Berlin, Germany: Springer Berlin Heidelberg, 2006.

[3] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013).

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[5] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. "Regret Minimization with Performative Feedback". In: *ICML*. 2022.

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: `https://openreview.net/forum?id=rJzIBfZAb`.

[7] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. "Performative Prediction". In: vol. 119. Proceedings of Machine Learning Research. PMLR, 2020.

[8] Scott Skinner-Thompson. "Performative Privacy". In: *UC Davis Law Review* 50.4 (2017). NYU School of Law, Public Law Research Paper No. 17-10; U of Colorado Law Legal Studies Research Paper No. 17-4.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014. URL: `http://arxiv.org/abs/1312.6199`.