Arbitrary Decisions are a Hidden Cost of Differentially Private Training

Bogdan Kulynych[1], Hsiang Hsu[2], Carmela Troncoso[1] and Flavio P. Calmon[2]

[1]*EPFL SPRING Lab*
[2]*Harvard University*

Abstract
Mechanisms used in privacy-preserving machine learning often aim to guarantee differential privacy (DP) during model training. Practical DP-ensuring training methods use randomization when fitting model parameters to privacy-sensitive data (e.g., adding Gaussian noise to clipped gradients). We demonstrate that such randomization incurs predictive multiplicity: for a given input example, the output predicted by equally-private models depends on the randomness used in training. Thus, for a given input, the predicted output can vary drastically if a model is re-trained, even if the same training dataset is used. The predictive-multiplicity cost of DP training has not been studied, and is currently neither audited for nor communicated to model designers and stakeholders. We derive a bound on the number of re-trainings required to estimate predictive multiplicity reliably. We analyze–both theoretically and through extensive experiments–the predictive-multiplicity cost of three DP-ensuring algorithms: output perturbation, objective perturbation, and DP-SGD. We demonstrate that the degree of predictive multiplicity rises as the level of privacy increases, and is unevenly distributed across individuals and demographic groups in the data. Because randomness used to ensure DP during training explains predictions for some examples, our results highlight a fundamental challenge to the justifiability of decisions supported by differentially private models in high-stakes settings. We conclude that practitioners should audit the predictive multiplicity of their DP-ensuring algorithms before deploying them in applications of individual-level consequence.

Extended Abstract[1]

In many high-stakes prediction tasks (e.g., lending, healthcare), training data used to fit parameters of machine-learning models are privacy-sensitive. A standard technical approach to ensure privacy is to use training procedures that satisfy *differential privacy* (DP) [1, 2]. DP is a formal condition that, intuitively, guarantees a degree of plausible deniability on the inclusion of an individual sample in the training data. In order to satisfy this condition, non-trivial differentially-private training procedures use some degree of randomization (see, e.g., Chaudhuri et al. [3], Abadi et al. [4]). The noisy nature of DP mechanisms is key to guarantee plausible deniability of a record's inclusion in the training data. Unfortunately, randomization comes at a cost: it often leads to decreased accuracy compared to non-private training. Reduced accuracy, however, is not the only cost incurred by differentially-private training. DP mechanisms can also increase *predictive multiplicity*, discussed next.

In a prediction task, there can exist multiple models that achieve comparable levels of accuracy yet output drastically different predictions for the same input. This phenomenon is known as predictive multiplicity [5], and has been documented in multiple realistic machine-learning

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1] The full version is available at https://arxiv.org/abs/2302.14517

CEUR-WS.org/Vol-3442/paper-15.pdf

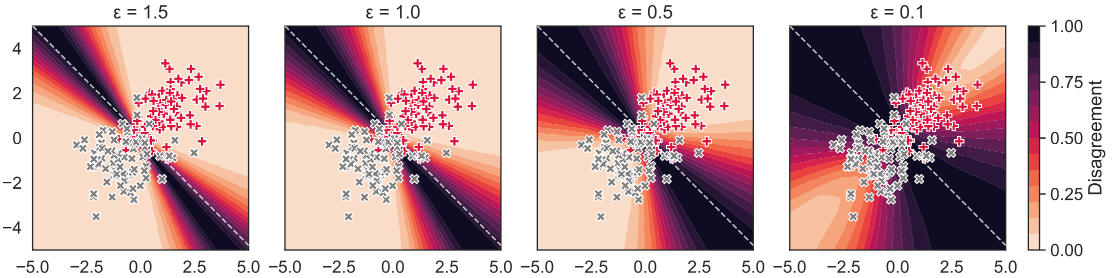

Figure 1: The region of examples which exhibit high variance of decisions (dark) across similarly-accurate models grows as the privacy level increases (lower ε). Each plot shows the level of decision disagreement across $m = 5000$ logistic-regression models (darker means higher disagreement) trained with varying levels of differential privacy (ε value, lower means more private) using the objective-perturbation method [3]. All models attain at least 72% accuracy on the test dataset (50% is the baseline). The disagreement value of 1.0 means that out of the m models, around 50% output the positive decision, whereas the other 50% output the negative one for a given example. The values of disagreement are shown for different possible two-dimensional examples, with x and y axes corresponding to the two dimensions. The markers show training data examples belonging to two classes (denoted as × and +, respectively). Without DP, there is a single optimal classification model. The dotted line - - shows the decision boundary of this optimal non-private model.

settings [5, 6, 7]. Predictive multiplicity can appear due to under-specification and randomness in the model's training procedure [8].

Predictive multiplicity formalizes the *arbitrariness* of decisions based on a model's output. In practice, predictive multiplicity can lead to questions such as "*Why has a model issued a negative decision on an individual's loan application if other models with indistinguishable accuracy would have issued a positive decision?*" or "*Why has a model suggested a high dose of a medicine for an individual if other models with comparable accuracy would have prescribed a lower dose?*" These examples highlight that acting on predictions of a single model without regard for predictive multiplicity can result in arbitrary decisions. Models produced by training algorithms that exhibit high predictive multiplicity face fundamental challenges to their credibility and justifiability in high-stakes settings [9, 8].

In this paper, we demonstrate a fundamental connection between privacy and predictive multiplicity: For a fixed training dataset and model class, DP training results in models that ensure the same degree of privacy and achieve comparable accuracy, yet assign conflicting outputs to individual inputs. DP training produces conflicting models even when non-private training results in a single optimal model. Thus, in addition to decreased accuracy, DP-ensuring training methods also incur an arbitrariness cost by exacerbating predictive multiplicity. We show that the degree of predictive multiplicity varies wildly across individual inputs and can disproportionately impact certain population groups. Fig. 1 illustrates the predictive-multiplicity cost of DP training in a simple synthetic example.

In summary, the level of privacy in DP training significantly impacts the level of predictive multiplicity. This, in turn, means that decisions supported by differentially-private models can have an increased level of arbitrariness: a given decision would have been different had we used a different random seed in training, even when all other aspects of training are kept fixed and the optimal non-private model is unique. Before deploying DP-ensuring models in high-stakes situations, we suggest that practitioners quantify the predictive multiplicity of these models over salient populations and—if possible to do so without violating privacy—measure predictive multiplicity of individual decisions during model operation. Such audits can help practitioners evaluate whether the increase in privacy threatens the justifiability of decisions, choose whether to enact a decision based on a model's output, and determine whether to deploy a model in the first place.

References

[1] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of cryptography conference, Springer, 2006.

[2] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy., Found. Trends Theor. Comput. Sci. (2014).

[3] K. Chaudhuri, C. Monteleoni, A. D. Sarwate, Differentially private empirical risk minimization., Journal of Machine Learning Research 12 (2011).

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

[5] C. Marx, F. Calmon, B. Ustun, Predictive multiplicity in classification, in: International Conference on Machine Learning, PMLR, 2020, pp. 6765–6774.

[6] H. Hsu, F. d. P. Calmon, Rashomon capacity: A metric for predictive multiplicity in probabilistic classification, Advances in Neural Information Processing Systems (2022).

[7] J. Watson-Daniels, D. C. Parkes, B. Ustun, Predictive multiplicity in probabilistic classification, in: AAAI, 2023.

[8] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., Underspecification presents challenges for credibility in modern machine learning, Journal of Machine Learning Research (2020).

[9] E. Black, M. Raghavan, S. Barocas, Model multiplicity: Opportunities, concerns, and solutions, in: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2022.