Author Response for Submission 8969: Differential Privacy Has Disparate Impact on Model Accuracy

Related work. Thanks for the pointers to recently released papers, we will acknowledge them. Yeom et al. [5] show that models with poor generalization are more vulnerable to inference attacks. They measure how DP bounds the leakage of information about training data, not its impact on model accuracy. They do not at all study (1) the accuracy of models on subgroups, nor (2) how accuracy on subgroups changes as a result of applying DP-SGD. (2) is our main result, which is completely independent and orthogonal to the analysis in Yeom et al.

We did not have room to discuss [3, 2] but the brief summary is they provide evidence that DP may be combined with fairness, but do not give algorithms that could be used to train practical DP neural networks.

Rényi differential privacy. We use Rényi DP only to estimate privacy loss. This does not change the DPSGD algorithm of Abadi et al. but rather provides tighter bounds on privacy loss [4], allowing to reduce the amount of added noise. The TF Privacy tool enables estimation of epsilon given the input parameters (dataset size, number of epochs, batch size, noise, delta) before starting the training, thus this computation is not part of Algorithm 1. We ensure that our training uses the same hyperparameters as used to estimate epsilon.

General statements about DP. We will clarify in the abstract and intro that our results apply to DPSGD, a popular way to train DP neural networks, and not necessarily to DP as a general concept.

Experiment details. Thanks for the comments about improving presentation (captions and trend lines). We used the UTK dataset as an additional source of darker-skinned faces because in the DiF dataset, some individuals with lighter skin were labeled as dark-skinned. We set the ratio between lighter- and darker-skinned individuals to measure the effect of DPSGD on underrepresented classes, not to reflect the demographic balance of any country or real-world dataset.

Size of the groups and complex classes. More items per class is indeed usually helpful. That said, our federated learning study shows that participants with simpler vocabularies get better accuracy with DPSGD, whereas participants with complex vocabularies contribute less to the model (Figure 3b). This is an example of how DPSGD negatively affects more complex data.

Impact of clipping. Clipping alone is responsible for slowing down the learning, similar to decreasing the learning rate. Without adding noise, both well- and under-represented classes converge to the same accuracy but much slower. Noise, however, prevents the model from converging to the same norm. We find this presentation to be more intuitive and perhaps a good starting point for future research on combining differential privacy with fairness.

Adversarial training. Adversarial training for fairness [1] overweights the loss for underrepresented groups. Sensitivity bounds imposed by DPSGD, DPGAN, and similar approaches hold only for specific loss functions and sampling strategies; if combined directly with adversarial training, the resulting models will not be DP. It is an open problem how to combine DP with censoring techniques such as adversarial training.

Training models with the same epsilon. Modifying the hyperparameters directly involved in estimating epsilon results in a big variance of results. Using TF Privacy, we observed that among all hyperparameters, the noise multiplier z has the highest impact on epsilon. Changing hyperparameters that do not affect privacy loss, such as the learning rate, model architecture, or optimizer, impacts the accuracy but does not affect fairness, thus we omitted these analyses due to lack of space.

Fairness measure. Equalized odds gives us the most direct way to measure the impact of DPSGD on a popular fairness measure. Equal opportunity requires equality on the "advantaged" outcome, but in the multi-label tasks in our experiments it is not always clear what outcome should be considered advantaged. Accuracy on each subgroup, on the other hand, is straightforward to measure.

References

[1] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv:1707.00075*, 2017.

[2] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the compatibility of privacy and fairness. `http://pwp.gatech.edu/rachel-cummings/wp-content/uploads/sites/679/2019/03/FairPrivate.pdf`, 2019.

[3] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially private fair learning. *arXiv:1812.02696*, 2018.

[4] I. Mironov. Rényi differential privacy. In *CSF*, 2017.

[5] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*, 2018.