Poster: DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?

Archit Uniyal[1, 2, 3], Rakshit Naidu[2, 4], Sasikanth Kotti[2, 5], Sahib Singh[2, 6]
Patrik Joslin Kenfack[2, 7], Fatemehsadat Mireshghallah[2, 8], Andrew Trask[2, 9]
[1] Panjab University, [2] Openmined, [3] University of Virginia
[4] Carnegie Mellon University, [5] IIT Jodhpur, [6] Ford Motor Company
[7] Innopolis University, [8] University of California, San Diego
[9] University of Oxford

***Abstract*—Recent advances in differentially private deep learning have demonstrated that application of differential privacy, specifically the DP-SGD algorithm, has a disparate impact on different sub-groups in the population, which leads to a significantly high drop-in model utility for sub-populations that are under-represented (minorities), compared to well-represented ones. In this work, we aim to compare PATE, another mechanism for training deep learning models using differential privacy, with DP-SGD in terms of fairness. We show that PATE does have a disparate impact too, however, it is much less severe than DP-SGD. We draw insights from this observation on what might be promising directions in achieving better fairness-privacy trade-offs.**

Index terms - Fairness, Differential Privacy.

I. Introduction and Approach

Most of the datasets being used to perform experiments nowadays are quite well balanced in terms of distribution across the various classes present quantitatively, but in real world scenario this ain't necessary. Most of the times there are underrepresented groups present which has a disparate impact on model accuracy, especially when privacy preserving algorithms like DP-SGD and PATE are applied. In this paper, we analyse and perform experiments to measure the fairness and accuracy of these algorithms.

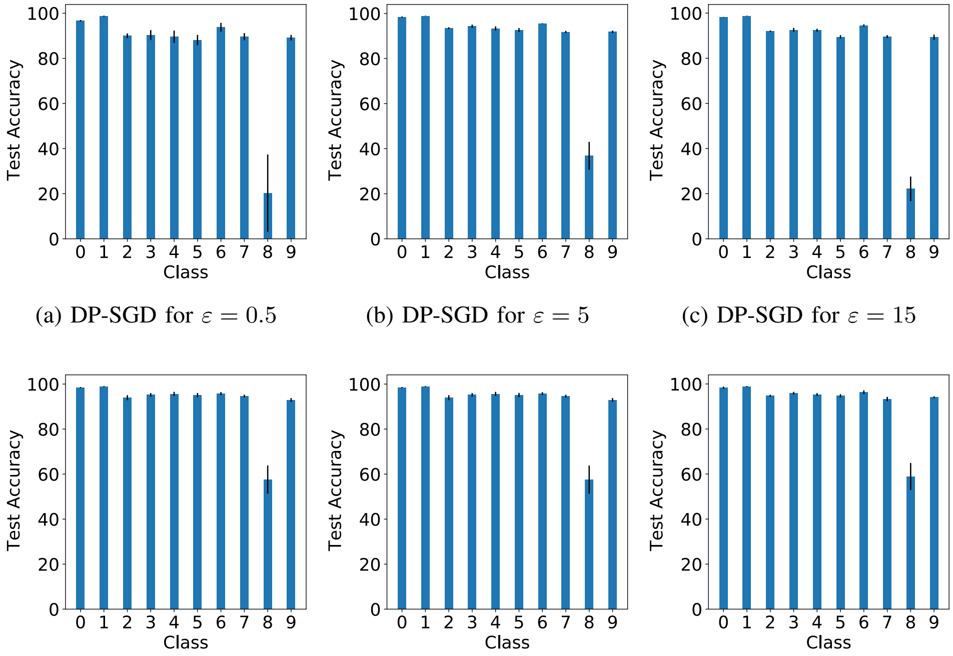
We have taken into consideration two datasets - MNIST and SVHN. For both these datsets, we have induced an imbalance in class '8'. In MNIST, the dataset is imbalanced in a 1:10 ratio i.e for every one image in class '8', there are 10 images in each of the other classes. Furthermore, SVHN dataset is quite imbalanced in itself. Therefore, we first balance the dataset to 5000 images in each class followed by imbalancing the class '8' to a 1:2 ratio i.e for every one image in class '8' there are two images in each class. This imbalance helps us portray the underrepresented groups in real world datasets where data across some groups is limited.

We device experiments around these datasets at different levels of privacy which we denote through a privacy budget parameter called epsilon ϵ. We train a 4-layer deep CNN for $\epsilon = 0.5, 5, 15$ on the imbalanced MNIST dataset. The reason for using such a small CNN for MNIST is that the dataset contains grayscale images and performing feature extraction on such images is quite easier in comparison to RGB images. For SVHN dataset, we utilized the ResNet-18 architecture. We use learning rates 0.01 and 0.05 for DP-SGD and PATE, respectively. We train PATE with 50 teachers and 30 students. We train each model 5 times and report the mean and standard deviation of accuracy in our experiments.

As shown in figure 1, in our experiments on imbalanced MNIST dataset we observe that the deviation of the test accuracy for the imbalanced class 8 decreased with the increase in the values of ϵ in DP-SGD. We also notice that over different values of ϵ, PATE exhibits "stable" results for the test accuracy of the imbalanced class. In PATE, the accuracy of the imbalanced class is almost (more than) twice the accuracy obtained by DP-SGD for the same ϵ value.

Furthermore, on the imbalanced SVHN dataset for $\epsilon = 5, 8$ (figure 2), we observe that both DP-SGD and PATE produce similar results with the averaged accuracy for the imbalanced class being less than 5%. PATE showcased more robust results than DP-SGD which exhibits higher standard deviation at different values of ϵ.

Through our experiments and ablation study, we are able to summarize our work in three key observations. Firstly, PATE and DP-SGD both have disparate impact on the under-represented groups but PATE has signifi-



(a) DP-SGD for $\varepsilon = 0.5$

(b) DP-SGD for $\varepsilon = 5$

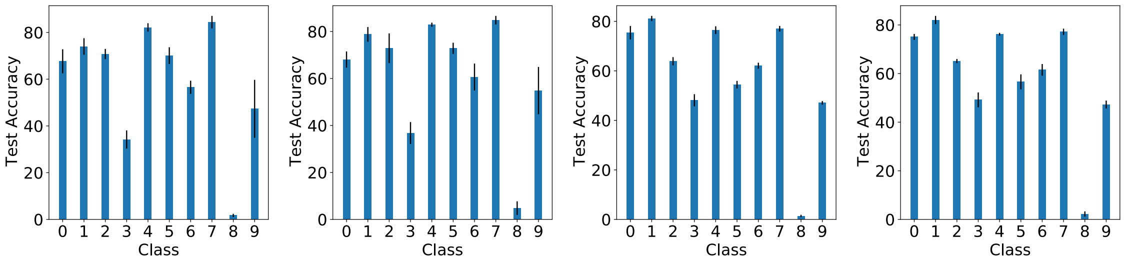
(c) DP-SGD for $\varepsilon = 15$

(d) PATE for $\varepsilon = 0.5$

(e) PATE for $\varepsilon = 5$

(f) PATE for $\varepsilon = 15$

Fig. 1: Average test accuracy of each digit (class) for models trained on imbalanced MNIST data, where samples of class “8” are decreased to 0.1 their original count.



(a) DP-SGD for $\varepsilon = 5$

(b) DP-SGD for $\varepsilon = 8$

(c) PATE for $\varepsilon = 5$

(d) PATE for $\varepsilon = 8$

Fig. 2: Average test accuracy of each digit (class) for models trained on imbalanced SVHN data, where samples of class “8” are decreased to half their original count.

cantly less disproportionate impact on the utility compared to DP-SGD. Secondly, we note that the standard deviation of the accuracy for each class over 5 runs was much lower in PATE compared to DP-SGD. Lastly, by experimenting with various teacher counts, we observe that having multiple teachers often provides a higher utility than a single teacher for underrepresented groups. However beyond the tipping point of this ensemble (10 teachers in our case), the utility stagnates and then starts dropping significantly. It is also worth noting since PATE is a semi-supervised approach, therefore the availability of public data is necessary for training without which PATE is not applicable.