Fairness under Privacy Constraints

Cuong Tran

Department of Computer Science
Syracuse University

May 1, 2020

Introduction

Fairness

- A machine learning model is considered fair if its outcome is independent w.r.t certain sensitive attributes like races, gender, ages,..
- Motivation of fairness is from anti-discrimination laws.

Privacy

- A public learning model can reveal private information of the training data the model is based on.
- We can quantify the level of privacy (or privacy loss) by the framework of differential privacy [1].

[1] C DWork, The Algorithmic Foundations of Differential Privacy, 2014

Relationship b.w Fairness vs Privacy

- R. Cummings showed that there are scenarios of data where we can not exact perfect fairness under privacy constraints. [2]
- C. Dwork indicated that individual fairness is a generalization notation of differential privacy(DP).[3]
- In 2019, E. Bagdasaryan argued that differential privacy under DP-SGD has negative impact toward model's fairness. [4]
- These works posed an open question. Under which conditions, fairness is at odds or align with privacy ?

[2] R. Cummings,et al, On the Compatibility of Privacy and Fairness, 2019

[3] C. Dwork, et al, Fairness Through Awareness, 2011

[4] E. Bagdasaryan, et al, Differential Privacy Has Disparate Impact on Model Accuracy, 2019

What's Class Project about ?

- We investigate under which data conditions, DP might have negative impact towards the fairness ?
- We initially hypothesized two conditions: (1) data distributions between minority and majority groups are different (2) the underlying mapping functions from input-to-output space across groups are not similar.
- We provide a new variant of DP, called sensitive feature level DP which provides privacy guarantees at individual feature level.
- We test our hypothesis on real datasets to affirm our claims.

Why DP has negative impact towards fairness ?

- In NIPS' 19, E. Bagdasaryan blamed the negative impact of DP due to the clipping norm in DP-SGD algorithm.
- In that work, authors argued that when the minority groups exhibit complex behaviours, during training the SGD algorithms will produce larger gradients. However, under gradient norm clipping in DP-SGD, the model's ability to explain minority group is reduced.
- Their claim was confronted by the recent work D. Xu, *Removing Disparate Impact of Differentially Private Stochastic Gradient Descent on Model Accuracy*, in 2020.

Possible Solutions to Increase Fairness

- We accept that unfairness(or bias) might happens under privacy constraints. To alleviate that issue, we propose to add the fairness constraints:

$$\min_{\theta} \sum_{i} \mathcal{L}(x_i, y_i; \theta)$$

$$\texttt{s.t}: \ \frac{1}{|Z_0|} \sum_{i \in Z_0} \mathcal{L}(x_i, y_i; \theta) = \frac{1}{|Z_1|} \sum_{i \in Z_1} \mathcal{L}(x_i, y_i; \theta)$$

- We will apply the Dual Lagrangian method to solve that constrained problem.
- During model's training, we apply DP-SGD mechanism to protect data's privacy.

Experiments

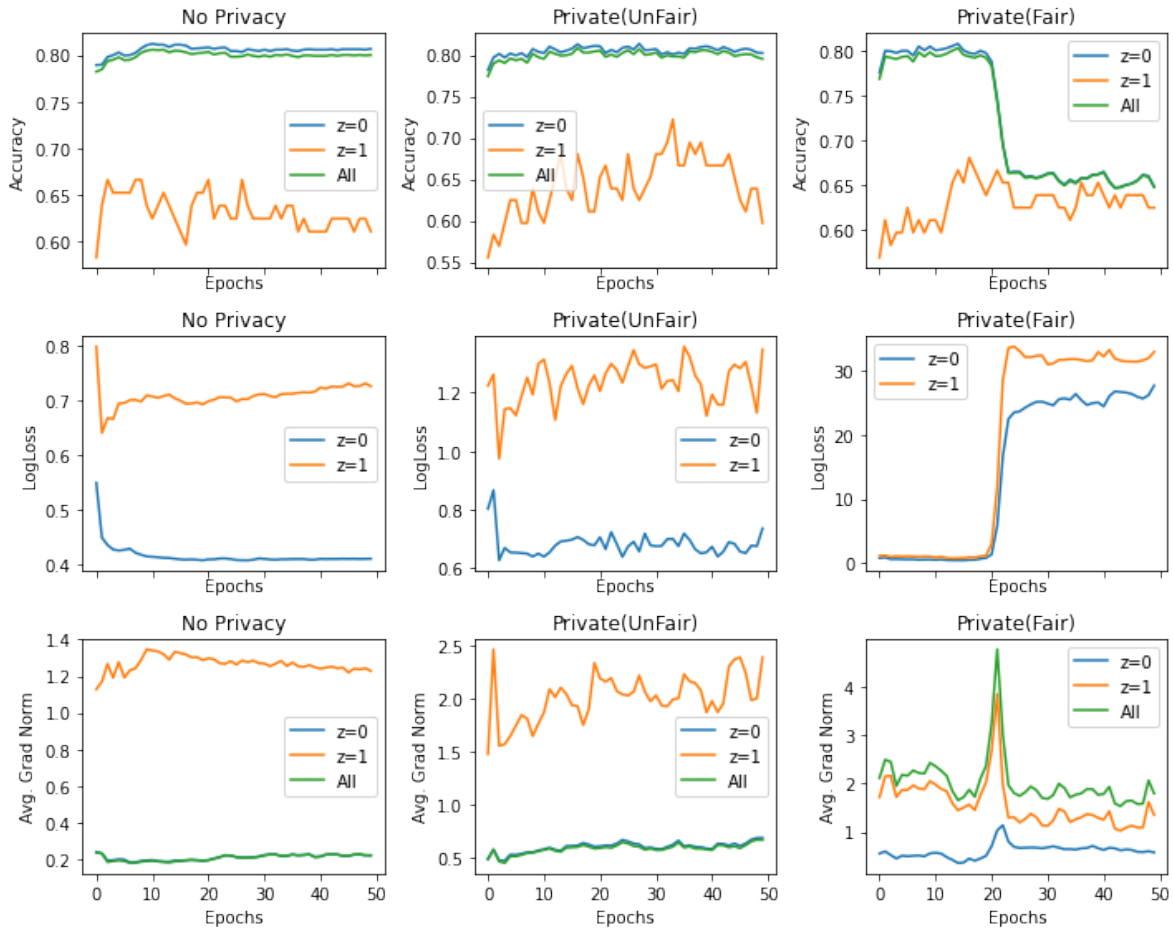
- Datasets. We consider Bank, Income, and Default which are considered as benchmark datasets in research about fairness.
- Selection Bias. For each dataset, we proceed the following steps to exacerbate data bias.
  - Pick the most correlated feature with the label. Compute its median.
  - For the minority group, like female in Income data, select only samples that has feature smaller than that median. We keep the original samples for majority group.

Experimental Settings

- We compare three models, (a) a simple classifier, (b) a private classifier (c) and a private(fair) classifier.
- For the privacy constraints, we fixed the gradient bound $C = 1$, and noise multiplier $\sigma = 1$ for two private models above.

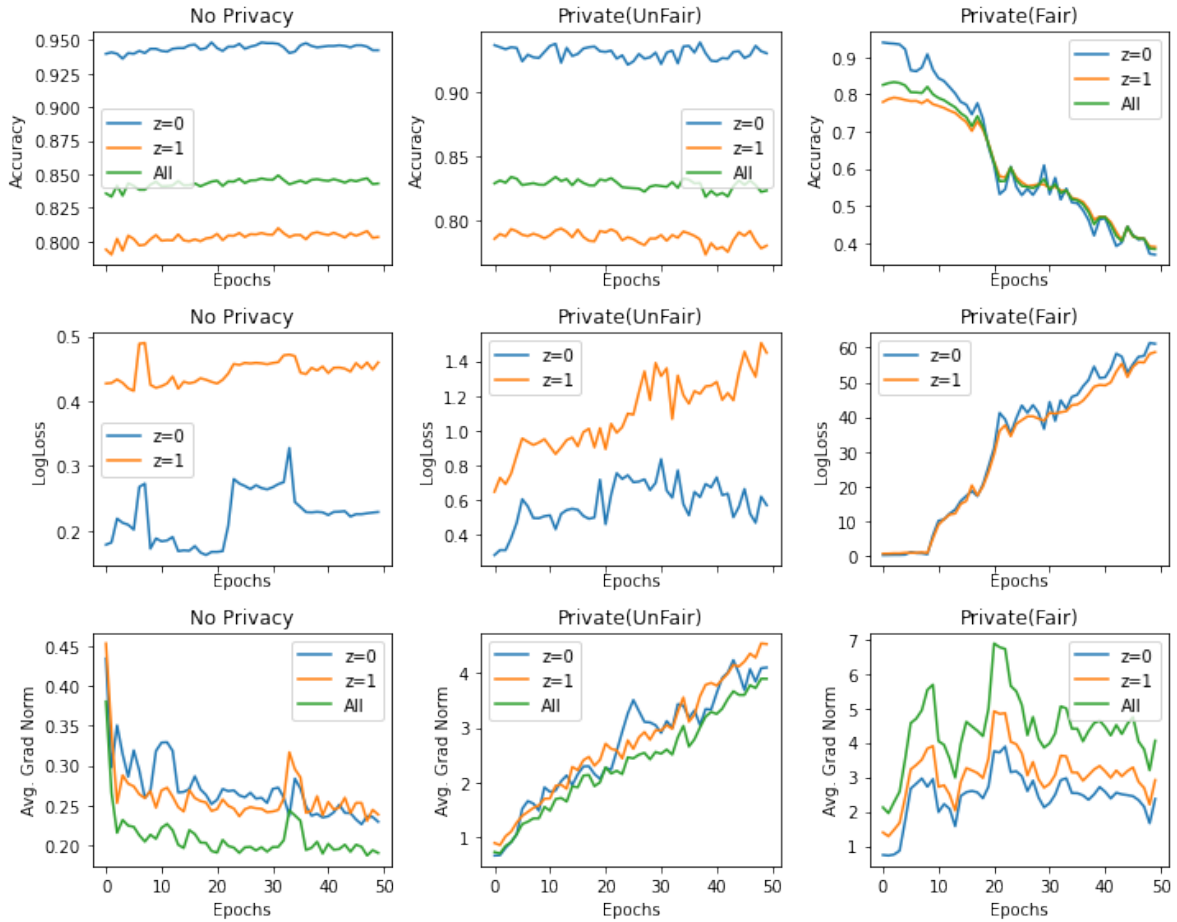
Bank dataset

We observed a surprising result on Bank data, when DP in fact has positive effect towards fairness.

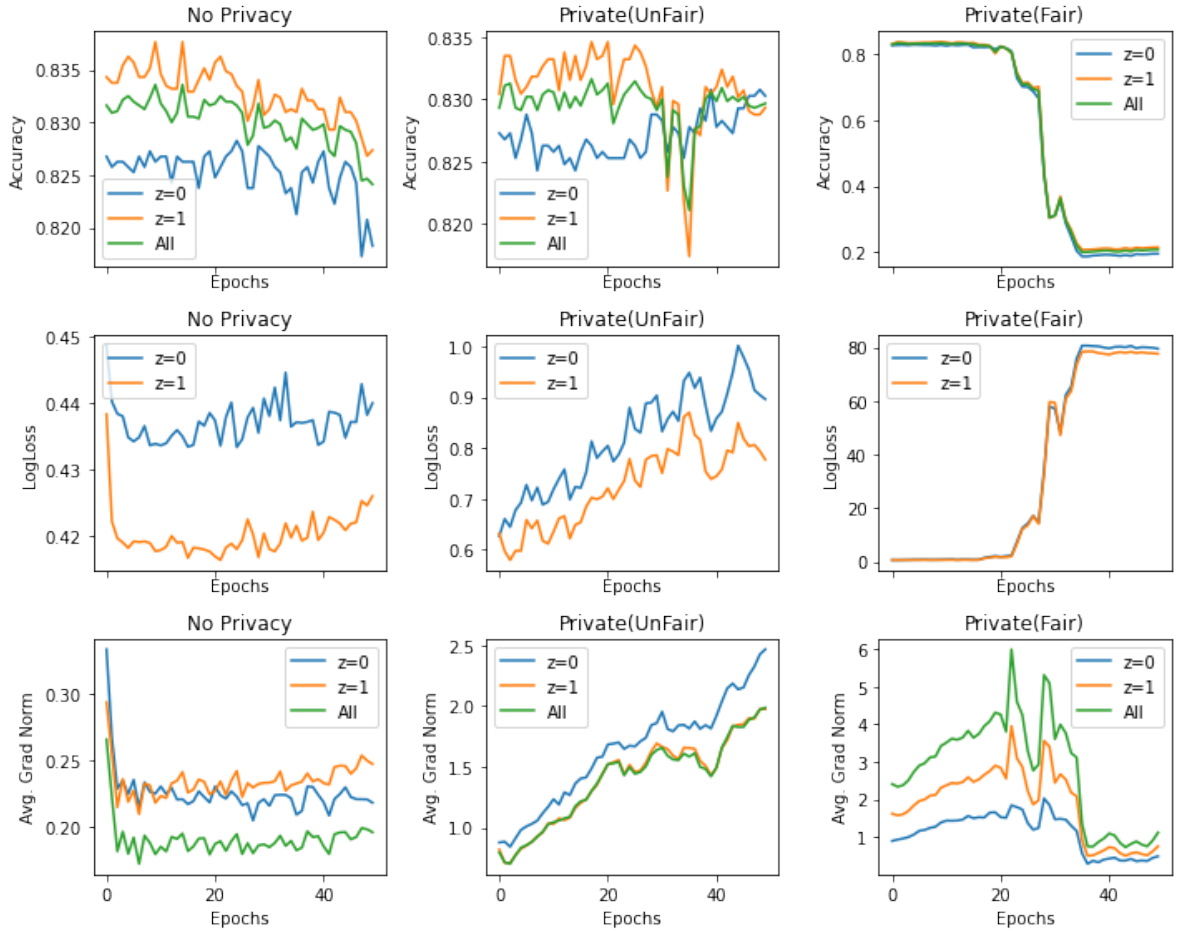


Income dataset

We see that (1) DP does not have negative impact towards fairness here. (2) The Dual Lagrangian method effectively reduced the accuracy gap between two groups.



Default dataset



Conclusions and Future Work

- Negative impact of DP towards fairness depends on many factors rather than only complex behaviours of minority groups. It can depends on privacy mechanism, model's hyper-parameters, privacy loss constraints.
- Dual Lagrangian method with equalized odd fairness constraints can effectively increase the model's fairness.