CY 7790, Lecture 21: Privacy risks in ML. Membership Inference

Pablo Kvitca, Zohair Shafi

December 8, 2021

The topic of the class today is the relationship between privacy and fairness in Machine Learning models. We look at the research on the effect differential-privacy has on fairness, the effect fairness constrains have on a model's vulnerability to privacy attacks (specifically membership inference). Finally we observe there is a, seemingly, necessary trade-off between fairness and privacy for getting usable (good accuracy) models.

1 Chang and Shokri. On the Privacy Risks of Algorithmic Fairness. In Euro SP 2021

Problem Statement This paper analysis both privacy (through membership inference) and group fairness concerns. It presents a framework for analyzing the privacy risks of group fairness algorithms for machine learning, through empirical analysis.

The work comes from the intuition that fair algorithms tend to memorize more data for the under-represented subgroups than others, since they aim to equalize the model's error across different groups. This memorization leads to an increased information leakage by the model for unprivileged groups (vs the privileged groups).

Examples of Fairness needs in ML

1. COMPAS dataset for recidivism
2. Amazon Recruiting tool
3. Apple Card Investigated After Gender Discrimination Complaints

Managing Bias Proposal There is a 2021 proposal by NIST on ”Identifying and Managing Bias within Artificial Intelligence”, presented standards and techniques for bias around AI.

Bias in ML The bias in an ML model can come from the training dataset (data bias, which reflects human bias) and from the algorithm (algorithmic bias, learned to minimize the overall loss).

Introduction to Fairness The process of data-driven Machine Learning can inherit bias from their datasets, which is problematic when the decisions of made by the algorithm (or using their output) can be discriminatory or harmful. There are two main types of harm: allocation harm (when an AI system extends/withholds opportunities, resources, or information) applies to tasks like hiring, admissions, lending; and quality-of-service harm (when a system does not work as well for one person as it does for another, even if no opportunities, resources, or information are extended/withheld) applies to accuracy on face recognition, document search, product recommendation, and others.

Main concepts for Fairness and Bias in ML

1. Group Fairness
2. Sensitive Features
3. Parity Constrains: including measures like *Demographic Parity*, *Equalized Odds*, *Equal Opportunity*, and others

Typical Steps for Approaching Fairness in Practice

1. Assessment (compute various metrics)

   (a) **Statistical Parity Difference**: the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group

   (b) **Equal Opportunity Difference**: the difference of true positive rates between the unprivileged and the privileged groups

   (c) **Average Odds Difference**: the average difference of false positive rate and true positive rate between unprivileged and privileged groups

   (d) **Disparate Impact**: the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group

   (e) **Theil Index**: measures the inequality in benefit allocation for individuals

   (f) **Distance**: the average Euclidean/Mahanalobis/Manhattan distance between the samples from two datasets

   (g) Others (see AIF360)

2. Mitigation

   (a) *Optimized Pre-processing*, *Reweighing*, *Adversarial Debiasing*, *Reject Option Classification*, *Disparate Impact Remover*, *Learning Fair Representations*, *Prejudice Remover*, *Calibrated Equalized Odds Post-processing*, *Equalized Odds Post-processing*, *Meta Fair Classifier*, others (see AIF360)

Fairness AND Privacy Fair ML aims at minimizing discrimination against protected groups by applying some measurement of fairness. For example, by imposing a constraint on learned models to equalize their behaviour across different groups. This has a disproportionate effect on the influence different training data points have on the (fair) model; which leads to the information leakage of the model about its training data. This paper analyzes the privacy risk of applying a group fairness constraint, equalized odds through the lens of membership inference privacy attacks. The question: ”Is there a privacy cost to fairness?”. This is done through analyzing models trained with differential privacy (DP) and fairness constraints. The privacy risk is formalized as ”the success of membership inference attacks against the ML models”

Definition: Membership Inference Attack This is a type of attack on an ML model that aims to infer whether an individual’s data is in the training dataset of the model or not.

Definitions

Definition 1 (δ - Equalized Odds Fairness). A classifier M satisfies δ-Equalized Odds with respective to the protected attribute $\mathcal{G}$, if for all $g, g' \in \mathcal{G}$, the false positive rate and false negative rate of the classifier for group $\{G = g\}$ and $\{G = g'\}$ are within δ range of one another.

$$\Delta(M, \mathcal{D}) \triangleq \max_{\substack{y \in \{-,+\} \\ g,g' \in \mathcal{G}}} \left| \Pr_{\mathcal{D}}[M(X) \neq y | S = g, Y = y] - \Pr_{\mathcal{D}}[M(X) \neq y | S = g', Y = y] \right| \leq \delta, \tag{1}$$

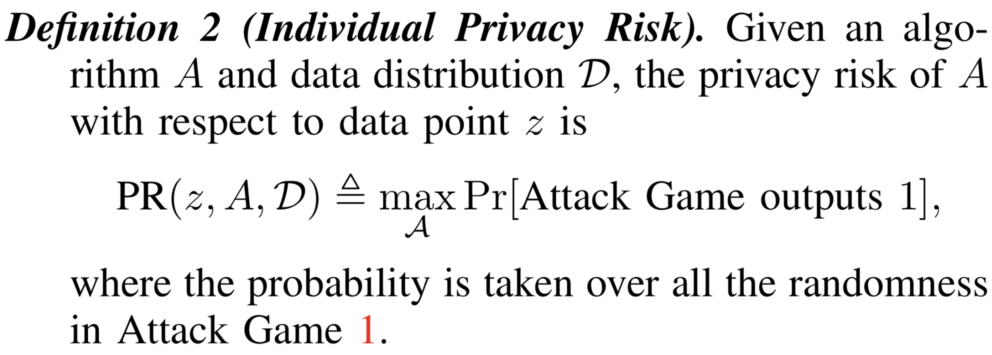
where the probabilities are computed over the data distribution $\mathcal{D}$. We refer to Δ as the model's **fairness gap** under equalized odds. A model satisfies exact fairness under equalized odds when $\delta = 0$.

Figure 1: Definition for delta equalized odds fairness

Attack Game 1 (Membership Inference).

1) Adversary chooses a data point z, and sends it to the challenger.
2) Challenger chooses a secret bit $b \leftarrow \{0, 1\}$ uniformly at random, and samples dataset $S \sim \mathcal{D}^n$. If $b = 1$, the challenger overwrites a random element in S with z.
3) Challenger runs algorithm A on S and sends its outputs A_S to the adversary.
4) Adversary runs an inference attack $\mathcal{A}$, and tries to infer the secret bit as $\hat{b} \in \{0, 1\}$.
5) The game outputs 1 (indicating that adversary wins) if $\hat{b} = b$, and 0 otherwise.

Figure 2: Attack Game 1

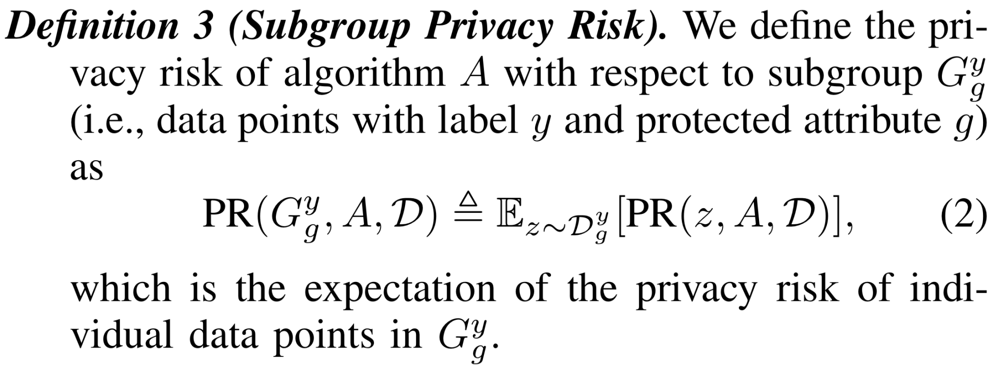


Definition 2 (Individual Privacy Risk). Given an algorithm A and data distribution $\mathcal{D}$, the privacy risk of A with respect to data point z is

$$\mathrm{PR}(z, A, \mathcal{D}) \triangleq \max_{\mathcal{A}} \Pr[\text{Attack Game outputs } 1],$$

where the probability is taken over all the randomness in Attack Game 1.

Figure 3: Individual Privacy Risk

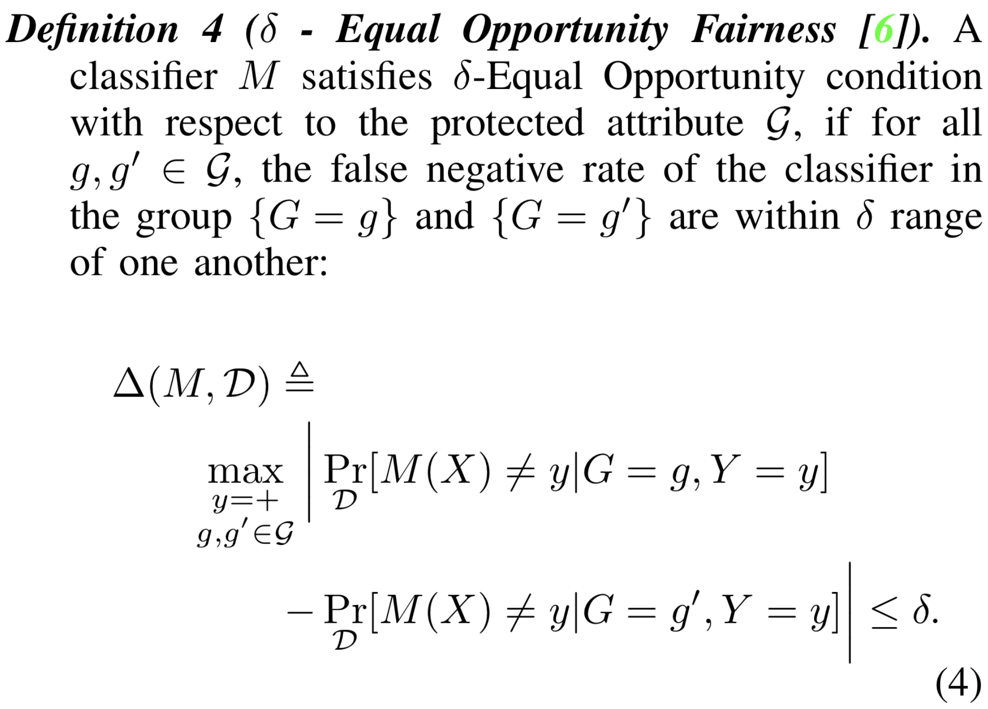


Definition 3 (Subgroup Privacy Risk). We define the privacy risk of algorithm A with respect to subgroup G_g^y (i.e., data points with label y and protected attribute g) as

$$\mathrm{PR}(G_g^y, A, \mathcal{D}) \triangleq \mathbb{E}_{z \sim \mathcal{D}_g^y}[\mathrm{PR}(z, A, \mathcal{D})], \qquad (2)$$

which is the expectation of the privacy risk of individual data points in G_g^y.

Figure 4: Subgroup Privacy Risk



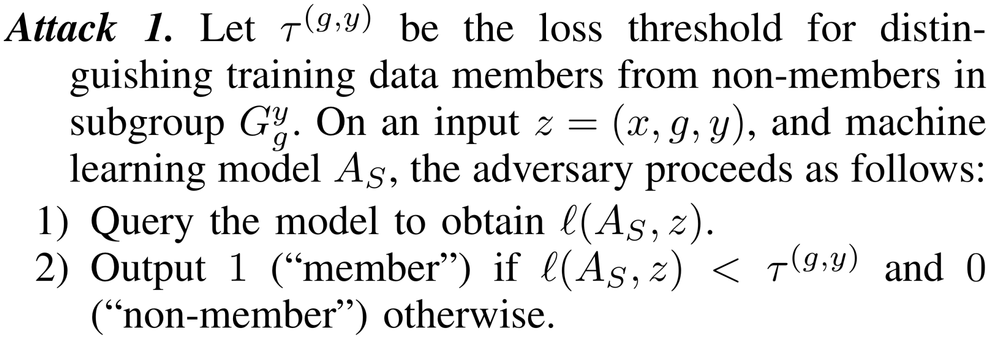
Definition 4 (δ - Equal Opportunity Fairness [6]). A classifier M satisfies δ-Equal Opportunity condition with respect to the protected attribute $\mathcal{G}$, if for all $g, g' \in \mathcal{G}$, the false negative rate of the classifier in the group $\{G = g\}$ and $\{G = g'\}$ are within δ range of one another:

$$\begin{aligned} &\Delta(M, \mathcal{D}) \triangleq \\ &\quad \max_{\substack{y=+ \\ g,g' \in \mathcal{G}}} \Big| \Pr_{\mathcal{D}}[M(X) \neq y | G = g, Y = y] \\ &\qquad - \Pr_{\mathcal{D}}[M(X) \neq y | G = g', Y = y] \Big| \leq \delta. \end{aligned} \tag{4}$$

Figure 5: Delta Equal Opportunity Fairness

Quantifying Privacy Risk

1. The adversary has black-box access to the model
2. The adversary can compute the loss of the model on any input
3. A simple attack model: compare the model's loss on an input with a threshold. The attack outputs ”member” if the loss if below a threshold, and ”nonmember' otherwise
4. The adversary can design separate membership inference attacks for each subgroups, making them stronger
5. The adversary can compute the loss threshold based on knowledge about the population (statistics) or through ”shadow models”
6. Intuition: the loss distribution would be different for each subgroups, specially underprivileged groups.



Attack 1. Let $\tau^{(g,y)}$ be the loss threshold for distinguishing training data members from non-members in subgroup G_g^y. On an input $z = (x, g, y)$, and machine learning model A_S, the adversary proceeds as follows:

1) Query the model to obtain $\ell(A_S, z)$.
2) Output 1 ("member") if $\ell(A_S, z) < \tau^{(g,y)}$ and 0 ("non-member") otherwise.

Figure 6: Attack 1

Experiments and Empirical Analysis The paper approaches the analysis empirically. First the create a synthetic dataset which is used to train one unconstrained model and three fair models (by applying equalized odds) on different delta levels (see definition 1)

Observations:

1. Points that are vulnerable to privacy attacks are mostly form the unprivileged subgroup (figure 2)
2. Members of the training set are more distinguishable form non-members on fair models compared to unfair models (figure 3)
3. Fair models memorize the points in the underprivileged subgroups (figure 4)
4. There is a correlation between accuracy gain and privacy cost (figure 5)
5. There is a large fairness gap versus the unconstrained model. Imposing fairness constrains results in a large privacy cost. (figure 8)
6. A smaller number of samples in the unprivileged subgroup results in a higher privacy cost for the unprivileged group (figure 9)
7. The post-processing algorithm doe snot improve the training accuracy over the unpriveleged subgroup
8. There are similar patterns when applying different group-fairness metrics (see table 4)

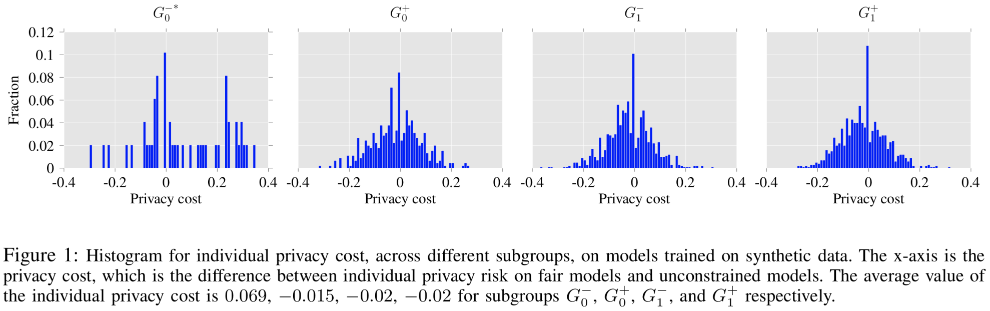


Figure 1: Histogram for individual privacy cost, across different subgroups, on models trained on synthetic data. The x-axis is the privacy cost, which is the difference between individual privacy risk on fair models and unconstrained models. The average value of the individual privacy cost is 0.069, −0.015, −0.02, −0.02 for subgroups G_0^-, G_0^+, G_1^-, and G_1^+ respectively.

Figure 7: Figure 1

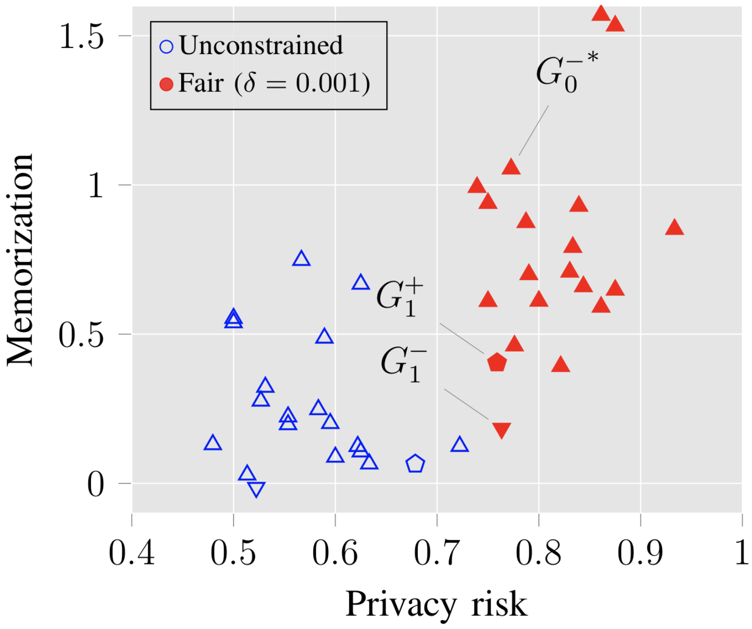


Figure 2: The most vulnerable points on fair models (trained on synthetic data). We find the top 20 vulnerable points that have the highest privacy risk on fair models. For each vulnerable point, we show its privacy risk and the memorization of models before (blue color) and after (red color) imposing fairness constraints. The marker shows which subgroup a point belongs to.

Figure 8: Figure 2

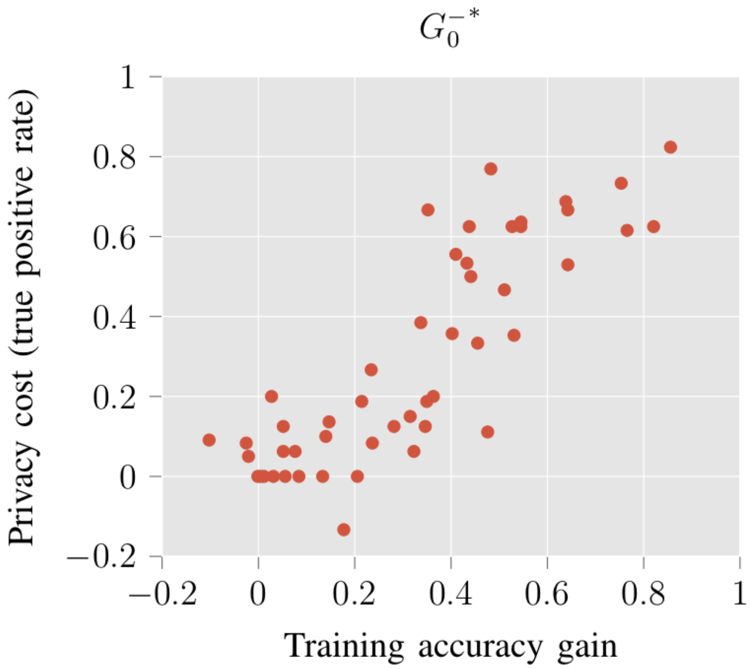


Figure 5: Accuracy gain versus privacy cost on the unprivileged subgroup G_0^-. Each point in the plot represents a data point in the training dataset. The training accuracy gain is the difference in the training accuracy between fair and unconstrained models. The y-axis is the difference in the attacker's true positive rate between fair and unconstrained models on each training point.

Figure 9: Figure 5

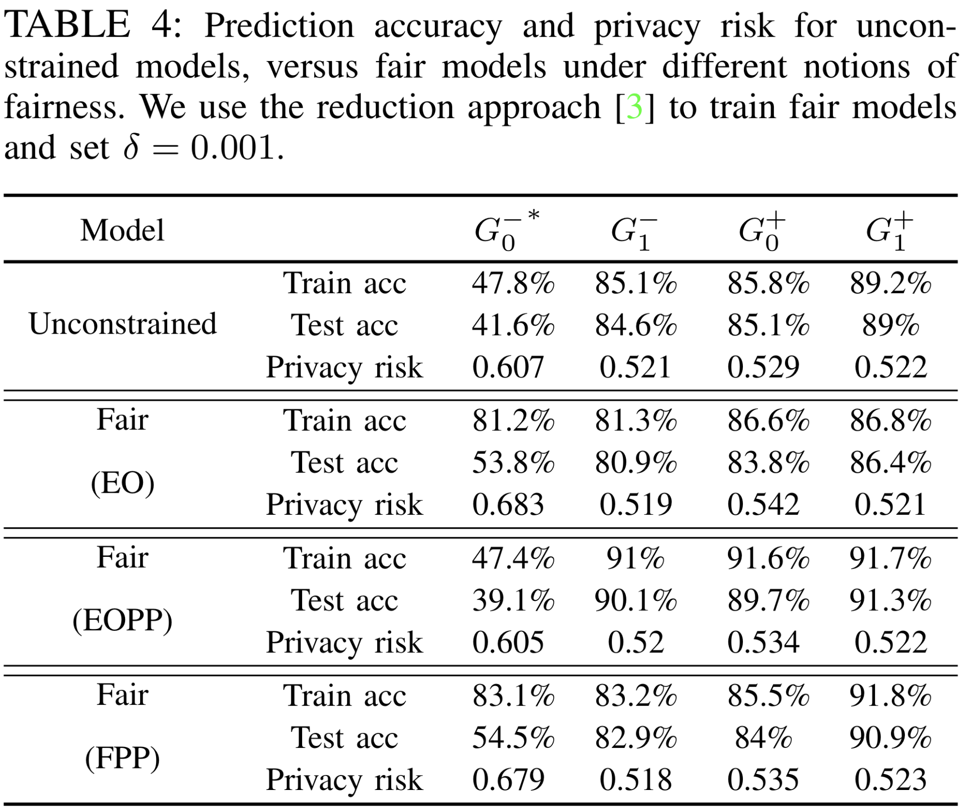


TABLE 4: Prediction accuracy and privacy risk for unconstrained models, versus fair models under different notions of fairness. We use the reduction approach [3] to train fair models and set $\delta = 0.001$.

Model		G_0^{-*}	G_1^-	G_0^+	G_1^+
Unconstrained	Train acc	47.8%	85.1%	85.8%	89.2%
	Test acc	41.6%	84.6%	85.1%	89%
	Privacy risk	0.607	0.521	0.529	0.522
Fair	Train acc	81.2%	81.3%	86.6%	86.8%
(EO)	Test acc	53.8%	80.9%	83.8%	86.4%
	Privacy risk	0.683	0.519	0.542	0.521
Fair	Train acc	47.4%	91%	91.6%	91.7%
(EOPP)	Test acc	39.1%	90.1%	89.7%	91.3%
	Privacy risk	0.605	0.52	0.534	0.522
Fair	Train acc	83.1%	83.2%	85.5%	91.8%
(FPP)	Test acc	54.5%	82.9%	84%	90.9%
	Privacy risk	0.679	0.518	0.535	0.523

Figure 10: Table 4

Strengths

1. This is a practical approach to quantifying privacy risk and evaluating the trade offs between fairness and privacy constraints.
2. It builds upon prior work
3. Is simple to implement

Possible Improvements

1. Mitigation of discrimination and how to address the privacy risks is still an open area of research and not addressed.

Class Discussion

Better model? We cannot directly say a ”fair model is better”. Depending on the task/application, a privacy concern might apply.

Trade off There is a trade off between fairness and privacy, while maintaining usable accuracy.

Memorization could be good? In some cases of unbalanced datasets, there could be a need for memorization of the samples for smaller groups to achieve generalization (Does Learning Require Memorization? A Short Tale about a Long Tail, Vitaly Feldman)

Regularization This paper does not evaluate the impact of regularization on fairness and privacy. This could improve memorization. How does it affect fairness? How does it affect privacy?

Real Data Experiments The experiments with real data only use decision tree models, not neural networks.

Improvement through fairness constraints On the real data experiments, the fairness constraints seem to improve the model accuracy, which is uncommon

2 Bagdasaryan and Shmatikov. Differential Privacy Has Disparate Impact on Model Accuracy

2.1 Introduction

This paper speaks about the cost of privacy. Unlike most other papers that speak about the cost of privacy in terms of accuracy alone, this paper looks at the cost in terms of fairness and how DP training algorithms impact underrepresented groups adversely. For the purposes of measuring this disparate impact, they use **Accuracy Parity**, a weaker form of Equal Odds.

The paper consists mostly of empirical experiments and results. They measure performance across the following datasets :

- Gender and Age Classification on Facial Images
- Sentiment Analysis for Tweets
- Species Classification
- Federated Learning of Language Models
- MNIST (to observe training dynamics)

The paper contains detailed information about the setup of each experiment which will be left out for the purpose of this summary. We encourage the reader to look at the original paper for more details.

2.2 Results and Observations

Gender and Age Classification on Facial Data : Data was collected from the Flickr-based Diversity in Faces (DiF) dataset and the UTKFace dataset as a source of darker-skinned faces. A ResNet18 model pretrained on ImageNet was used.

The data was picked to ensure an imbalance between lighter and darker skinned people - 29500 images of people with ligher skin tones and 500 images of people with darker skin tones. A 5000 images long test set was considered with the same ratio.

Figure 11(a) shows how the disparity between the groups increases when going from a non-DP model to a DP model with varying ϵ values.

The same dataset is to train on small subgroups defined by the intersection of (age, gender, skin color) attributes. 60,000 images are randomly sampled from DiF and their accuracies are measured on each of the 72 intersections. Figure 11(b) shows that the DP model tends to be less accurate on the smaller subgroups. Figure 11(c) shows “the poor get poorer” effect: classes that have relatively lower accuracy in the non-DP model suffer the biggest drops in accuracy as a consequence of applying DP.

Sentiment Analysis of Tweets : This task involves classifying Twitter posts from the recently proposed corpus of African American English as positive or negative. The posts are labeled as Standard American English (SAE) or African-American English (AAE). 60,000 tweets labeled SAE and 1,000 labeled AAE are sampled, each subset split equally between positive and negative sentiments. A bidirectional 2 layer LSTM is used to train on the data - further details can be found in the full paper.

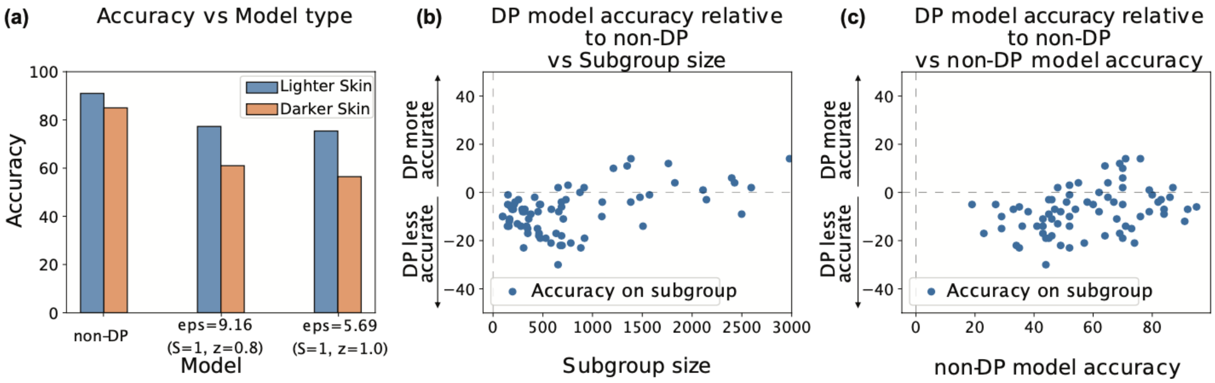


Figure 11: Gender and age classification on facial images.

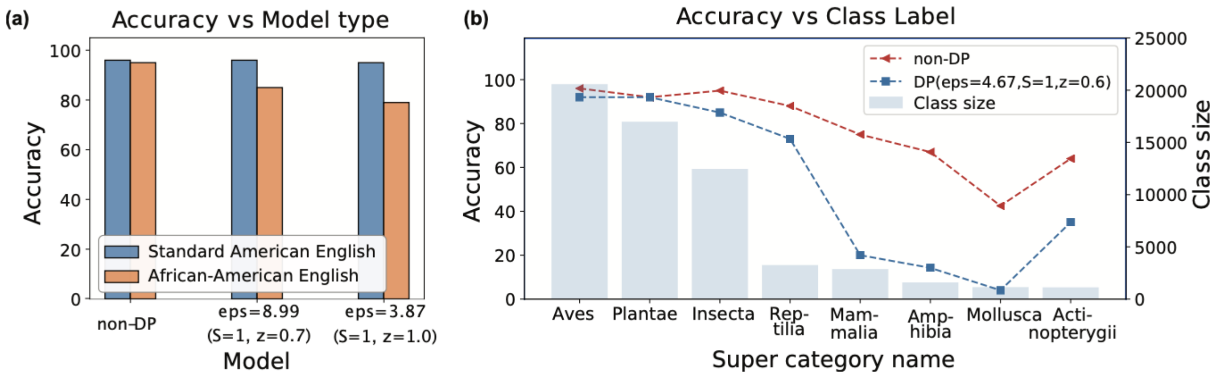


Figure 12: Sentiment analysis of tweets and species classification.

Figure 12(a) shows that all models learn the SAE subgroup almost perfectly, but on the AAE subgroup, accuracy of the DP models drops much more than the non-DP model.

Species Classification on Nature Images 60,000 images are sampled from the iNaturalist dataset of hierarchically labeled plants and animals in natural environments to predict the top level class. 8 out of 14 classes with low counts of images are dropped to make the task easier. An Inception V3 model is used to train on the data.

Figure 12(b) shows that the DP model almost matches the accuracy of the non-DP model on the well-represented classes but performs significantly worse on the smaller classes. Moreover, the accuracy drop doesn't depend only on the size of the class. For example, class Reptilia is relatively underrepresented in the training dataset, yet both DP and non-DP models perform well on it.

Federated Learning of a Language Model : Reddit data for the month of November 2017 is used for users with between 150 and 500 posts for a total of 80,000 users with 247 posts on average. The vocabulary is limited to 50k words with unpopular and rare words replaced by an unknown token. Every participant in the federated learning uses a two-layer, 10M-parameter LSTM. Interested readers are encouraged to look to the full paper for complete implementational details.

To illustrate the difference between trained models that have similar test accuracy, the diversity of the words the models output is measured. Figure 13(a) shows that all models have a limited vocabulary, but the vocabulary of the non-DP model is larger.

The accuracies of the models on participants whose vocabularies have different sizes is calculated. Figure 13(b)

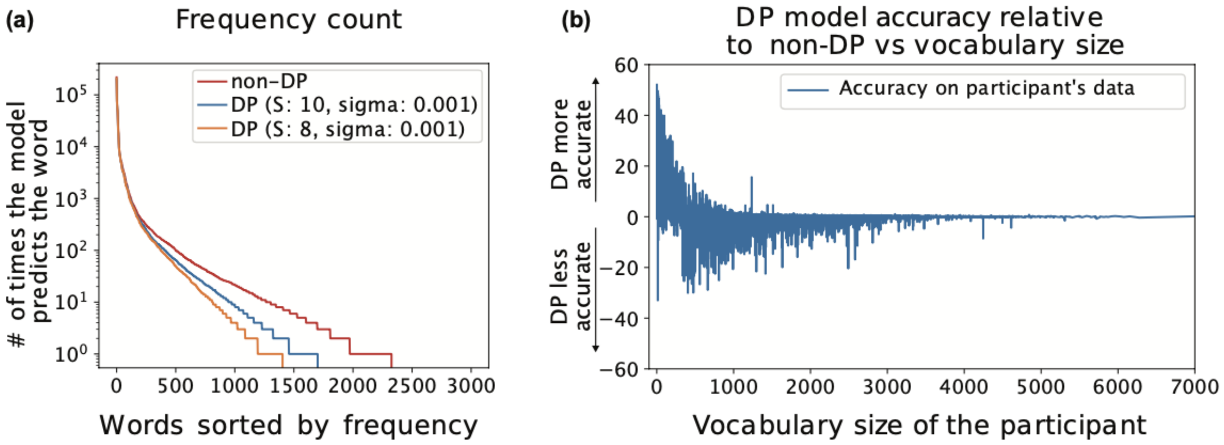


Figure 13: Federated learning of a language model.

shows that the DP model has worse accuracy than the non-DP model on participants with moderately sized vocabularies (500-1000 words) and similar accuracy on large vocabularies. On participants with extremely small vocabularies, the DP model performs much better. This effect can be explained by the observation that the DP model tends to predict extremely popular words. Participants who appear to have very limited vocabularies mostly use emojis and special symbols in their Reddit posts, and these symbols are replaced by ¡unk¿ during preprocessing. Therefore, their “words” become trivial to predict.

In federated learning, as in other scenarios, DP models tend to focus on the common part of the distribution, i.e., the most popular words. This effect can be explained by how clipping and noise addition act on the participants’ model updates. In the beginning, the global model predicts only the most popular words. Simple texts that contain only these words produce small update vectors that are not clipped and align with the updates from other, similar participants. This makes the update more “resistant” to noise and it has more impact on the global model. More complex texts produce larger updates that are clipped and significantly affected by noise and thus do not contribute much to the global model.

Hyperparameter Study on MNIST : MNIST was used to study the effects of different hyper parameters. Class ”8” was used as the underrepresented class and class ”2” as the majority class.

We see in 14 that the non-DP model (no clipping and no noise) converges to 97% accuracy on “8” vs. 99% accuracy on “2”. By contrast, the DP model achieves only 77% accuracy on “8” vs. 98% for “2”, exhibiting a disparate impact on the underrepresented class. We also see that the combination of Clipping and Noise in DP-SGD disproportionately impacts underrepresented classes.

To understand how the gradients of different classes behave, experiments are run without clipping or noise. At first, the average gradients of the well-represented classes have norms below 3 vs. 12 for the underrepresented class. After 10 epochs, the norms for all classes drop below 1 and the model converges to 97% accuracy for the underrepresented class and 99% for the rest. Next, experiments are run with clipped gradients but without adding noise. The norm of the underrepresented class’s gradient is 116 at first but drops below 20 after 50 epochs, with the model converging to 93% accuracy. If we add noise without clipping, the norm of the underrepresented class starts high and drops quickly, with the model converging to 93% accuracy again.

If, however, both clipping and noise are applied, the average gradients for all classes do not decrease as fast and stabilize at around half of their initial norms. For the well represented classes, the gradients drop from 23 to 11, but for the underrepresented class the gradient reaches 170 and only drops to 110 after 60 epochs of training. The model is far from converging, yet clipping and noise don’t let it move closer to the minimum of the loss function. Furthermore, the addition of noise whose magnitude is similar to the update vector prevents the clipped gradients of the underrepresented class from sufficiently updating the relevant parts of the model. This results in only a minor decrease in

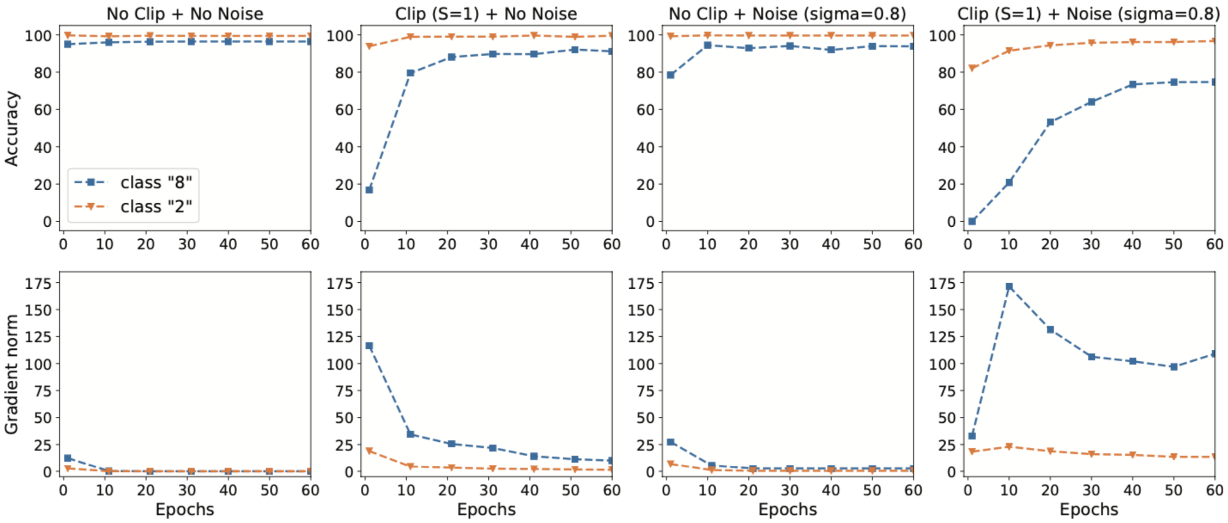


Figure 14: Effect of clipping and noise on MNIST training.

accuracy on the well-represented classes (from 99% to 98%) but accuracy on the underrepresented class drops from 93% to 77%. Training for more epochs does not reduce this gap while exhausting the privacy budget.

Figure 15 looks at the effects of various hyper parameters on training.

Noise multiplier z : This parameter enforces a ratio between the clipping bound S and noise σ: σ = zS. Figure 15(a) shows the accuracy of the model under different ϵ. Different values of S and σ are used that result in the same privacy loss. For example, large values of z require smaller S, otherwise the model is destroyed by noise, but smaller z allows an increase in S and obtains a more accurate model. In all cases, the accuracy gap between the underrepresented and well-represented classes is at least 20% for the DP model vs. under 3% for the non-DP model.

Batch size b : Larger batches mitigate the impact of noise. Figure 15(b) shows that increasing the batch size decreases the accuracy gap at the cost of increasing the privacy loss ϵ. Overall accuracy still drops.

Number of epochs T : Training a model for longer may produce higher accuracy at the cost of a higher privacy loss. Figure 15(c) shows, however, that longer training can still saturate the accuracy of the DP model without matching the accuracy of the non-DP model.

Size of the under represented class : In all preceding MNIST experiments, the classes were unbalanced with a 12 : 1 ratio, i.e., 500 images of class “8” vs. 6,000 images for the other classes. Figure 15(d) demonstrates that accuracy depends on the size of the underrepresented group for both DP and non-DP models.

From all these experiments, what we see is conclusive proof that DP disproportionately hurts underrepresented groups.

Discussion

Large models note that the models used for testing are large neural networks that might encode and overfit details of the dataset.

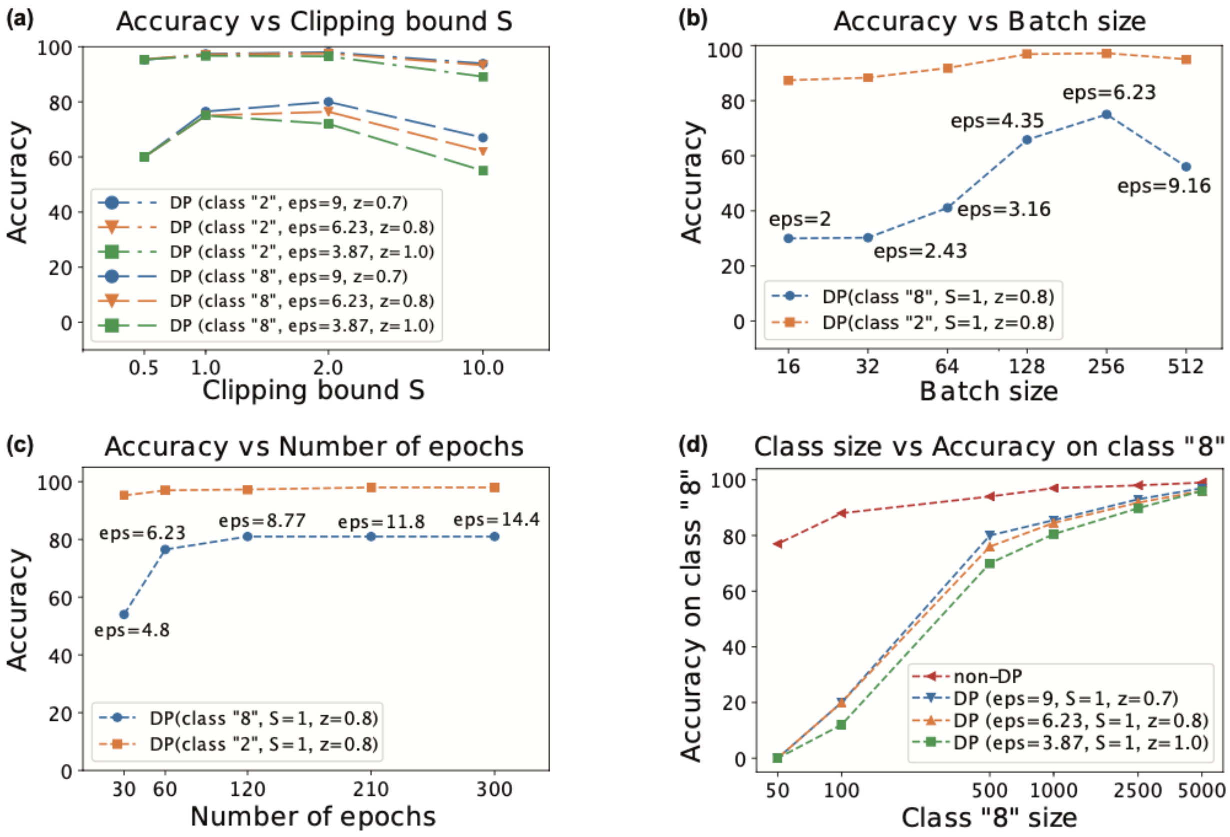


Figure 15: Effect of hyperparameters on MNIST training.