CY 7790

Special Topics in Security and Privacy: Machine Learning Security and Privacy
Fall 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science

December 2 2021

On the Privacy Risks of Algorithmic Fairness

CY7790
Paper review
Sri Krishnamurthy

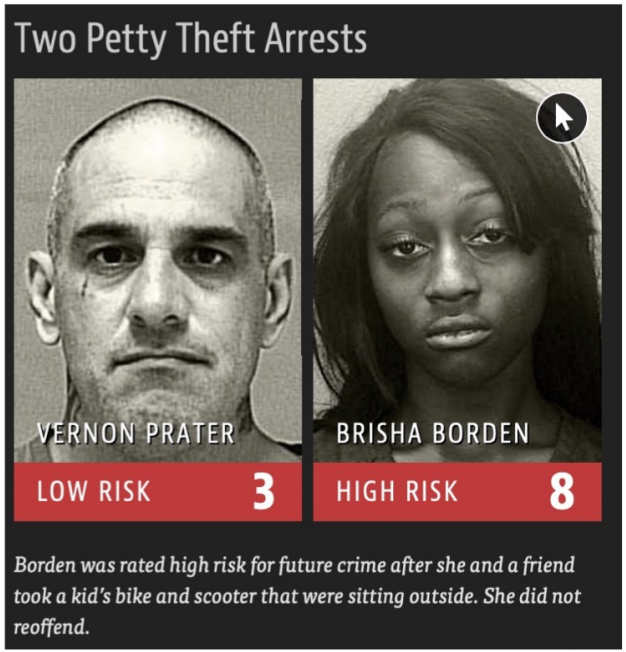
Paper summary:

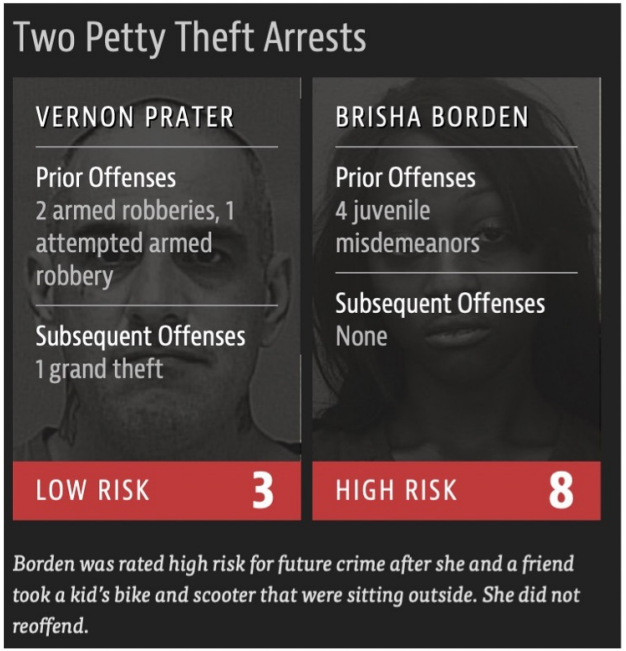
Privacy and Fairness

- Framework for analyzing privacy risks of group fairness algorithms for machine learning
- Fair algorithms tend to memorize data from the under-represented subgroups, while trying to equalize the model's error across groups
- Memorization leads to an increase in the model's information leakage about unprivileged groups

ML models are not neutral

- Recidivism prediction





from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

REUTERS

World Business Markets Breakingviews Video Mor

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

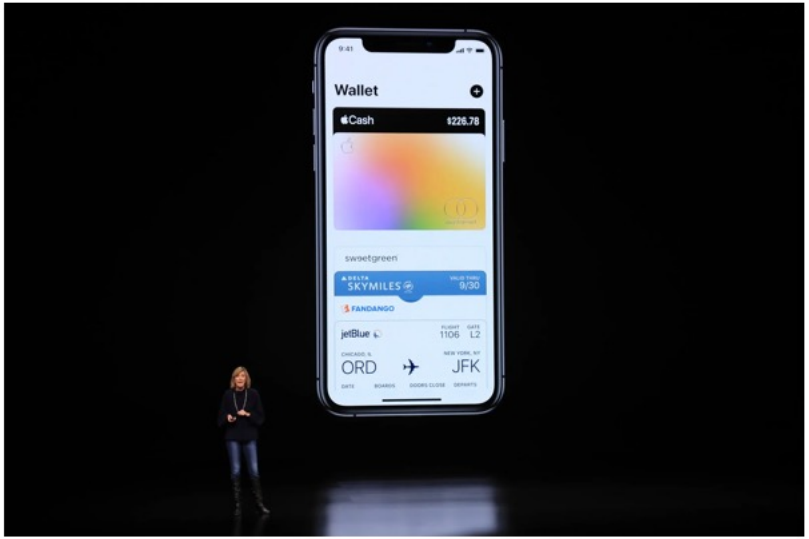
By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card's algorithm, which is used to determine applicants' creditworthiness. Jim Wilson/The New York Times

Related work

Draft NIST Special Publication 1270

A Proposal for Identifying and Managing Bias within Artificial Intelligence

Reva Schwartz
National Institute of Standards and Technology
Information Technology Laboratory

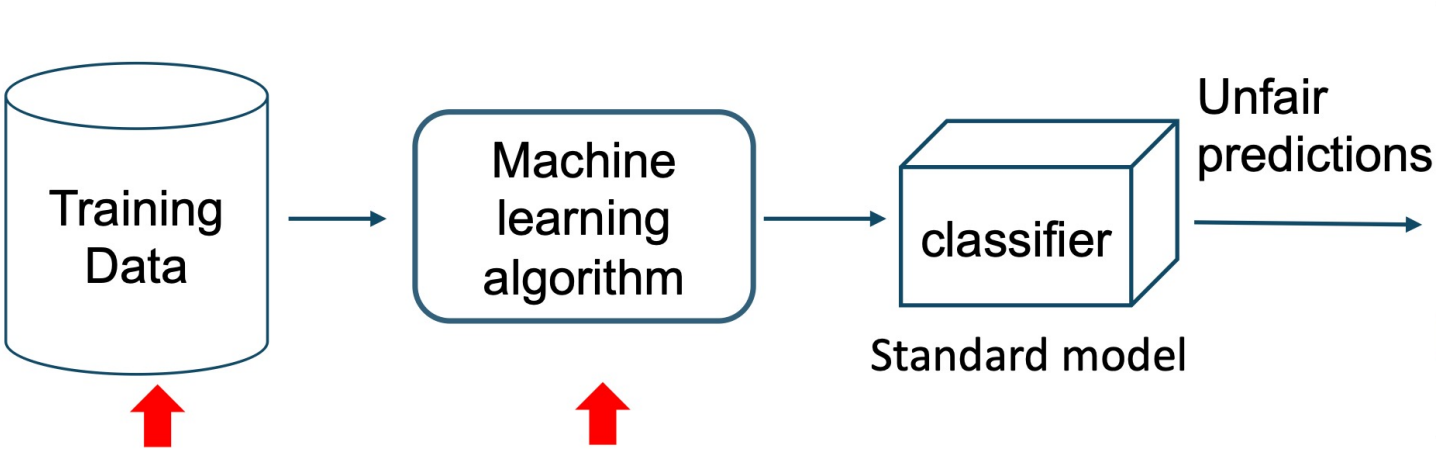
Leann Down
Adam Jonas
Parenthetic, LLC

Elham Tabassi
National Institute of Standards and Technology
Information Technology Laboratory

This draft publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1270-draft

June 2021

Bias in ML



- **Data bias** : training data can reflect the human bias
- **Algorithmic bias:** the model is learned to minimize the overall loss

Ref: [2]

Fairness: An Introduction

- Data driven ML algorithms inherit biases from datasets
- Decisions by algorithms can be discriminatory and harmful[1]
- ***Allocation harms*** can occur when AI systems extend or withhold opportunities, resources, or information. Some of the key applications are in hiring, school admissions, and lending.
- ***Quality-of-service harms*** can occur 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 or withheld. Examples include varying accuracy in face recognition, document search, or product recommendation.

Concepts {Ref 1}

Group fairness, sensitive features

There are many approaches to conceptualizing fairness. In Fairlearn, we follow the approach known as group fairness, which asks: *Which groups of individuals are at risk for experiencing harms?*

The relevant groups (also called subpopulations) are defined using **sensitive features** (or sensitive attributes), which are passed to a Fairlearn estimator as a vector or a matrix called `sensitive_features` (even if it is only one feature). The term suggests that the system designer should be sensitive to these features when assessing group fairness. Although these features may sometimes have privacy implications (e.g., gender or age) in other cases they may not (e.g., whether or not someone is a native speaker of a particular language). Moreover, the word sensitive does not imply that these features should not be used to make predictions – indeed, in some cases it may be better to include them.

Fairness literature also uses the term *protected attribute* in a similar sense as sensitive feature. The term is based on anti-discrimination laws that define specific *protected classes*. Since we seek to apply group fairness in a wider range of settings, we avoid this term.

Concepts {Ref 1}

Parity constraints

Group fairness is typically formalized by a set of constraints on the behavior of the predictor called **parity constraints** (also called criteria). Parity constraints require that some aspect (or aspects) of the predictor behavior be comparable across the groups defined by sensitive features.

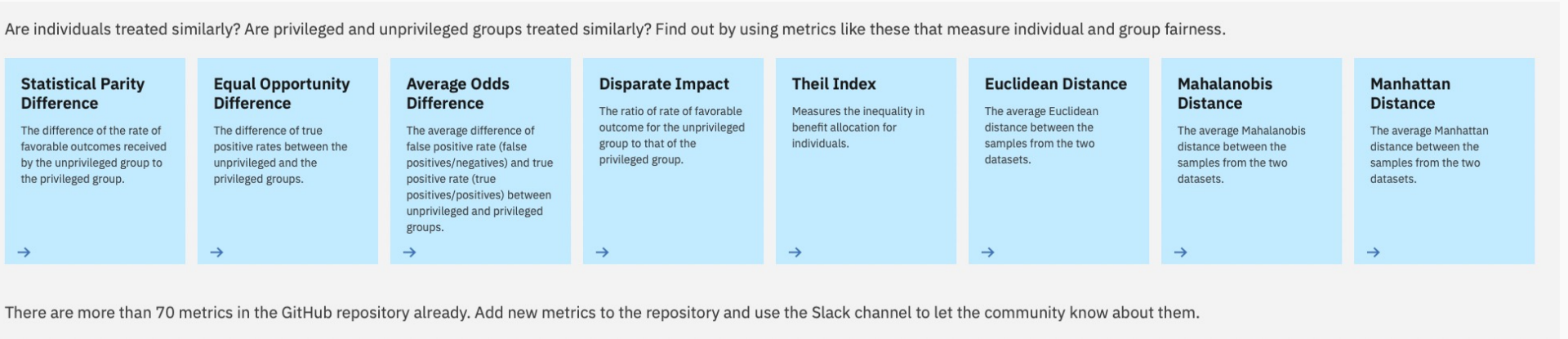
Let X denote a feature vector used for predictions, A be a single sensitive feature (such as age or race), and Y be the true label. Parity constraints are phrased in terms of expectations with respect to the distribution over (X, A, Y). For example, in Fairlearn, we consider the following types of parity constraints.

Binary classification:

- *Demographic parity* (also known as *statistical parity*): A classifier h satisfies demographic parity under a distribution over (X, A, Y) if its prediction $h(X)$ is statistically independent of the sensitive feature A. This is equivalent to $\mathbb{E}[h(X) \mid A = a] = \mathbb{E}[h(X)] \quad \forall a$. [8]
- *Equalized odds*: A classifier h satisfies equalized odds under a distribution over (X, A, Y) if its prediction $h(X)$ is conditionally independent of the sensitive feature A given the label Y. This is equivalent to $\mathbb{E}[h(X) \mid A = a, Y = y] = \mathbb{E}[h(X) \mid Y = y] \quad \forall a, y$. [8]
- *Equal opportunity*: a relaxed version of equalized odds that only considers conditional expectations with respect to positive labels, i.e., $Y = 1$. [7]

Typical steps

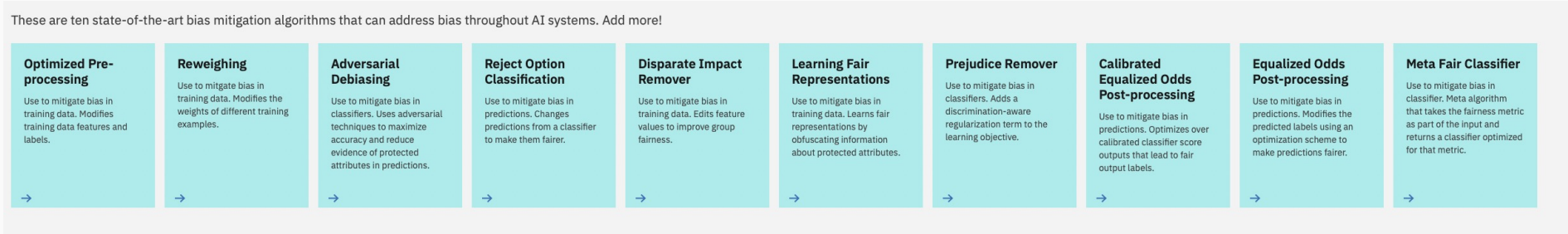
- Assessment
  - Compute various metrics



Source: AIF360

Typical steps

- Mitigation



Source: AIF360

Typical steps

- Mitigation
  - Different algorithms in Fairlearn

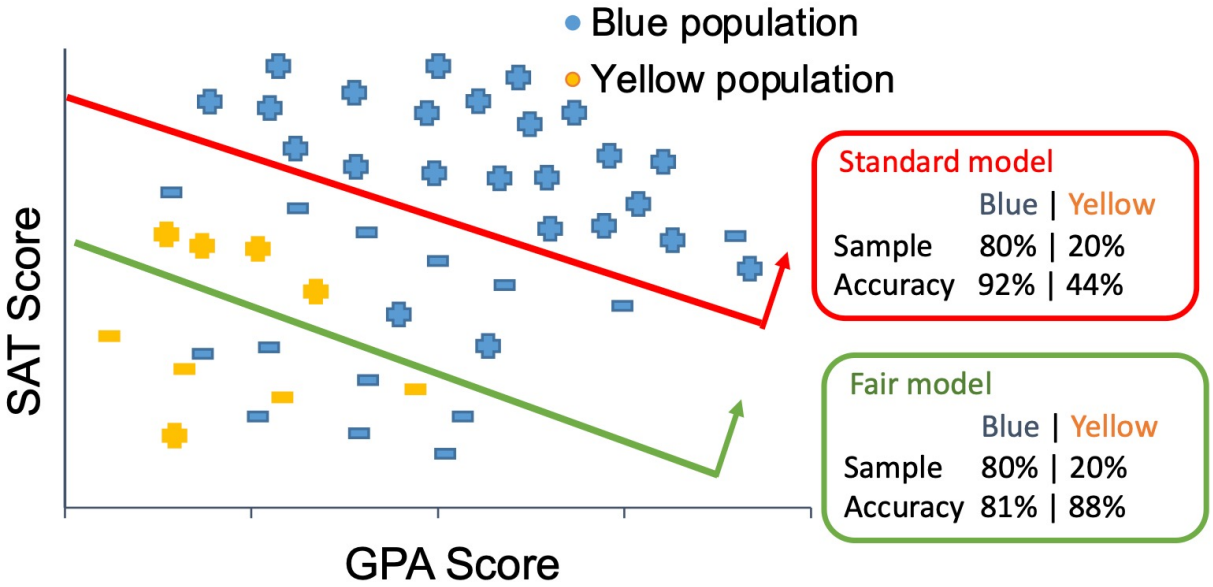
algorithm	description	binary classification	regression	supported fairness definitions
`ExponentiatedGradient`	A wrapper (reduction) approach to fair classification described in *A Reductions Approach to Fair Classification* [5].	✔	✔	DP, EO, TPRP, FPRP, ERP, BGL
`GridSearch`	A wrapper (reduction) approach described in Section 3.4 of *A Reductions Approach to Fair Classification* [5]. For regression it acts as a grid-search variant of the algorithm described in Section 5 of *Fair Regression: Quantitative Definitions and Reduction-based Algorithms* [4].	✔	✔	DP, EO, TPRP, FPRP, ERP, BGL
`ThresholdOptimizer`	Postprocessing algorithm based on the paper *Equality of Opportunity in Supervised Learning* [6]. This technique takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the classifier's prediction to enforce the specified parity constraints.	✔	✘	DP, EO, TPRP, FPRP
`CorrelationRemover`	Preprocessing algorithm that removes correlation between sensitive features and non-sensitive features through linear transformations.	✔	✔	✘

Reductions[1]

- On a high level, the reduction algorithms within Fairlearn enable unfairness mitigation for an arbitrary machine learning model with respect to user-provided fairness constraints.
- The reductions approach for classification seeks to reduce binary classification subject to fairness constraints to a sequence of weighted classification problems

Fairness in ML

- **Equalized odds:** TPR and TNR should be similar across protected groups that are defined by a sensitive attribute (e.g., race, gender)



Side effect: **increase** the influence of the training data from underprivileged group on the learned model

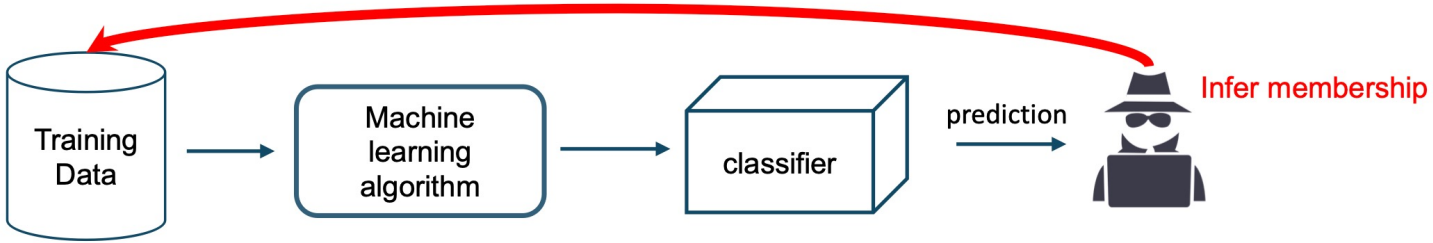
Hardt, Moritz, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." NeurIPS 2016
Example is from Michael Kearns & Aaron Roth talk at Google

Fairness and privacy

- Fair machine learning aims at minimizing discrimination against protected groups by, for example, imposing a constraint on models to equalize their behavior across different groups.
- Change the influence of training data points on the fair model, in a disproportionate way
- But this can lead to information leakage of the model about its training data
- This paper analyzes the privacy risks of group fairness (e.g., equalized odds) through the lens of membership inference attacks

Fairness meets privacy

Membership inference attack: infer whether an individual's data is in the training dataset or not



Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." TCC, 2006.
Shokri, Reza, et al. "Membership inference attacks against machine learning models." *SP,* 2017.

Is there a privacy cost for achieving fairness?

Approaches:

- Analyzing models which are trained with differential privacy and fairness constraints and evaluating the compatibility of the two measures
- Formalize privacy risk as the success of membership inference attacks against machine learning models

Key conclusions

- Empirically show that the fairness-aware learning has a disparate impact on the privacy risk of subgroups, and in particular, it increases the privacy risk of the unprivileged subgroup
- When the underlying data and the corresponding unconstrained model are more “unfair”, the trained fair models leak more information about the unprivileged subgroups
- The more fair a model is, the higher the privacy risk of the model on the unprivileged subgroups will be

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.

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

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 $\delta = 0$.

Privacy risk

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.

We define privacy risk of algorithm A with respect to an individual data point z as the probability that the most powerful adversary wins the attack game.

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.

The individual privacy risk is equivalent to the average true positive and true negative rates $\frac{1}{2}(\Pr[\hat{b} = 1|b = 1] + \Pr[\hat{b} = 0|b = 0])$ of the adversary.

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.

Definitions

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:

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

Quantifying privacy risk

- Adversary has black-box access to the model, and can compute the loss of the model on any input data
- A simple attack model is to compare the model's loss on an input with a threshold. The attack outputs "member" if the loss is below the threshold, and "nonmember" otherwise
- The adversary can design a separate membership inference attack for each subgroup
- The adversary can compute the loss threshold based on the knowledge about the population or through using shadow models

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.

Experiments and Empirical analysis

Data and Models. We generate synthetic datasets, of size $2,500$ records, similar to the prior work on analyzing fairness [33]. Specifically, we generate binary sensitive attributes, for each record, from a Bernoulli distribution $P_g = \Pr[G = g]$, for $g \in \{0, 1\}$. We generate binary labels from a Bernoulli distributions $P_g^y = \Pr[Y = y|G = g]$, for $y \in \{-, +\}$ and for all $g \in \{0, 1\}$. We generate a 2-dimensional feature vector from four different Gaussian distributions:

$$\begin{aligned}
&\text{For subgroup } G_0^- : \; X \sim \mathcal{N}([0, -1], [7, 1; 1, 7]) \\
&\text{For subgroup } G_1^- : \; X \sim \mathcal{N}([-5, 0], [5, 1; 1, 5]) \\
&\text{For subgroup } G_0^+ : \; X \sim \mathcal{N}([1, 2], [5, 2; 2, 5]) \\
&\text{For subgroup } G_1^+ : \; X \sim \mathcal{N}([2, 3], [10, 1; 1, 4]) \qquad (3)
\end{aligned}$$

where, $\mathcal{N}(\mu, \Sigma)$ represents a Gaussian distribution with mean vector μ and covariance matrix Σ.

We set $P_0 = 0.2$, $P_0^- = 0.1$ and $P_1^- = 0.5$. Accordingly, $P_1 = 0.8$, $P_0^+ = 0.9$ and $P_1^+ = 0.5$. Group G_0,

Accuracy and fairness gap

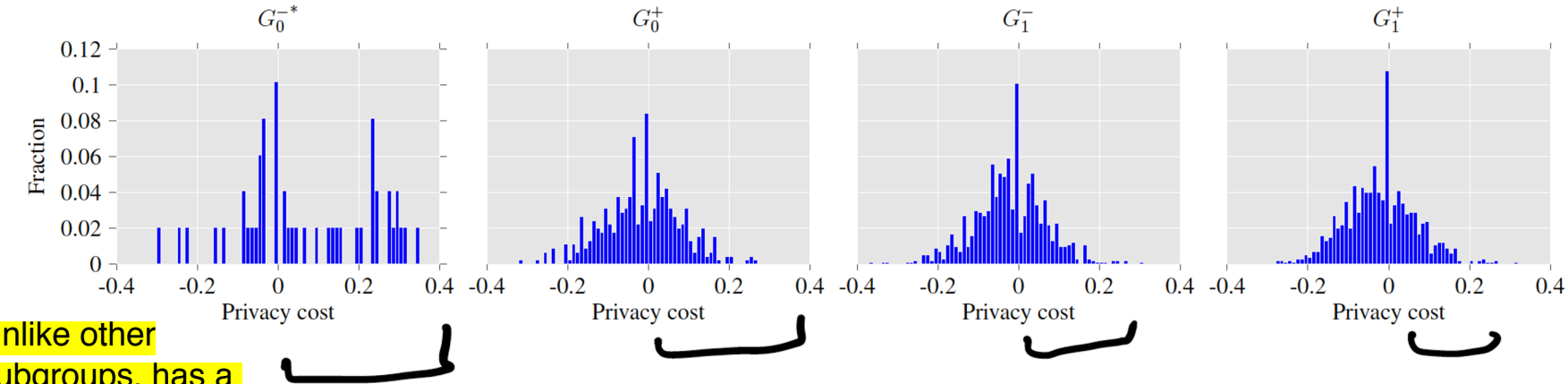
TABLE 1: Accuracy and fairness gap of unconstrained and fair models (with three different fairness constraints δ) on the synthetic datasets. The "Train Δ" and "Test Δ" columns show the fairness gap (as defined in (1)) on the training and test data.

Model	Train acc	Test acc	Train Δ	Test Δ
Unconstrained	86.2%	85.5%	0.373	0.430
Fair ($\delta = 0.1$)	89.1%	87.8%	0.105	0.332
Fair ($\delta = 0.01$)	85.6%	84.0%	0.014	0.283
Fair ($\delta = 0.001$)	84.5%	83.8%	0.001	0.275

TABLE 2: Accuracy of membership inference attacks with a fixed loss threshold for all data points, versus using multiple thresholds one for each subgroup - Synthetic dataset.

Model	Attack	G_0^-	G_1^-	G_0^+	G_1^+
Unconstrained	Single	52.9%	51.2%	51.8%	51.2%
	Group-based	61.8%	52.8%	52.4%	52.2%
Fair ($\delta = 0.001$)	Single	60.8%	51.9%	51.6%	50.8%
	Group-based	69.2%	53.4%	52.5%	51.6%

Histogram for individual privacy cost, across different subgroups, on models trained on synthetic data



Unlike other subgroups, has a larger fraction of samples with positive privacy cost

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 privacy cost is 0.069, −0.015, −0.02, −0.02 for subgroups G_0^-, G_0^+, G_1^-, and G_1^+ respectively.

Privacy cost = **Privacy risk on fair model** - **Privacy risk on unconstrained model**

Vulnerable points on fair models

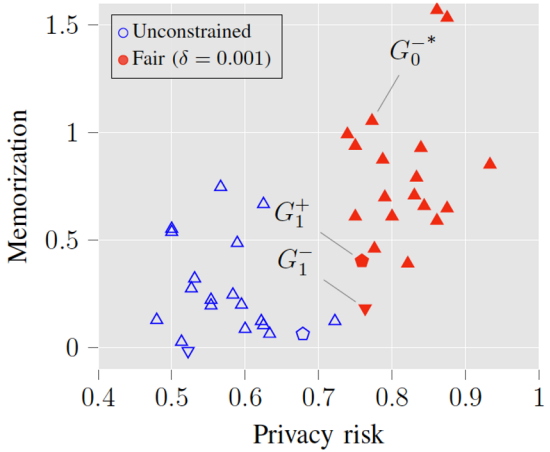


Figure 2: The most vulnerable points on fair models (trained on synthetic data). We find the top 20 vulnerable points that have the highest privacy risk on fair models. For each vulnerable point, we show its privacy risk and the memorization of models before (blue color) and after (red color) imposing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

points are mainly
from the unprivileged subgroup

Loss distribution of unconstrained and fair model on subgroup G_0^-

members of training set
are more distinguishable from non-members on fair models compared with unconstrained models

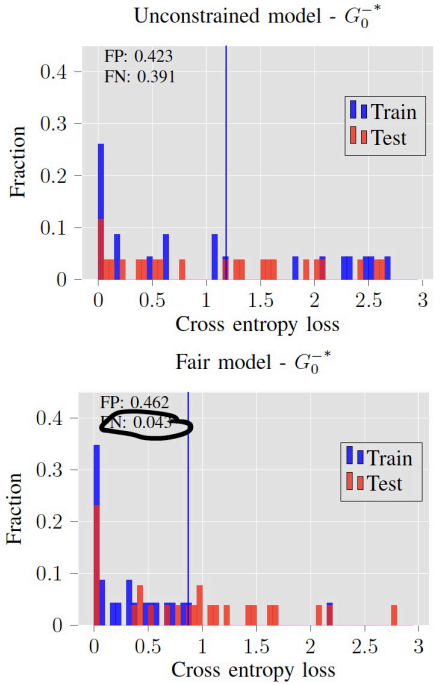


Figure 3: Loss distribution of an unconstrained model and a fair model on subgroup G_0^-. The vertical blue line shows the loss threshold used in the membership inference attack. FN is the false negative, and FP is the false positive rate for the attack.

Memorization and training accuracy of fair and unconstrained models on all subgroups

Fair models memorize the points in the underprivileged subgroup

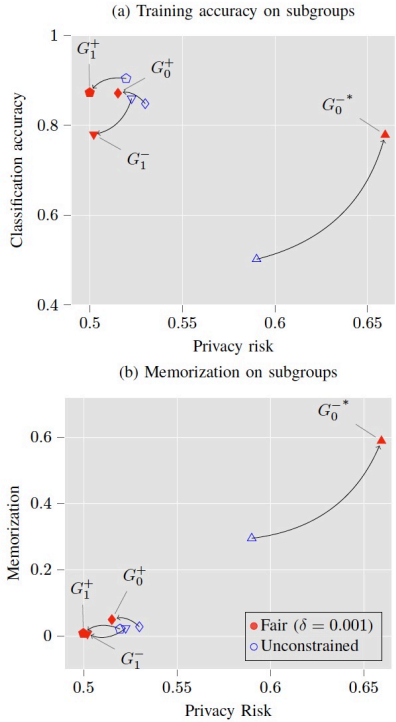


Figure 4: Memorization and training accuracy of fair and unconstrained models on all subgroups - trained on synthetic data. The blue/red color show the results on the unconstrained/fair models. The markers represent different subgroups.

Accuracy gain versus privacy cost on the unprivileged subgroup

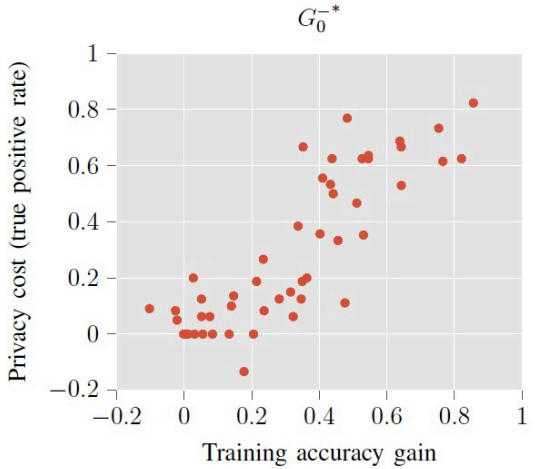


Figure 5: Accuracy gain versus privacy cost on the unprivileged subgroup G_0^-. Each point in the plot represents a data point in the training dataset. The training accuracy gain is the difference in the training accuracy between fair and unconstrained models. The y-axis is the difference in the attacker's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

clear correlation between the accuracy gain and the privacy cost

Effect of enforced fairness level

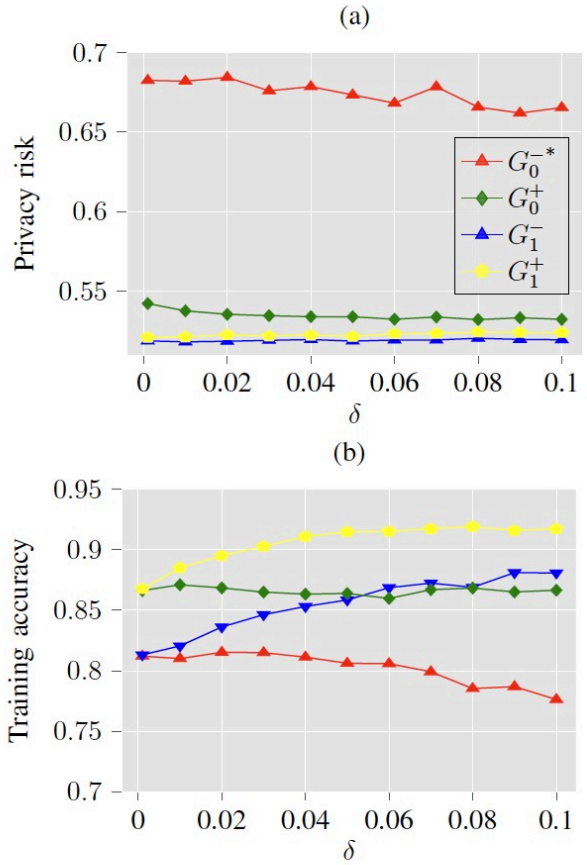


Figure 6: (a) The effect of the enforced fairness level δ on the privacy risk of fair models for different subgroups. (b) The effect of enforced fairness level on the classification accuracy of fair models for different subgroups.

Effect of enforced fairness level

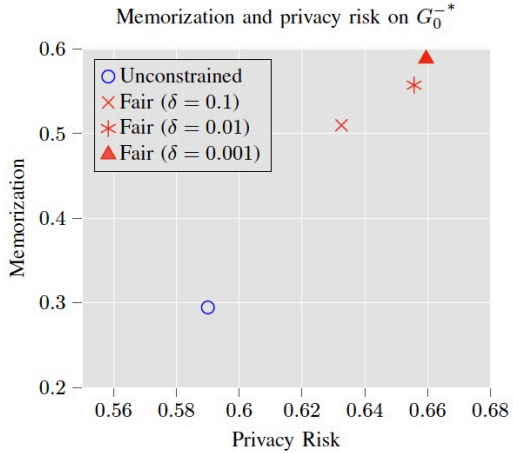


Figure 7: The effect of the enforced fairness gap δ on the privacy risk and memorization of the model - Synthetic dataset.

Effect of underlying unfairness on privacy cost.

Large fairness gap in the underlying unconstrained model results in a large privacy cost (of imposing fairness constraints)

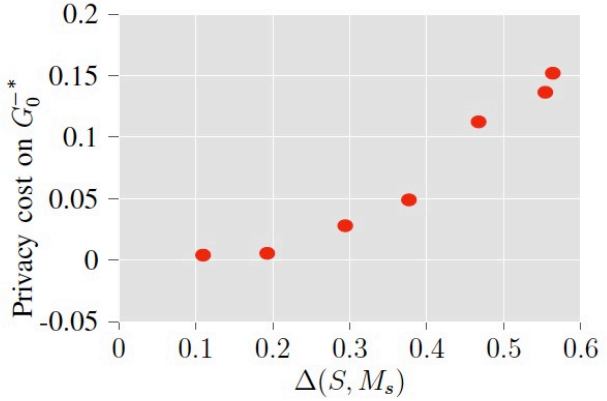


Figure 8: The effect of the unconstrained model's fairness gap (which captures the unfairness that needs to be removed by the fair algorithm) on the privacy cost of G_0^-. The x-axis is the fairness gap of unconstrained models on the training dataset.

Smaller number of samples in the unprivileged subgroup results in a higher privacy cost for the unprivileged group.

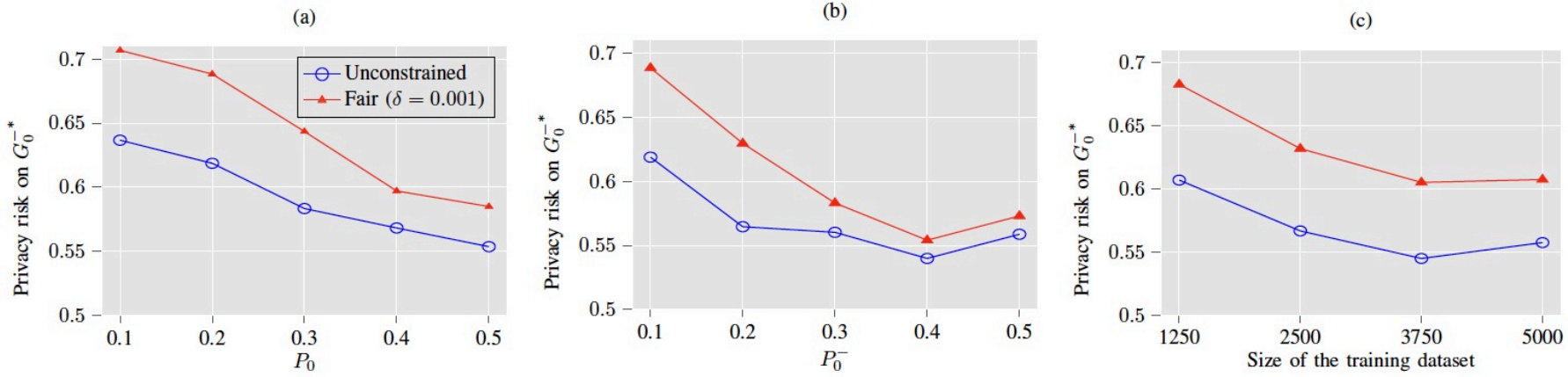


Figure 9: Privacy risk on the unprivileged subgroup G_0^-, for various faction of data in group G_0, various fraction of data in subgroup G_0^-, and for various training set size.

Post-processing algorithm does not improve the training accuracy of the unprivileged subgroup

TABLE 3: Prediction accuracy and privacy risk for unconstrained models, and fair models trained using reduction approach [3] and post-processing (PP) approach [6].

Model		G_0^{-*}	G_1^-	G_0^+	G_1^+
Unconstrained	Train acc	52.2%	85.6%	85.3%	89.8%
	Test acc	39.1%	84.7%	86.2%	89.5%
	Privacy risk	0.639	0.521	0.518	0.523
Fair (reduction) ($\delta = 0.001$)	Train acc	80.1%	80.2%	88.4%	88.5%
	Test acc	49.2%	79.2%	88.1%	88.3%
	Privacy risk	0.688	0.521	0.542	0.518
Fair (PP) ($\delta = 0$)	Train acc	48.4%	48.4%	90.7%	90.7%
	Test acc	41%	47.9%	89.4%	90.5%
	Privacy risk	0.55	0.507	0.509	0.504

Similar patterns across different group-fairness metrics

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 (EO)	Train acc	81.2%	81.3%	86.6%	86.8%
	Test acc	53.8%	80.9%	83.8%	86.4%
	Privacy risk	0.683	0.519	0.542	0.521
Fair (EOPP)	Train acc	47.4%	91%	91.6%	91.7%
	Test acc	39.1%	90.1%	89.7%	91.3%
	Privacy risk	0.605	0.52	0.534	0.522
Fair (FPP)	Train acc	83.1%	83.2%	85.5%	91.8%
	Test acc	54.5%	82.9%	84%	90.9%
	Privacy risk	0.679	0.518	0.535	0.523

Real data Experiments

Data and models. We conduct experiments on the Law School dataset (Law) [27][2] (with 19 features and $16,672$ data points), Bank Marketing dataset (Bank) [28] (with 58 features and $24,391$ data points), and COMPAS dataset [29] (with 11 features and $4,302$ data points). We use the same preprocessing on these datasets as in IBM's AI Fairness 360 [35]. For COMPAS and Law datasets, we consider two versions for each dataset, one where the protected attribute is "race" (white versus non-white) and the other where the protected attribute is "gender" (male versus female). For Bank dataset, the protected attribute is "age" (age ≥ 25 versus age < 25). Table 5 shows the distribution of data points across different subgroups. For all datasets, we use 50% of the available data for training and the remaining 50% for test.

TABLE 5: The data partitioning based on the protected attributes, and the percentage of data in different subgroups for the real-world datasets.

Name	G_0^-	G_1^-	G_0^+	G_1^+
Bank (age)	2.2%	85.2%	0.7%	12.0%
COMPAS (race)	28.3%	24.4%	31.7%	15.6%
COMPAS (gender)	12.8%	39.9%	7.3%	40.0%
Law (race)	2.3%	2.7%	13.5%	81.5%
Law (gender)	2.5%	2.5%	41.4%	53.6%

TABLE 6: Accuracy and fairness gap Δ of unconstrained models and fair models (with different enforced fairness level δ) on the training and test dataset – Decision tree model with max depth 10.

Dataset	Model	Train acc	Test acc	Train Δ	Test Δ
Bank (age)	Unconstrained	92.5%	87.8%	0.063	0.074
	Fair ($\delta = 0.1$)	94.7%	89.3%	0.057	0.078
	Fair ($\delta = 0.01$)	94.6%	89.3%	0.011	0.063
	Fair ($\delta = 0.001$)	94.6%	89.3%	0.001	0.066
COMPAS (race)	Unconstrained	72.4%	60.1%	0.097	0.133
	Fair ($\delta = 0.1$)	79.4%	64.3%	0.108	0.131
	Fair ($\delta = 0.01$)	79%	64%	0.017	0.073
	Fair ($\delta = 0.001$)	78.7%	64%	0.002	0.067
COMPAS (gender)	Unconstrained	72.4%	60.2%	0.117	0.107
	Fair ($\delta = 0.1$)	79.3%	64.4%	0.081	0.1
	Fair ($\delta = 0.01$)	78.6%	64.4%	0.013	0.083
	Fair ($\delta = 0.001$)	78.5%	64.3%	0.001	0.08
Law (race)	Unconstrained	95.9%	92.2%	0.236	0.165
	Fair ($\delta = 0.1$)	97.6%	93.6%	0.148	0.156
	Fair ($\delta = 0.01$)	97.5%	93.4%	0.018	0.12
	Fair ($\delta = 0.001$)	97.5%	93.5%	0.002	0.112
Law (gender)	Unconstrained	95.9%	92.2%	0.035	0.039
	Fair ($\delta = 0.1$)	97.6%	93.6%	0.097	0.039
	Fair ($\delta = 0.01$)	97.6%	93.6%	0.019	0.04
	Fair ($\delta = 0.001$)	97.6%	93.6%	0.002	0.033

TABLE 7: Privacy risk of unconstrained and fair models (with $\delta = 0.001$) across different subgroups – Decision tree models with max depth 10. We indicate the protected attribute for each dataset. Unprivileged subgroups are identified by asterisks.

Dataset	Model	G_0^-	G_1^-	G_0^+	G_1^+
Bank	Unconstrained	0.545	0.516	0.645	0.611*
(age)	Fair	0.574	0.521	0.707	0.644
COMPAS	Unconstrained	0.582	0.565	0.579	0.611*
(race)	Fair	0.599	0.589	0.601	0.648
COMPAS	Unconstrained	0.583	0.569*	0.643	0.576
(gender)	Fair	0.572	0.591	0.643	0.599
Law	Unconstrained	0.726	0.711*	0.541	0.510
(race)	Fair	0.745	0.818	0.555	0.515
Law	Unconstrained	0.721	0.724*	0.514	0.514
(gender)	Fair	0.774	0.788	0.521	0.519

TABLE 8: Prediction accuracy and privacy risk of unconstrained and fair models with different enforced fairness gap δ on decision tree models with max depth 10 – Law (race) dataset.

	Model	G_0^-	G_1^{-*}	G_0^+	G_1^+
Train acc	Unconstrained	70.1%	43.4%	99.0%	99.8%
	Fair ($\delta = 0.1$)	64.7%	49.7%	99.3%	99.8%
	Fair ($\delta = 0.01$)	55.6%	53.8%	99.6%	99.8%
	Fair ($\delta = 0.001$)	54.5%	54.3%	99.6%	99.7%
Test acc	Unconstrained	28.6%	10.8%	92.2%	98.6%
	Fair ($\delta = 0.1$)	26.8%	10.9%	92.8%	98.4%
	Fair ($\delta = 0.01$)	23.7%	10.9%	93.5%	98.2%
	Fair ($\delta = 0.001$)	23.3%	11.8%	94.0%	98.1%
Privacy risk	Unconstrained	0.726	0.711	0.541	0.510
	Fair ($\delta = 0.1$)	0.744	0.777	0.550	0.513
	Fair ($\delta = 0.01$)	0.743	0.810	0.553	0.514
	Fair ($\delta = 0.001$)	0.745	0.818	0.555	0.515

TABLE 9: Privacy risk of unconstrained and fair models on Law (race) dataset (with $\delta = 0.001$) – Decision tree models. The "DT-x" row shows the results on decision tree models with max depth x.

Model type	Model	G_0^-	G_1^{-*}	G_0^+	G_1^+
DT-5	Unconstrained	0.585	0.561	0.515	0.503
	Fair	0.574	0.583	0.517	0.504
DT-10	Unconstrained	0.726	0.711	0.541	0.510
	Fair	0.745	0.818	0.555	0.515
DT -15	Unconstrained	0.815	0.874	0.557	0.516
	Fair	0.879	0.955	0.589	0.527

Strengths

- Practical approach to quantifying privacy risk and evaluating tradeoffs between fairness and privacy
- Builds upon prior work and simple to implement

Improvements

Mitigating and how to address privacy risks – area of research.

References

1. https://fairlearn.org/v0.7.0/api_reference/index.html
2. Author slides source: https://www.ieee-security.org/TC/EuroSP2021/slides/Hongyan%20Chang%20-%20Hongyan%20Chang-On%20the%20Privacy%20Risks%20of%20Algorithmic%20Fairness.pdf

Differential Privacy Has Disparate Impact on Model Accuracy (May 2019 v1)

By : Eugene Bagdasaryan
Vitaly Shmatikov

Presented by: Manjit Ullal
December 02, 2021

Motivation

Privacy preserving technique, Differential privacy (DP) does not distribute the cost of reduction in model's accuracy equally across the subgroups in data. This paper demonstrates that in the neural networks the accuracy of the model drops much more for the underrepresented classes.

Examples

The Machine
Making sense of AI

Audit finds gender and age bias in OpenAI's CLIP model

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Table 2. Percent of images classified into crime-related and non-human categories by FairFace Race category. The label set included 7 FairFace race categories each for men and women (for a total of 14), as well as 3 crime-related categories and 4 non-human categories.

AUTO FINANCE NEWS

HOME NEWS EVENTS EXCELLENCE MAGAZINE PODCAST DATA + AFN PLUS

CFPB's 'BISG' Algorithm Not Designed to Determine Race, Creator Says

Related work

- Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness - Michael Kearns et al.
- Differentially Private Fair Learning - Michael Kearns, Matthew Jagielski, Alina Oprea et al.
- Other techniques like oversampling, cost sensitive learning, adversarial learning and re-sampling cannot be combined with DP-SGD due to sensitivity bounds enforced by DP-SGD.

Methodology

Techniques

1. Differential privacy
2. Federated learning

Tasks

1. Gender and Age classification
2. Sentiment analysis
3. Species classification
4. Word Prediction

Metric

1. Model accuracy

Methodology - Differential Privacy

A randomised mechanism $M : D \to R$, with domain D and range R.

Satisfies (ϵ, δ) differential privacy for any two adjacent datasets $d, d^{'} \in D$ and for any subset of outputs we have

$S \subseteq R, Pr[M(d) \in S] \leq e^{\epsilon} Pr[M(d^{'}) \in S] + \delta$

Algorithm 1: Differentially Private SGD (DP-SGD)

Input: Dataset $(x_1, y_1), ..., (x_N, y_N)$ of size N, batch size b, learning rate η, sampling probability q, loss function $\mathcal{L}(\theta(x), y)$, K iterations, noise σ, clipping bound S, $\pi_S(x) = x * \texttt{min}(1, \frac{S}{||x||_2})$

Initialize: Model θ_0

1 **for** $k \in [K]$ **do**

2 randomly sample $batch$ from dataset N with probability q

3 **foreach** (x_i, y_i) *in batch* **do**

4 $g_i \leftarrow \nabla \mathcal{L}(\theta_k(x_i), y_i)$

5 $g_{batch} = \frac{1}{qN}(\sum_{i \in batch} \pi_S(g_i) + \mathcal{N}(0, \sigma^2 \boldsymbol{I}))$

6 $\theta_{k+1} \leftarrow \theta_t - \eta g_{batch}$

Output: Model θ_K and accumulated privacy cost (ϵ, δ)

Methodology - Federated learning

Distributed learning framework, n participants jointly train a model. At each round t, global server distributes the current model G_t to small subgroup d_C, where each local model produces a new model L_{t+1}. Then the global server aggregates these model updates as below.

$G_{t+1} = G + t + \frac{\eta_g}{n} \sum_{i \in d_C} (L^i_{t+1} - G_t)$

DP-FedAvg algorithm

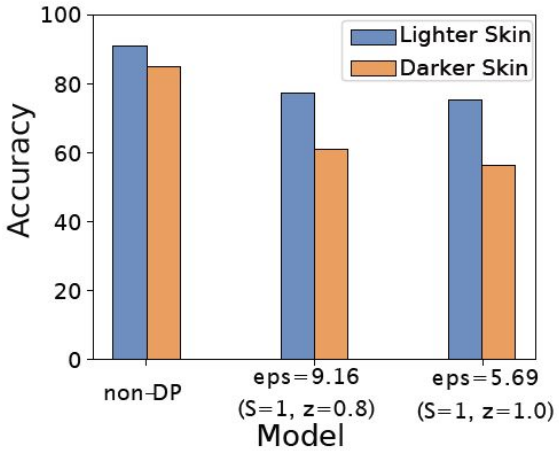
$G_{t+1} = G + t + \frac{\eta_g}{n} \sum_{i \in d_C} \pi_S(L^i_{t+1} - G_t) + \mathcal{N}(0, \sigma^2 I)$, where $\sigma = \frac{zS}{C}$

Experiments

$\epsilon < 10$, $\delta < 10^{-6}$

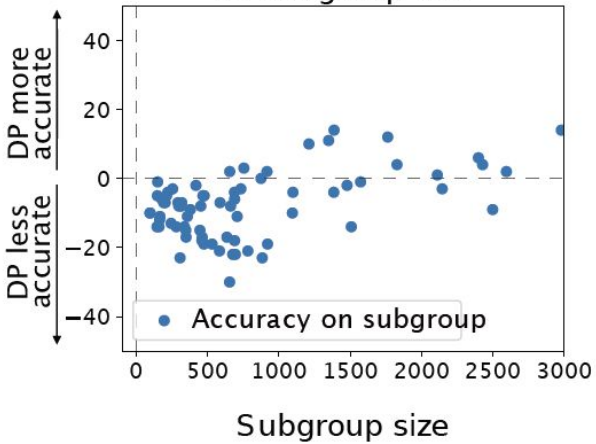
Gender Classification
Resnet 18

(a) Accuracy vs Model type



Age Classification
Resnet 18

(b) DP model accuracy relative to non-DP vs Subgroup size



Age Classification
Resnet 18

(c) DP model accuracy relative to non-DP vs non-DP model accuracy

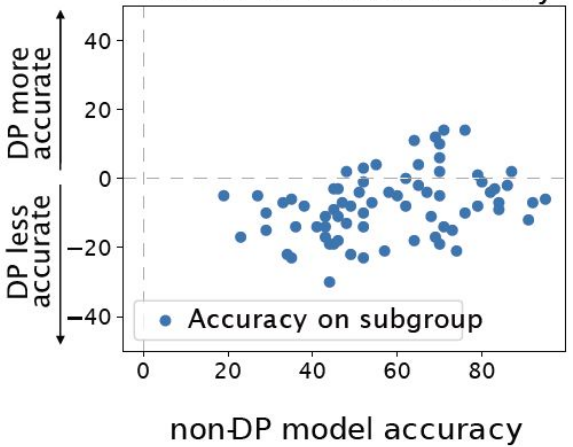


Figure 1: Gender and age classification on facial images.

Experiments

$\epsilon = [3.87, 8.99]$, $\delta < 10^{-6}$

$\epsilon = 4.67$, $\delta < 10^{-6}$

Sentiment Analysis
Bi-LSTM

Species Classification
Inception V3

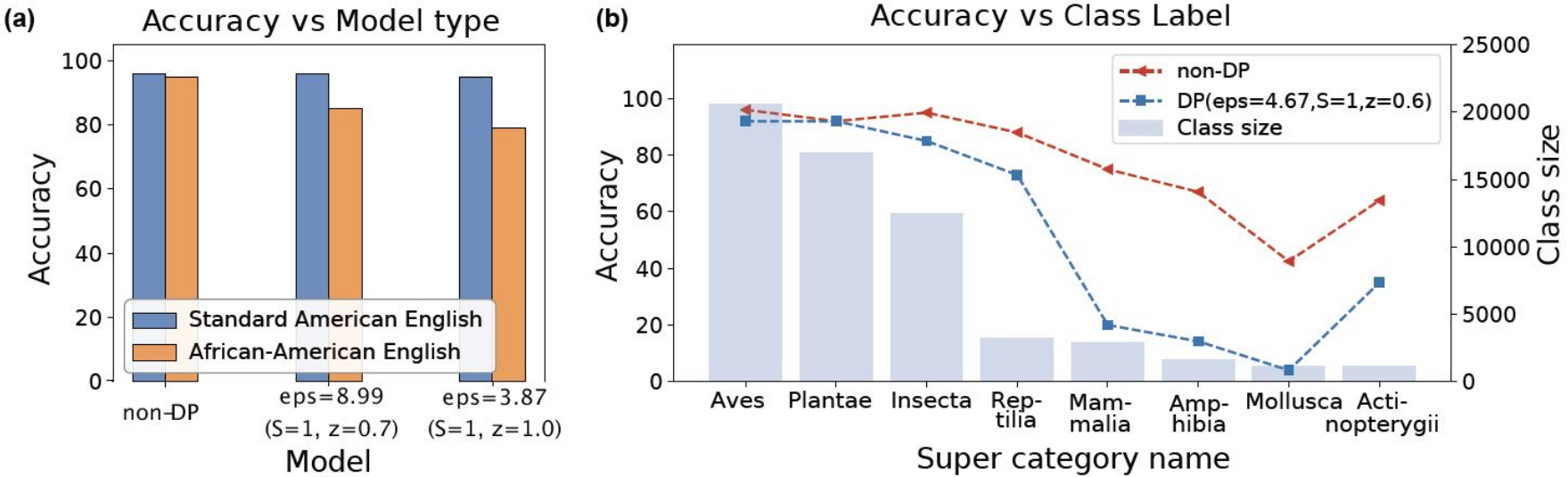
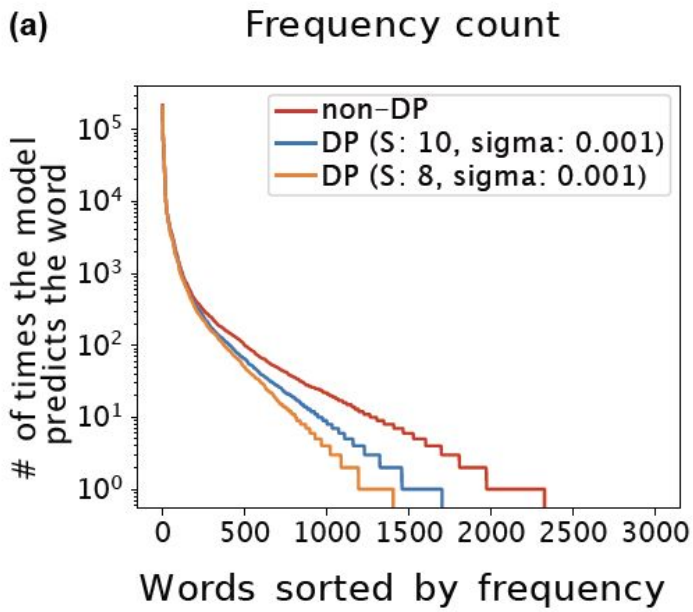


Figure 2: Sentiment analysis of tweets and species classification.

Experiments

δ = 0.001

Word Prediction
LSTM



Word Prediction
LSTM

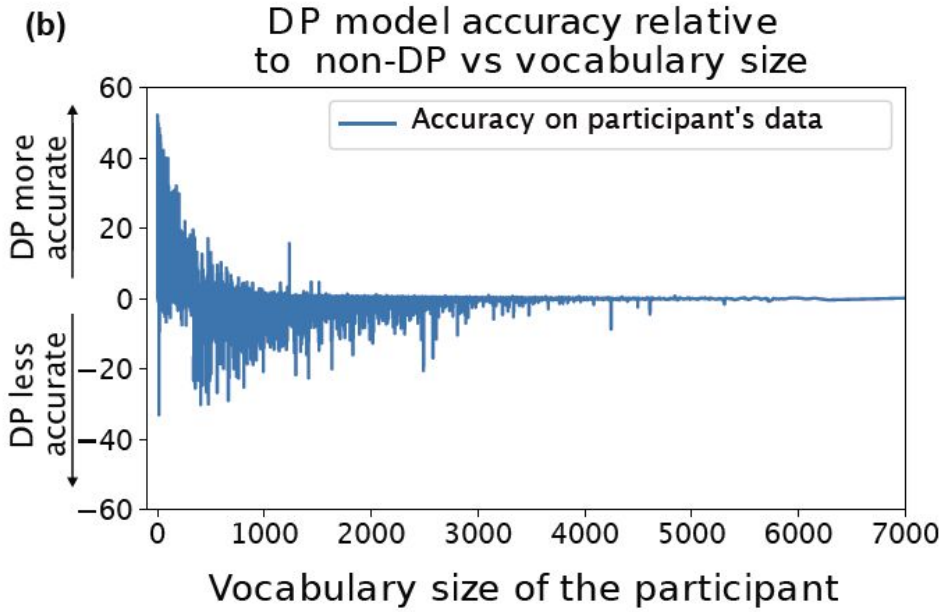


Figure 3: Federated learning of a language model.

Effect of hyperparameters

Image Classification
CNN

$\epsilon = 6.23$, $\delta < 10^{-6}$

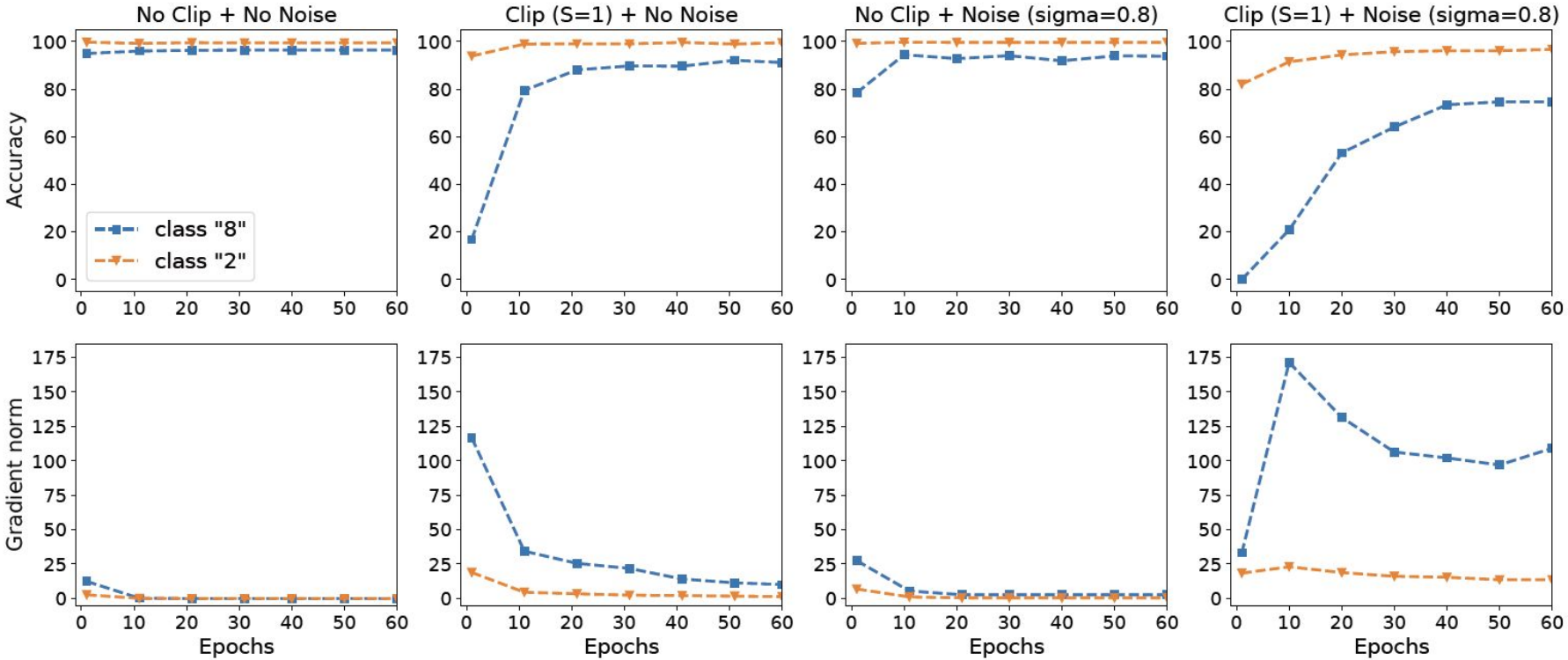


Figure 4: Effect of clipping and noise on MNIST training.

Effect of hyperparameters

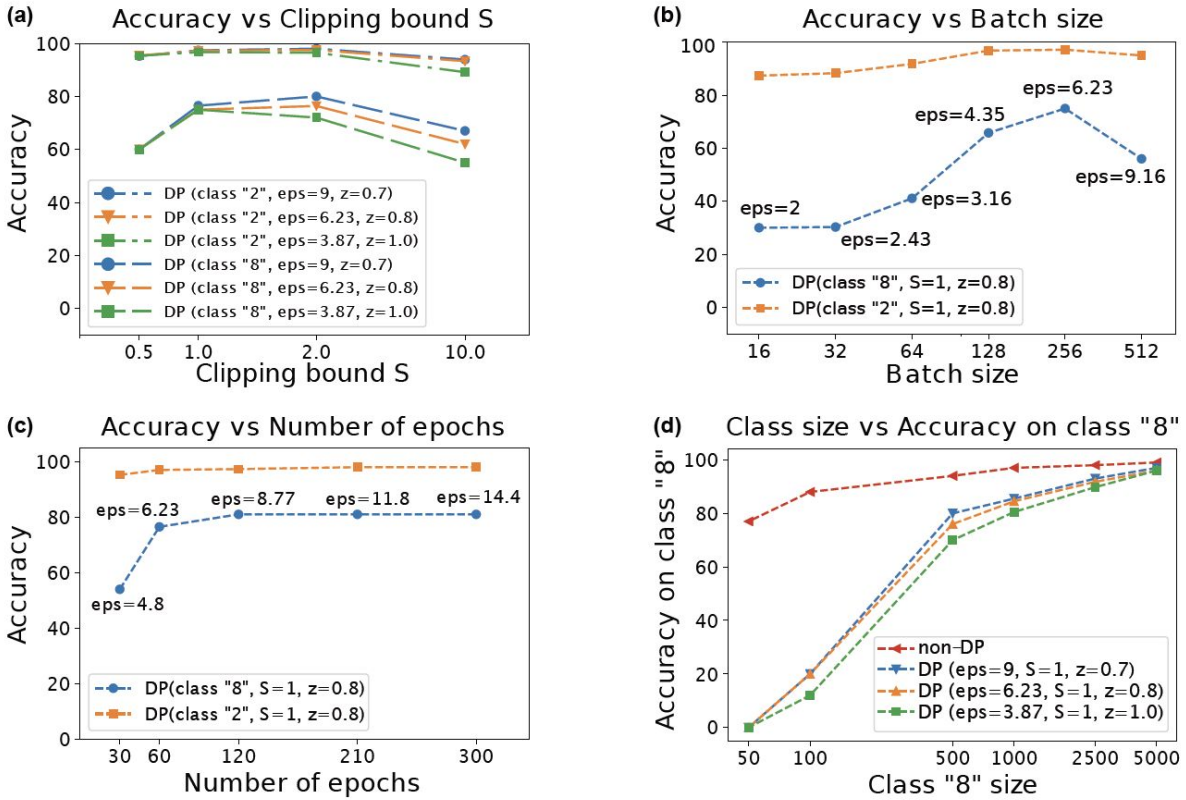


Figure 5: Effect of hyperparameters on MNIST training.

Conclusions

- DP-SGD computes gradient for each training example and averages them per class so there will be fewer representation on minority class in a randomised small batches.
- Clipping along with noise has the maximum effect on causing the the imbalance in the cost of model accuracy.

Strengths

1. Provides empirical evidence of bias against minority class during differential private (DP) training.
2. Performs Thorough experiments to validate above hypothesis.
3. Experiments with various hyper parameters of DP training and provides an explanation for the trend.

Limitations

1. The datasets are intentionally biased and it may not be realistic.
2. In general the minority classes have lower accuracy so lower accuracy in DP trained model can be expected and it is not a novel discovery.
3. Datasets and labeling procedure in few of the experiments are questionable.