How unfair is private learning?

Amartya Sanyal[*1,2], Yaxi Hu [*3], Fanny Yang[1,2]

[*]Equal contribution; [1]Department of Computer Science; [2]ETH AI Center; [3] Department of Mathematics, ETH Zürich.

Overview

Machine learning (ML) methods are regularly used in sensitive and impactful applications.

We want ML algorithms to satisfy various qualities:

- **Accuracy**: Have high overall accuracy.
- **Privacy**: Not leak private training data.
- **Fairness**: Perform equitably on different subpopulations.

Question: Is it possible to satisfy these three properties simultaneously in real world data ?

This work: We study **fairness** of **private** and **accurate** algorithms for data with **multiple subpopulations**.

Data with subpopulations

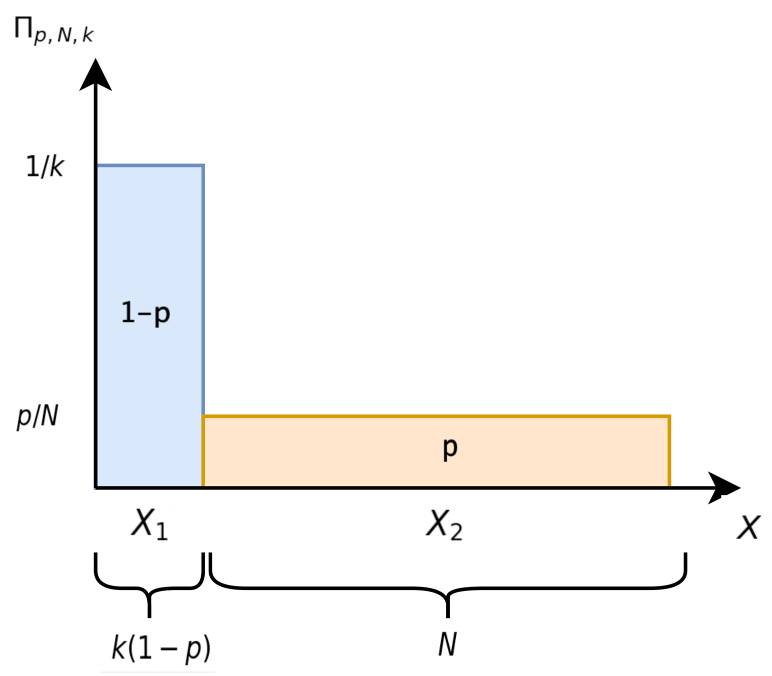
Consider a discrete set X without any structure. Each subpopulation is an element of X.

Given $p \in (0,1)$, $1 < k \ll N \in \mathbb{N}$, define two groups

- $X_1 \subset X$: Majority subpopulations $|X_1| = (1-p)\,k$
- $X_2 := X \setminus X_1$: Minority subpopulations $|X_2| = N$.

Define the **distribution**

$$\Pi_{p,N,k}(x) = \begin{cases} \frac{1}{k} & x \in X_1 \\ \frac{p}{N} & x \in X_2. \end{cases}$$



Label prior $\mathcal{F}$ is a distribution over labelling functions $\mathcal{Y}^X$.

Generating a dataset of size m

- Sample unlabelled dataset $S = \{x_1, \cdots, x_m\} \sim \Pi^m_{p,N,k}$.
- Sample labelling function $f \sim \mathcal{F}$.
- Create labelled dataset $S_f = \{(x_1, f(x_1)), \ldots, (x_m, f(x_m))\}$.

Privacy and Fairness

•Privacy: Differential Privacy

An algorithm $\mathcal{A}$ is (ϵ,δ)-DP if for any two neighboring datasets S_1, S_2 and for all subsets Q in $\mathrm{im}(\mathcal{A})$.

$$\mathbb{P}[\mathcal{A}(S_1) \in Q] \le e^{\epsilon}\mathbb{P}[\mathcal{A}(S_2) \in Q] + \delta$$

•Error of an algorithm $\mathcal{A}$ on a distribution Π with label prior $\mathcal{F}$ is

$$\mathrm{err}(\mathcal{A}, \Pi, \mathcal{F}) = \mathbb{E}_{x,h,f,S_m}[\mathbb{I}\{h(x) \neq f(x)\}]$$

•(Un)-Fairness: Accuracy Discrepancy

The accuracy discrepancy of an algorithm $\mathcal{A}$ on the distribution $\Pi_{p,N}$ with label $\mathcal{F}$ is

$$\Gamma(A, \Pi_{p,N}) = \mathrm{err}\left(\mathcal{A}, \Pi^2_{p,N}, \mathcal{F}\right) - \mathrm{err}(\mathcal{A}, \Pi_{p,N}, \mathcal{F})$$

where $\mathrm{err}\left(\mathcal{A}, \Pi^2_{p,N}, \mathcal{F}\right)$ is the marginal distribution over minority subpopulations X_2.

•Asymptotic regime

All metrics are evaluated with $\frac{N}{m} \to c$ as $N, m \to \infty$. Intuitively, c quantifies the hardness of the problem.

Main Theoretical Results

Theorem A (Informal) *For any $p \in (0, 1/2)$, consider*

- *Any distribution $\Pi_{p,N}$ where $\frac{N}{m} \to c$ as N, m goes to ∞.*
- *Any* sufficiently *entropic label prior $\mathcal{F}$:*
  i.e. $\max_{x\in X_2, y\in\mathcal{Y}} \mathbb{P}_{f\sim\mathcal{F}}[f(x) = y]$ is small.
- *Any (ϵ,δ)-DP algorithm $\mathcal{A}$ that* is *highly accurate.*

Then, the accuracy discrepancy is lower bounded as

$$\Gamma(\mathcal{A}, \Pi_{p,N}, \mathcal{F}) \gtrsim (1-p)\gamma_0$$

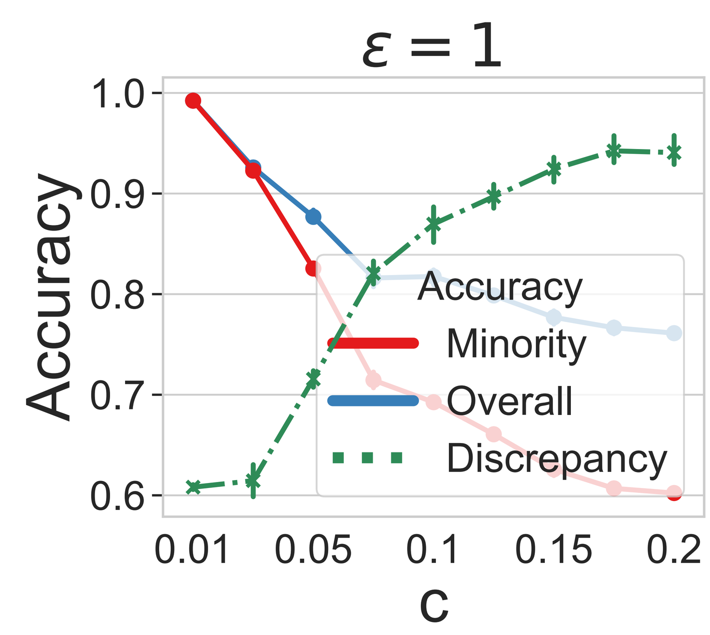
where γ_0, in the limit $c, m \to \infty$, increases as $1 - O\left(\epsilon e^{-c/\epsilon}/\sqrt{c}\right)$.

Theorem B (Informal) *When the private algorithm has low accuracy, then the more private the algorithm is, the fairer it is.*

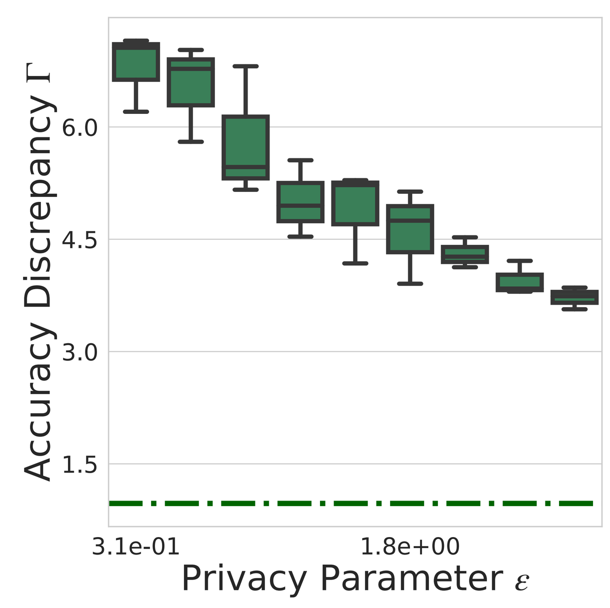
Experimental validation of Theorem A

Interpretation of Theorem A: For a private and accurate algorithm on data with subpopulations:
Unfairness (accuracy discrepancy) increases ↑ as

- **Privacy** increases i.e. $\epsilon \downarrow$
- Relative **number of subpopulations** increases i.e. $c \uparrow$.



Synthetic data



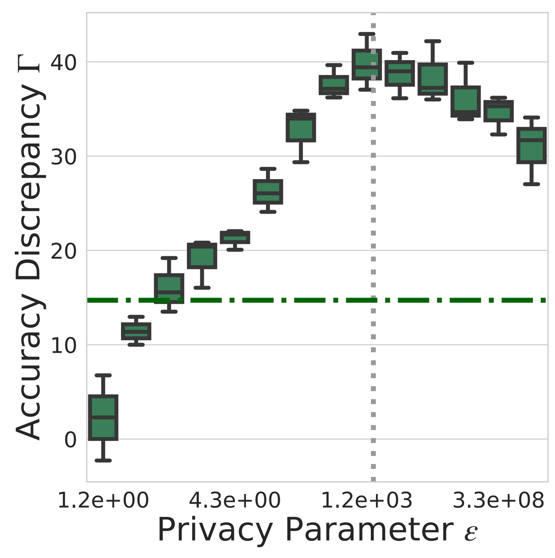
CELEBA (Acc≥ 0.95)

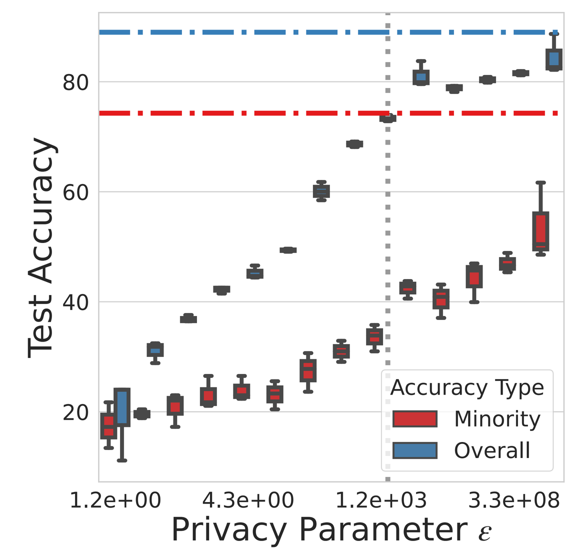
Experiments on real data

We conduct experiments using deep neural networks and Random forests on vision and tabular data respectively.

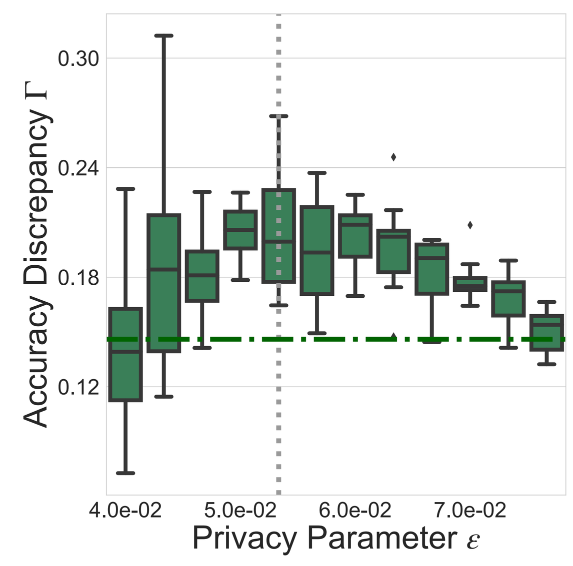
Similar trends across both cases support the universality of this phenomenon.

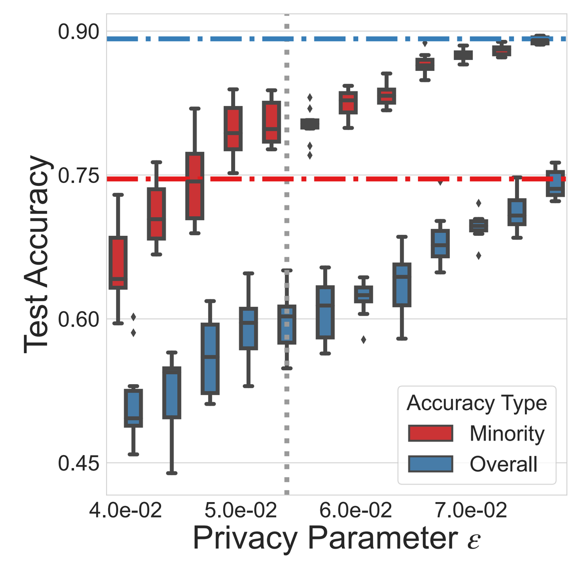
CIFAR10 with ResNet18





Law School with Random Forests





In both datasets, **Theorem A** explains to the right of the vertical dashed bar and **Theorem B** explains to the left.