How unfair is private learning?

Amartya Sanyal, Yaxi Hu, Fanny Yang
Privacy and Fairness
Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.
Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:
Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:


Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:


THIS WORK: The interaction of Privacy and Fairness of nearly accurate algorithms.
Differential Privacy
Differential Privacy

Neighbouring Datasets

- 0 0 0 0 0 0 0
- 1 1 1 1 1 1 1
- 2 2 2 2 2 2 2
- 3 3 3 3 3 3 3

- 0 0 0 0 0 0 0
- 1 1 1 1 1 1 1
- 2 2 2 2 2 2 2
- 3 3 3 3 3 3 3

- 0 0 0 0 0 0 0
- 1 1 1 1 1 1 1
- 2 2 2 2 2 2 2
- 3 3 3 3 3 3 3

- 0 0 0 0 0 0 0
- 1 1 1 1 1 1 1
- 2 2 2 2 2 2 2
- 3 3 3 3 3 3 3
Differential Privacy

Neighbouring Datasets
Differential Privacy

Neighbouring Datasets

$(\epsilon, \delta) \rightarrow$ DP Algorithm

$(\epsilon, \delta) \rightarrow$ DP Algorithm
Differential Privacy

Neighbouring Datasets

$(e, \delta) \rightarrow$ DP Algorithm

$(e, \delta) \rightarrow$ DP Algorithm
Differential Privacy

Neighbouring Datasets

Model 1

$\epsilon, \delta \rightarrow \text{DP Algorithm}$

$\epsilon, \delta \rightarrow \text{DP Algorithm}$
Differential Privacy

Model 1

Model 2

Neighbouring Datasets
Differential Privacy

$(\epsilon, \delta) - \text{DP Algorithm}$

Neighbouring Datasets

Model 1

Basically same

Model 2
Differential Privacy

Model 1

Model 2

Basically same

Neighbouring Datasets

\((\epsilon, \delta) \rightarrow \text{DP Algorithm}\)
Differential Privacy

Basically same

Model 1

Model 2
(Un) Fairness (Accuracy Discrepancy)
(Un) Fairness (Accuracy Discrepancy)
(Un) Fairness (Accuracy Discrepancy)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Proportion</th>
<th>B&amp;W</th>
<th>Mimes</th>
<th>Silent</th>
<th>Puppet</th>
<th>Ostern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrillers</td>
<td>40%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Superhero</td>
<td>40%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Genre: B&W, Mimes, Silent, Puppet, Ostern
## (Un) Fairness (Accuracy Discrepancy)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Thrillers</th>
<th>Superhero</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>40%</td>
<td>B&amp;W: 4%</td>
</tr>
</tbody>
</table>

- **40%** of the total population are Thrillers.
- **40%** of the total population are Superhero.
- *Minority subpopulations* include B&W, Mimes, Silent, Puppet, and Ostern, each comprising 4% of the population.
### (Un) Fairness (Accuracy Discrepancy)

<table>
<thead>
<tr>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genre</strong></td>
<td><strong>Proportion</strong></td>
</tr>
<tr>
<td>Thrillers</td>
<td>40%</td>
</tr>
<tr>
<td>Superhero</td>
<td>40%</td>
</tr>
<tr>
<td>B&amp;W</td>
<td>4%</td>
</tr>
<tr>
<td>Mimes</td>
<td>4%</td>
</tr>
<tr>
<td>Silent</td>
<td>4%</td>
</tr>
<tr>
<td>Puppet</td>
<td>4%</td>
</tr>
<tr>
<td>Ostern</td>
<td>4%</td>
</tr>
</tbody>
</table>

40% of the population consists of Majority subpopulations, and 4% of the population consists of Minority subpopulations.
ML Problem: Is the movie safe to watch for kids?

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrillers</td>
<td>40%</td>
<td>B&amp;W 4%</td>
</tr>
<tr>
<td>Superhero</td>
<td>40%</td>
<td>Mimes 4%</td>
</tr>
</tbody>
</table>

Proportion: 40% for both Majority and Minority subpopulations.

(Un) Fairness (Accuracy Discrepancy)
### ML Problem: Is the movie safe to watch for kids?

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thrillers</td>
<td>B&amp;W</td>
</tr>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Superhero</td>
<td>Mimes</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Silent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Puppet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ostern</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td>Error</td>
<td>5%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
</tr>
</tbody>
</table>
ML Problem: Is the movie safe to watch for kids?

(Un) Fairness (Accuracy Discrepancy)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thrillers</td>
<td>B&amp;W</td>
</tr>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td>Error</td>
<td>5%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>Superhero</td>
<td>Mimes</td>
</tr>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td>Error</td>
<td>5%</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Silent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Puppet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ostern</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
</tr>
</tbody>
</table>

Majority Error = 5%

Minority Error = 70%
### (Un) Fairness (Accuracy Discrepancy)

**ML Problem: Is the movie safe to watch for kids?**

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thrillers</td>
<td>B&amp;W</td>
</tr>
<tr>
<td></td>
<td>Superhero</td>
<td>Mimes</td>
</tr>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>4%</td>
</tr>
<tr>
<td>Error</td>
<td>5%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>75%</td>
</tr>
</tbody>
</table>

**Total Error = 18%**
## (Un) Fairness (Accuracy Discrepancy)

**ML Problem:** Is the movie safe to watch for kids?

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genre</strong></td>
<td>Thrillers</td>
<td>Superhero</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td><strong>Error</strong></td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

- **Majority Error** = 5%
- **Minority Error** = 70%

**Total Error** = 18%

**Accuracy Discrepancy** = Minority Error - Total Error
## (Un) Fairness (Accuracy Discrepancy)

**ML Problem:** Is the movie safe to watch for kids?

<table>
<thead>
<tr>
<th>Genre</th>
<th>Majority subpopulations</th>
<th>Minority subpopulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thrillers</td>
<td>Superhero</td>
</tr>
<tr>
<td>Proportion</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>Error</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

- **Majority Error** = 5%
- **Minority Error** = 70%
- **Total Error** = 18%

**Accuracy Discrepancy** = 70 - 18 = 52%
Example dataset
CelebA
Example dataset

CelebA
Example dataset
CelebA

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise
Example dataset
CelebA

40 binary attributes with each image

Eyeglasses
Bangs
Pointy Noise

40 binary attributes -> $2^{40}$ subpopulations.
Example dataset
CelebA

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

40 binary attributes -> $2^{40}$ subpopulations.

- Subpopulation 1: Eyeglasses, bangs, ..., pointy nose.
Example dataset
CelebA

40 binary attributes with each image

40 binary attributes \( \rightarrow 2^{40} \) subpopulations.

- **Subpopulation 1**: Eyeglasses, bangs, ..., pointy nose.
- **Subpopulation 2**: No eyeglasses, bangs, ..., pointy noise.
Example dataset

CelebA

40 binary attributes with each image

- Subpopulation 1: Eyeglasses, bangs, ..., pointy nose.
- Subpopulation 2: No eyeglasses, bangs, ..., pointy nose.
- ...
- ...
- Subpopulation $2^{40}$: No eyeglasses, no bangs, ..., no pointy nose.

40 binary attributes -> $2^{40}$ subpopulations.
Example dataset
CelebA

40 binary attributes with each image

40 binary attributes \(\rightarrow 2^{40}\) subpopulations.

• **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
• **Subpopulation 2**: No eyeglasses, bangs,……,pointy nose.
• ...
• ...
• **Subpopulation \(2^{40}\)**: No eyeglasses, no bangs,…, no pointy nose.
Example dataset
CelebA

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

40 binary attributes -> $2^{40}$ subpopulations.

- **Subpopulation 1**: Eyeglasses, bangs, ..., pointy nose.
- **Subpopulation 2**: No eyeglasses, bangs, ..., pointy nose.
  - ...
- **Subpopulation $2^{40}$**: No eyeglasses, no bangs, ..., no pointy nose.
Example dataset
CelebA

40 binary attributes with each image

• Subpopulation 1: Eyeglasses, bangs, ..., pointy nose.
• Subpopulation 2: No eyeglasses, bangs, ...., pointy noise.
• ...
• ...
• Subpopulation $2^{40}$: No eyeglasses, no bangs, ..., no pointy nose.

Eyeglasses

Bangs

Pointy Noise

40 binary attributes -> $2^{40}$ subpopulations.
Privacy vs Fairness
CelebA
Privacy vs Fairness
CelebA

![Graph showing test accuracy vs privacy parameter ε for CelebA dataset. The graph compares minority and overall accuracy types.]
Privacy vs Fairness

CelebA

More Private
Privacy vs Fairness
CelebA

More Private
Privacy vs Fairness

CelebA

More Private

More Private
Privacy vs Fairness
CelebA

More Private
Less Fair

More Private
Privacy vs Fairness
CelebA
Is this trend systematic?
Is this trend systematic?

Main Contribution:
We prove this trend in a model-agnostic setting
Is this trend systematic?

Main Contribution:
We prove this trend in a model-agnostic setting for long-tailed distribution.
Is this trend systematic?

Main Contribution:
We prove this trend in a model-agnostic setting for long-tailed distribution.
Definitions of error and fairness
Definitions of error and fairness

- Error
Definitions of error and fairness

- Error

\[ \text{err} (A, \Pi, F) = \]
Definitions of error and fairness

• Error

\[
\text{err} (A, \Pi, F) =
\]

Learning Algorithm
Definitions of error and fairness

- Error

\[
\text{err}(A, \Pi, F) = \]

Learning Algorithm

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$\text{err} (A, \Pi, F) =$

Learning Algorithm

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions \( \subseteq Y^X \)

- **Error**
  \[
  \text{err}(A, \Pi, F) = \mathbb{P}[h(x) \neq f(x)]
  \]

Learning Algorithm

Data Distribution

Probability is over \( S \sim \Pi^m, f \sim F, h \sim A(S_f), \) and \( x \sim \Pi_{p,N} \)
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\text{err} \left( A, \Pi, F \right) = \mathbb{P} \left[ h(x) \neq f(x) \right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\text{err}(A, \Pi, F) = \mathbb{P} [h(x) \neq f(x)]$$

Learning Algorithm

Data Distribution

Probability is over $S \sim \Pi^m[f \sim F, h \sim A(S_f), \text{and } x \sim \Pi_{p,N}]$
Definitions of error and fairness

Prior distribution over labelling functions \( \subseteq Y^X \)

- Error

\[
\text{err} (A, \Pi, F) = \mathbb{P} \left[ h(x) \neq f(x) \right]
\]

Learning Algorithm

Probability is over \( S \sim \Pi^m, f \sim F, h \sim A(S_f) \) and \( x \sim \Pi_{p,N} \)

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\text{err} (A, \Pi, F) = \mathbb{P} [h(x) \neq f(x)]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f), \text{and } x \sim \Pi_{p,N}$

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

• Error

\[ \text{err } (A, \Pi, F) = \mathbb{P} [h(x) \neq f(x)] \]

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution
Definitions of error and fairness

Prior distribution over labelling functions \( \subseteq Y^X \)

- Error

\[
\text{err} \left( A, \Pi, F \right) = \mathbb{P} \left[ h(x) \neq f(x) \right]
\]

- Accuracy Discrepancy

Probability is over \( S \sim \Pi^m, f \sim F, h \sim A(S_f), \text{ and } x \sim \Pi_{p,N} \)
Definitions of error and fairness

- Error

\[
\text{err} (A, \Pi, F) = \mathbb{P} [h(x) \neq f(x)]
\]

- Accuracy Discrepancy

\[
\Gamma (A, \Pi, F) = \text{err}_{\text{Minority}} (A, \Pi, F) - \text{err} (A, \Pi, F)
\]
Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- **Error**
  $$\text{err} (A, \Pi, F) = \mathbb{P} [h(x) \neq f(x)]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

- **Accuracy Discrepancy**
  $$\Gamma (A, \Pi, F) = \text{err}_{\text{Minority}} (A, \Pi, F) - \text{err} (A, \Pi, F)$$

Marginalised over minority subpopulations
Privacy at the cost of fairness
Privacy at the cost of fairness

Consider any $(\epsilon, \delta)$-DP algorithm that obtains low error on a long-tailed distribution.
Privacy at the cost of fairness

Consider any $(\epsilon, \delta)$-DP algorithm that obtains low error on a long-tailed distribution.

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter $\epsilon$. 
Privacy at the cost of fairness

Consider any $(\epsilon, \delta)$-DP algorithm that obtains low error on a long-tailed distribution.

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter $\epsilon$. 

$N$: # Minority subpopulations

$m$: # Training points
Privacy at the cost of fairness

Consider any $\left(\epsilon, \delta\right)$-DP algorithm that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \to c$ as $N, m \to \infty$. 

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter $\epsilon$. 

N: # Minority subpopulations  
m: # Training points
Privacy at the cost of fairness

Consider any \((\epsilon, \delta)\)-DP algorithm that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let \(\frac{N}{m} \rightarrow c\) as \(N, m \rightarrow \infty\).

\(N\): # Minority subpopulations
\(m\): # Training points

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter \(\epsilon\).
- (Long-tailed) Increases with (relative) # of minority subpopulations \(c\).
Privacy at the cost of fairness

Consider any $(\epsilon, \delta)$-DP algorithm that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \rightarrow c$ as $N, m \rightarrow \infty$. 

Informal Theorem A: We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter $\epsilon$.
- (Long-tailed) Increases with (relative) # of minority subpopulations $c$. 

$N$: # Minority subpopulations
$m$: # Training points
$F$: Label prior
Privacy at the cost of fairness

Consider any \((\varepsilon, \delta)\)-DP algorithm that obtains low error on a long-tailed distribution.

\[(\text{Minority Subpopulations}) \text{ Let } \frac{N}{m} \to c \text{ as } N, m \to \infty.\]

\[(\text{Label prior Entropy}) \text{ Define } \|F\|_\infty = \max_x \max_y \mathbb{P}_{f \sim F} [f(x) = y].\]

**Informal Theorem A** We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter \(\varepsilon\).
- (Long-tailed) Increases with (relative) \# of minority subpopulations \(c\).
Privacy at the cost of fairness

Consider any \((\epsilon, \delta)\)-DP algorithm that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let \(\frac{N}{m} \to c\) as \(N, m \to \infty\).

(Label prior Entropy) Define \(\|F\|_{\infty} = \max_{x,y} \mathbb{P}_{f \sim F}[f(x) = y]\) as # Minority subpopulations : # Training points.

\(F\) : Label prior

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter \(\epsilon\).
- (Long-tailed) Increases with (relative) # of minority subpopulations \(c\).
- (Label prior) Increases with entropy of the label prior.
Thank you