Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, Michaël Perrot, Aurélien Bellet, Marc Tommasi
Magnet Team

Workshop @ Comète on Ethical AI

November 23, 2023
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
Notations

• Features space $\mathcal{X}$, labels space $\mathcal{Y} = \{-1, 1\}$, and sensitive attributes space $\mathcal{S}$.
  • Features: images, tabular data, graphs, . . .
  • Labels: hired/not hired, profession, disease, . . .
  • Sensitive attributes: gender, race, age, . . .

• Decision function $h : \mathcal{X} \mapsto \mathbb{R}$ taken in a set $\mathcal{H}$.

• Binary decision function $H(x) = \begin{cases} 1 & \text{if } h(x) > 0, \\ -1 & \text{otherwise}. \end{cases}$

• A set $D = \{(x_i, s_i, y_i)\}_{i=1}^n$ of $n$ examples drawn i.i.d. a distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Goal: Measure the fairness of the model $H$. 
Accuracy Parity [Zafar et al., 2017]

For all $s \in S$, $F_s(h, D) = P(H(X) = Y | S = s) - P(H(X) = Y)$

- Accuracy of $h$ on the group $s$
- Accuracy of $h$. 

M. Perrot
November 23, 2023 4 / 16
Accuracy Parity [Zafar et al., 2017]

For all $s \in S$, $F_s(h, D) = P(H(X) = Y | S = s) - P(H(X) = Y)$

Accuracy of $h$ on the group $s$  

Accuracy of $h$. 

$P(H(X) = Y | S = \text{red}) = 1$

$P(H(X) = Y) = \frac{9}{10}$

$F_{\text{red}}(h, D) = \frac{1}{10}$
Accuracy Parity [Zafar et al., 2017]

For all $s \in S$, $F_s(h, D) = \underbrace{\Pr(H(X) = Y | S = s)} - \Pr(H(X) = Y)$

Accuracy of $h$ on the group $s$  
Accuracy of $h$.

$\Pr(H(X) = Y | S = \text{red}) = 1$

$\Pr(H(X) = Y) = 1$

$F_{\text{red}}(h, D) = 0$
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
A Story of Squirrels
Differential Privacy [Dwork, 2006]

A randomized algorithm \( A^{\text{priv}} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \rightarrow \mathcal{H} \) is \((\epsilon, \delta)\)-private if for all neighboring datasets \( D, D' \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \) and all subsets of hypotheses \( \mathcal{H}' \subseteq \mathcal{H} \)

\[
\mathbb{P}(A^{\text{priv}}(D) \in \mathcal{H}') \leq \exp(\epsilon) \mathbb{P}(A^{\text{priv}}(D') \in \mathcal{H}') + \delta
\]
Compute the $\ell_2$-sensitivity of the algorithm $A$:

$$\Delta(A) = \sup_{D \approx D'} \| A(D) - A(D') \|_H$$

Add noise to $A(D)$, calibrated to its sensitivity and the desired level of privacy:

$$A_{\text{priv}}(D) = A(D) + N(0, \frac{\Delta(A)^2}{2 \log(1.25/\delta)})$$
Gaussian Mechanism

Compute the $\ell_2$-sensitivity $\Delta$ of the algorithm $A$:

$$
\Delta(A) = \sup_{D \approx D'} \| A(D) - A(D') \|_H
$$

Add noise to $A(D)$, calibrated to its sensitivity and the desired level of privacy:

$$
A^{\text{priv}}(D) = A(D) + \mathcal{N}\left(0, \frac{2\Delta(A)^2 \log(1.25/\delta)}{\epsilon^2} \mathbb{I}_p\right)
$$
Gaussian Mechanism

Compute the $\ell_2$-sensitivity of the algorithm $A$:

$$\Delta(A) = \sup_{D \approx D'} \| A(D) - A(D') \|_H$$

Add noise to $A(D)$, calibrated to its sensitivity and the desired level of privacy:

$$A_{\text{priv}}(D) = A(D) + \mathcal{N}\left(0, \frac{2\Delta(A)^2 \log(1.25/\delta)}{\epsilon^2} \mathbb{I}_p\right)$$
A Story of Squirrels
A Story of Squirrels
What is the impact of differential privacy on fairness?
What is the impact of differential privacy on fairness?

\[ \mathbb{P}(H(X) = Y | S = \text{red}) = \frac{4}{5} \]

\[ \mathbb{P}(H(X) = Y) = \frac{9}{10} \]

\[ F_{\text{red}}(h, D) = -\frac{1}{10} \]
Pointwise Lipschitzness of Accuracy Parity

For all $s \in S$, $F_s(h, D) = \underbrace{\mathbb{P}(H(X) = Y \mid S = s)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbb{P}(H(X) = Y)}_{\text{Accuracy of } h}$.

Let $h, h' \in \mathcal{H}$ be two models

$$|F_s(h, D) - F_s(h', D)| \leq \chi_s(h, D) \|h - h'\|_{\mathcal{H}}$$

- $\chi_s(h, D) = \mathbb{E}\left(\frac{L_X}{|h(X)|} \mid h(X) \neq 0\right) + \mathbb{E}\left(\frac{L_X}{|h(X)|} \mid S = s\right)$
- $L_X$ is a lipschitz constant of the model: $|h(X) - h'(X)| \leq L_X \|h - h'\|_{\mathcal{H}}$
Pointwise Lipschitzness of Accuracy Parity

For all \( s \in S \), \( F_s(h, D) = \underbrace{P(H(X) = Y | S = s)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{P(H(X) = Y)}_{\text{Accuracy of } h} \)

Let \( h, h' \in \mathcal{H} \) be two models

\[
|F_s(h, D) - F_s(h', D)| \leq \chi_s(h, D) \|h - h'\|_\mathcal{H}
\]

- \( \chi_s(h, D) = \mathbb{E}\left(\frac{L_X}{|h(X)|} \right) + \mathbb{E}\left(\frac{L_X}{|h'(X)|} \mid S = s\right) \)
- \( L_X \) is a lipschitz constant of the model: \( |h(X) - h'(X)| \leq L_X \|h - h'\|_\mathcal{H} \)

**Key quantity:** Ratio between \( L_X \) (\( \downarrow \)) and the margin \( |h(X)| \) (\( \uparrow \)).
Pointwise Lipschitzness of Accuracy Parity

\[ A(D) \quad \text{and} \quad h(X) \]
Output Perturbation

\( \mathcal{A}(D) \) is the following \textbf{optimization problem}

\[
    h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; x_i, s_i, y_i)
\]

with \( \ell : \mathcal{H} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \to \mathbb{R} \) is \( \Lambda \)-lipschitz, and \( \mu \)-strongly-convex

\[
    \Delta(\mathcal{A}) = \frac{2\Lambda}{\mu n}
\]

Using the \textbf{Gaussian Mechanism}, we can then release

\[
    h^{\text{priv}} = h^* + \mathcal{N} \left( \frac{8\Lambda^2 \log(1.25/\delta)}{\mu^2 n^2 \epsilon^2} \mathbb{I}_p \right)
\]
Output Perturbation

\( \mathcal{A}(D) \) is the following optimization problem

\[
h^* = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; x_i, s_i, y_i)
\]

with \( \ell : \mathcal{H} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \to \mathbb{R} \) is \( \Lambda \)-lipschitz, and \( \mu \)-strongly-convex

\[
\Delta(\mathcal{A}) = \frac{2\Lambda}{\mu n}
\]

Using the Gaussian Mechanism, we can then release

\[
h^{\text{priv}} = h^* + \mathcal{N}\left(0, \frac{8\Lambda^2 \log(1.25/\delta)}{\mu^2 n^2 \epsilon^2} \mathbb{I}_p\right)
\]

**Key result:** With probability at least 1 \(- \zeta\), \( \|h^{\text{priv}} - h^*\|_2 \leq \frac{\Lambda \sqrt{32p \log(1.25/\delta) \log(2/\zeta)}}{\mu n \epsilon} \).
Bounded Distance

With probability at least $1 - \zeta$, $\| h^{\text{priv}} - h^* \|_2 \leq O\left(\frac{\sqrt{p}}{n\epsilon}\right)$.
Main result

- \(|F_s(h, D) - F_s(h', D)| \leq \chi_s(h, D) \|h - h'\|_H\)
- With probability \(1 - \zeta\), \(\|h^{\text{priv}} - h^*\|_2 \leq \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)\)

With probability \(1 - \zeta\), \(|F_s(h^{\text{priv}}, D) - F_s(h^*, D)| \leq \chi_s(h^{\text{priv}}, D) \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)\)
Main result

- \(|F_s(h, D) - F_s(h', D)| \leq \chi_s(h, D) \|h - h'\|_H\)
- With probability \(1 - \zeta\), \(\|h^{\text{priv}} - h^*\|_2 \leq O\left(\frac{\sqrt{p}}{n\epsilon}\right)\)

With probability \(1 - \zeta\), \(|F_s(h^{\text{priv}}, D) - F_s(h^*, D)| \leq \chi_s(h^{\text{priv}}, D)O\left(\frac{\sqrt{p}}{n\epsilon}\right)\)

**Auditing:** The right hand side only depends on the private model!
Conclusion

The story so far...

• Accuracy Parity is pointwise lipschitz and margins are key quantities.
• Using output perturbation the private and non-private models are close.
• Differential privacy has bounded impact on fairness.
• The squirrels are safe if they learned their model on sufficiently many examples.
Conclusion

The story so far...

- Accuracy Parity is **pointwise lipschitz** and **margins** are key quantities.
- Using output perturbation the **private and non-private models are close**.
- Differential privacy has **bounded impact** on fairness.
- The **squirrels are safe** if they learned their model on **sufficiently many examples**.

but also...

- Similar results for **other fairness measures** (e.g. Demographic Parity [Calders et al., 2009], Equality of Opportunity [Hardt et al., 2016], Equalized Odds [Hardt et al., 2016]) and for **Accuracy**.
- Multi-class, multi-groups problems, DP-SGD, **tighter but harder to parse bounds**.
- A few **experiments** to check the tightness of our bounds.
- A **finite sample analysis** showing that our results also hold in **generalization**.
You can fetch the paper on arXiv!
References

All animal images were taken from commons.wikimedia.org. No animals were harmed during the preparation of this presentation.


Experiments

Folktables [Ding et al., 2021], $\ell_2$-regularized logistic regression, $(\epsilon, \delta) = (1, \frac{1}{n^2})$

---

Theoretical Upper Bound
Bound with Empirical Distance
Improved Bound Knowing both Models
Non-private Model Fairness
Private Models Fairness
Experiments

Folktables [Ding et al., 2021], $\ell_2$-regularized logistic regression, $\delta = \frac{1}{n^2}$, $n = 1,498,050$

![Graph showing Accuracy and Accuracy Parity vs Value of $\epsilon$]

- Theoretical Upper Bound
- Bound with Empirical Distance
- Improved Bound Knowing both Models
- Non-private Model Fairness
- Private Models Fairness

<table>
<thead>
<tr>
<th>Value of $\epsilon$</th>
<th>Accuracy</th>
<th>Accuracy Parity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-2}$</td>
<td>0.70</td>
<td>0.00</td>
</tr>
<tr>
<td>$10^{0}$</td>
<td>0.72</td>
<td>0.01</td>
</tr>
<tr>
<td>$10^{1}$</td>
<td>0.74</td>
<td>0.02</td>
</tr>
<tr>
<td>$10^{2}$</td>
<td></td>
<td>0.04</td>
</tr>
</tbody>
</table>

M. Perrot
November 23, 2023 19 / 16