Inria
Université de Lille

Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, Michaël Perrot, Aurélien Bellet, Marc Tommasi
Inria Lille, France

Classifiers and Margin [1]

Feature space $\mathcal{X}$, sensitive attributes $\mathcal{S}$, labels $\mathcal{Y}$.
Decision function $h \in \mathcal{H} \subseteq \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

We classify $x \in \mathcal{X}$ as:

$$H(x) = \arg\max_{y \in \mathcal{Y}} h(x, y) \qquad \text{for } x \in \mathcal{X} \quad .$$

Confidence margin of h for label y on input x:

$$\boxed{\rho(h, x, y) = h(x, y) - \max_{y' \neq y} h(x, y')} \quad .$$

Summary

The difference of fairness between private and optimal models vanishes since:

1. Group fairness notions are pointwise Lipschitz.
2. Models learned by output perturbation or DP-SGD converge to non-private one at a rate $O(\sqrt{p}/n\epsilon)$.

Main Assumption: Lipschitz Margins

For $x, y \in \mathcal{X} \times \mathcal{Y}$, there exists $L_{x,y}$ such that for all $h, h' \in \mathcal{H}$

$$\boxed{\left\|\rho(h, x, y) - \rho(h', x, y)\right\| \leq L_{x,y} \left\|h - h'\right\|} \quad .$$

Group Fairness (Example of Equality of Opportunity [2])

Fairness level of $h \in \mathcal{H}$, for $(y, k) \in \mathcal{Y} \times \mathcal{S}$, for "desirable" labels y:

$$F_{(y,k)}(h, D) = \mathbb{P}(H(X) = Y \mid Y = y, S = k) - \mathbb{P}(H(X) = Y \mid Y = y) \quad .$$

(Equalized odds, accuracy parity, and demographic parity have similar expressions.)

Average fairness level: $Fair(h, D) = \frac{1}{|\mathcal{Y} \times \mathcal{S}|} \sum_{(y,k) \in \mathcal{Y} \times \mathcal{S}} F_{(y,k)}(h, D)$.

Private Empirical Risk Minimization [3]

Assume **strongly-convex** loss. Release an (ϵ, δ)-DP value:

$$h^{\text{priv}} \approx h^* \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{(x,s,y) \in D} \ell(h(x), y) \ , \qquad \text{(ERM)}$$

▶ Output Perturbation [3]:

$$h^{priv} = h^* + \mathcal{N}\left(0, O\left(\frac{p}{n^2 \epsilon^2}\right)\right) \quad .$$

▶ DP-SGD [4], compute for $t = 0 \dots, T-1$:

$$h^{t+1} = h^t - \eta\left(\left(\nabla f_i(h^t) + \mathcal{N}\left(0, O\left(\frac{pT}{n^2 \epsilon^2}\right)\right)\right)\right) \quad ,$$

and return $h^{priv} = h^T$.

In both cases: $\boxed{\left\|h^{priv} - h^*\right\| = O\left(\frac{\sqrt{p}}{n\epsilon}\right)}$ with high proba .

(Some) Group Fairness Notions are Pointwise Lipschitz

For $h, h' \subseteq \mathcal{H}$, and any event E: $\left|\mathbb{P}(H(X) = Y \mid E) - \mathbb{P}(H'(X) = Y \mid E)\right| \leq \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \middle| E\right) \|h - h'\|$.

$$\Longrightarrow \left|F_{(y,k)}(h, D) - F_{(y,k)}(h', D)\right| \leq \chi_{(y,k)}(h) \cdot \|h - h'\| \ , \text{ with } \boxed{\chi_{(y,k)}(h) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \middle| Y = y, S = k\right) + \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \middle| Y = y\right)} \quad .$$

Fairness Loss due to Privacy Vanishes in $O(\sqrt{p}/n\epsilon)$

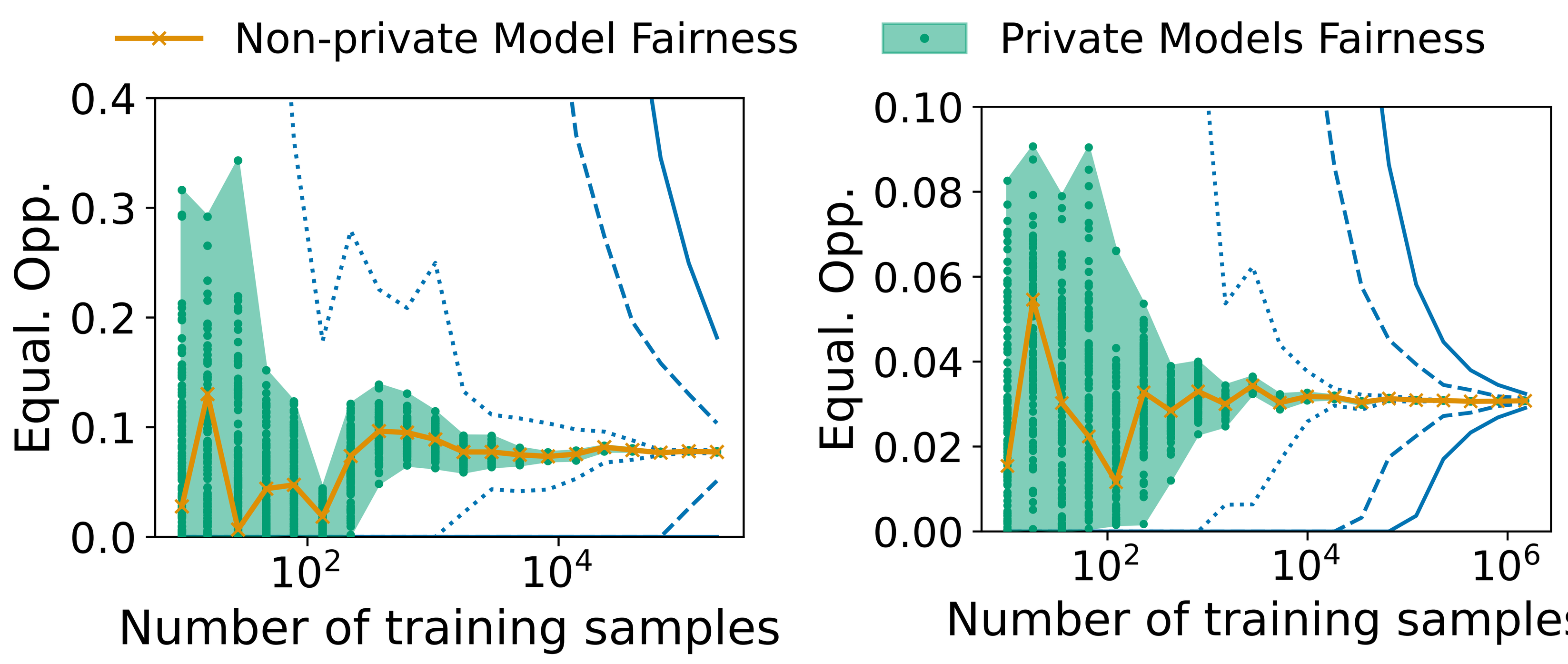
$$\left|F_{(y,k)}(h^*, D) - F_{(y,k)}(h^{priv}, D)\right| \leq \chi_{(y,k)}(h^{ref}) \cdot O(\sqrt{p}/n\epsilon) \ , \qquad \text{for } h^{ref} \in \{h^{priv}, h^*\} \ .$$

This guarantees that the fairness level of h^{priv} is close to the one of h^*, even when the latter is unknown.

Numerical Results on Logistic Regression

With $\epsilon = 1$, $\delta = \frac{1}{n^2}$ on `celebA` ($n = 182k$) and `folktables` ($n = 1,600k$).
Three variants of the bound, depending on knowledge of h, h':

- (solid line) knowing only theoretical bound on $\left\|h^{\text{priv}} - h^*\right\|$,
- (dashed line) knowing empirical value of $\left\|h^{\text{priv}} - h^*\right\|$,
- (dotted line) knowing actual values of h^{priv} and h^*.



(a) `celebA` (b) `folktables`

More generally

With $F_k(h, D) = C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \mathbb{P}(H(X) = Y \mid D_{k'})$,
(e.g. for equalized odds, accuracy parity, demographic parity...)

$$\chi_k(h) = \sum_{k'=1}^{K} C_k^{k'} \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \middle| D_{k'}\right) \quad .$$

References

[1] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. "Multi-Class Classification with Maximum Margin Multiple Kernel". In: *ICML*. 2013.
[2] Moritz Hardt et al. "Equality of Opportunity in Supervised Learning". In: *NeurIPS*. 2016.
[3] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. "Differentially Private Empirical Risk Minimization". In: *JMLR* (2011).
[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. "Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds". In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. 2014.