ON THE IMPACT OF DIFFERENTIAL PRIVACY ON FAIRNESS IN MACHINE LEARNING

Aurélien Bellet (Inria, France)
Joint work with Paul Mangold, Michaël Perrot and Marc Tommasi

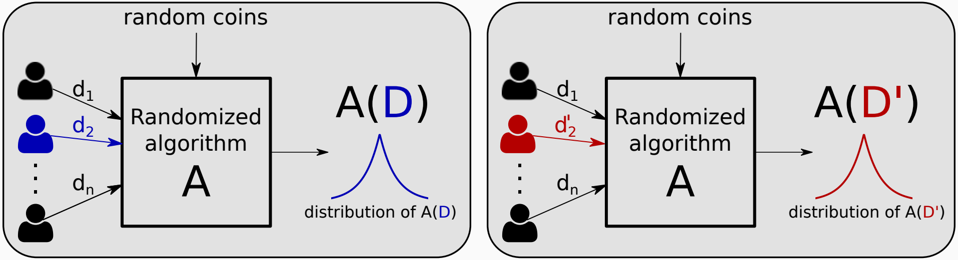
Workshop on Privacy and Fairness in AI for Health — AI UK 2023
March 27, 2023

- **Privacy** and **fairness** are two critical concerns when AI systems are deployed in high-stakes applications like health (e.g., for AI-assisted medical diagnosis)
    - The prediction model should not leak sensitive information about individuals whose data was used to train the model
    - Model predictions should not unjustly discriminate against some individuals or subgroups of the population
- Unfortunately, privacy and fairness are sometimes conflicting objectives [Bagdasaryan et al., 2019, Cummings et al., 2019, Chang and Shokri, 2020, Tran et al., 2021]
    - Privacy: “prevent the model from learning too much about a single individual”
    - Fairness: “make sure that underrepresented individuals have sufficient weight”
- **Our work:** provably bound the impact of privacy on fairness in classification, and uncover some of the key factors that govern this impact

- We consider a multi-class classification problem with a feature space $\mathcal{X}$, a finite set of labels $\mathcal{Y}$, and a finite set of sensitive attributes $\mathcal{S}$
- We denote by $\mathcal{D}$ the data distribution of variables (X, Y, Z) over $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$
- We denote by $D = \{(x_1, s_1, y_1), \ldots, (x_n, s_n, y_n)\}$ the training set of n examples drawn i.i.d. from $\mathcal{D}$
- Let $h : \mathcal{X} \to \mathcal{Y}$ be a model that predicts a label $h(x) \in \mathcal{Y}$ from features $x \in \mathcal{X}$

GROUP FAIRNESS

- We focus on group fairness, which requires that decisions made by machine learning models do not unjustly discriminate against subgroups of the population
  - Accuracy parity:
$$\Pr[h(X) = Y \mid S = s] = \Pr[h(X) = Y]$$
  - Equality of opportunity (assuming $Y = 1$ is the desirable outcome):
$$\Pr[h(X) = Y \mid Y = 1, S = s] = \Pr[h(X) = Y \mid Y = 1]$$
  - Also demographic parity and equalized odds
- Our results are general and hold for these 4 classic group fairness measures
- Given a partition of D into K groups $D_1, \ldots, D_K$, we will use $F_k(h)$ to denote the fairness level of model h for group k (when $F_k(h) < 0$, group k is disadvantaged)



- Differential Privacy (DP) requires that the distribution of outputs should be "similar" for two neighboring datasets $D = \{d_1, d_2, d_3, \ldots, d_n\}$ and $D' = \{d_1, d'_2, d_3, \ldots, d_n\}$
- Formally, for $\epsilon > 0$ and $\delta \in (0, 1)$, $\mathcal{A}$ satisfies (ϵ, δ)-DP if for all pairs of neighboring datasets D and D', and all $S \subseteq \text{range}(\mathcal{A})$, we have:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D') \in S] + \delta$$

TRAINING A MODEL WITH DIFFERENTIAL PRIVACY

- Consider the classic Empirical Risk Minimization (ERM) framework:

$$h^* = \text{arg min}_{h \in \mathcal{H} \subseteq \mathbb{R}^p} \left\{ f(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h; x_i, s_i, y_i) \right\}$$

- Differential privacy requires to add some noise to the model
- Output perturbation [Chaudhuri et al., 2011]: $h^{\text{priv}} = h^* + \mathcal{N}(0, \sigma^2 \mathbb{I}_p)$
- Differentially Private SGD [Bassily et al., 2014, Abadi et al., 2016]: iterate over

$$h^{t+1} = h^t - \gamma(\nabla \ell(h^t; x_i, si, y_i) + \mathcal{N}(0, \sigma^2 \mathbb{I}_p))$$

- In both cases, we know how to choose σ to achieve the desired (ϵ, δ)-DP guarantee (under suitable assumptions)

- Previous work has empirically shown that differential privacy can exacerbate unfairness, see e.g. the results below for accuracy parity [Bagdasaryan et al., 2019]

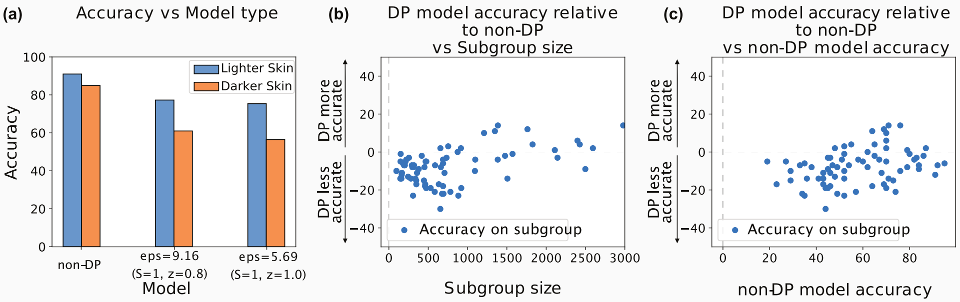


Figure 1: Gender and age classification on facial images.

- **Question:** when does this happen? how bad can it get?

SIMILAR MODELS HAVE SIMILAR FAIRNESS LEVELS!

Theorem (Pointwise Lipschitzness of group fairness)

For any two models $h, h' \in \mathcal{H}$, we have, for all $k \in [K]$,

$$|F_k(h) - F_k(h')| \leq \sum_{k'=1}^{K} |C_k^{k'}| \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \,\middle|\, D_{k'} \right) \|h - h'\|_{\mathcal{H}},$$

where $\rho(h,X,Y)$ is the confidence margin, $L_{X,Y}$ is the Lipschitz constant of $\rho(h,X,Y)$, and the C's are constants independent of h and h'.

- If two models h and h' are close, then their fairness levels are similar
- The difference in fairness is smaller if h is confident in its prediction for the true label

Theorem (Fairness loss due to privacy)

Let the loss function ℓ be Λ-Lipschitz and μ-strongly convex. Let h^* be the optimal model, and h^{priv} its private estimate obtained by output perturbation. Let $h^{\mathrm{ref}} \in \{h^{\mathrm{priv}}, h^*\}$. Then, for all $k \in [K]$ and any $0 < \zeta < 1$, we have with probability at least $1 - \zeta$:

$$|F_k(h^{\mathrm{priv}}) - F_k(h^*)| \leq \frac{\chi_k(h^{\mathrm{ref}})\Lambda\sqrt{32p\log(1.25/\delta)\log(2/\zeta)}}{\mu n \epsilon},$$

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

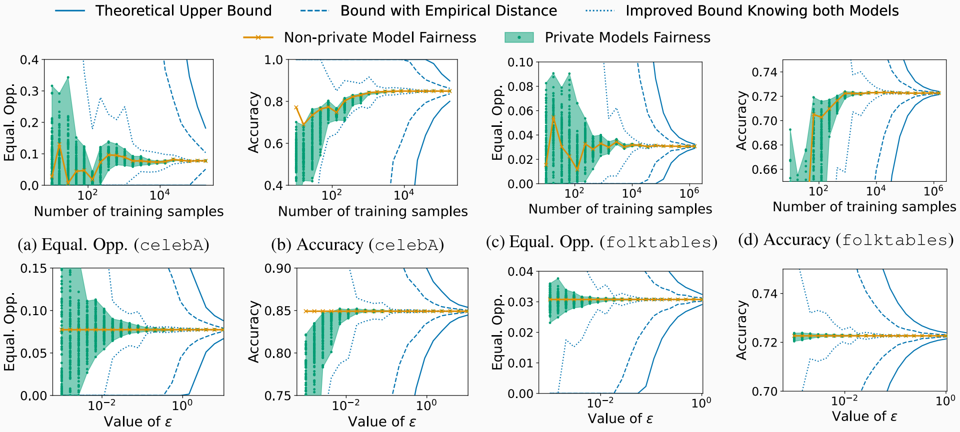
- The unfairness due to privacy vanishes at a $\widetilde{O}(\sqrt{p}/n)$ rate!
- A similar result holds for DP-SGD
- Note: we can pick the “reference” model to be either h^{priv} or h^* (e.g., depending on which model is known)

- For sufficiently large datasets, our bound gives useful guarantees

Table 1. Upper bound, with 99% probability, on the difference of fairness between private and non-private models for different fairness measures and accuracy. Privacy parameters are $\epsilon = 1$ and $\delta = 1/n^2$ where n is the number of samples in the training data.

Dataset	Equality of Opportunity	Equalized Odds	Demographic Parity	Accuracy Parity	Accuracy
celebA ($n = 182,339$)	0.1044	0.0975	0.0975	0.0975	0.0487
folktables ($n = 1,498,050$)	0.0017	0.0026	0.0026	0.0026	0.0013

EMPIRICAL ILLUSTRATIONS



(a) Equal. Opp. (celebA) (b) Accuracy (celebA) (c) Equal. Opp. (folktables) (d) Accuracy (folktables)

(e) Equal. Opp. (celebA) (f) Accuracy (celebA) (g) Equal. Opp. (folktables) (h) Accuracy (folktables)

- Our bounds appear to capture the right dependence in p and n
- With additional knowledge on models, we can get quite tight bounds

Take-home messages

- We can bound the impact of differential privacy on the fairness of classifiers
- The fairness loss due to privacy depends on the size of the training set, the number of model parameters, and the confidence margin of the model

Perspectives

- Apply our results to other privacy-preserving methods (and beyond): one only needs to derive a high-probability bound on the distance between the models of interest
- Extensions to nonconvex settings (what should the reference model be?)
- Design fairer privacy-preserving algorithms: combine our results with fairness-promoting regularizers [Lohaus et al., 2020], privately learn models with large-margin guarantees [Bassily et al., 2022]

THANK YOU FOR YOUR ATTENTION!

QUESTIONS?

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016).
Deep Learning with Differential Privacy.
In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA. Association for Computing Machinery.
1130 citations (Crossref) [2022-08-19].

[Agarwal, 2020] Agarwal, S. (2020).
Trade-offs between fairness and privacy in machine learning.

[Bagdasaryan et al., 2019] Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019).
Differential privacy has disparate impact on model accuracy.
In *Advances in Neural Information Processing Systems 32, 2019, Vancouver, BC, Canada*, pages 15453–15462.

[Bassily et al., 2022] Bassily, R., Mohri, M., and Suresh, A. T. (2022).
Differentially private learning with margin guarantees.
arXiv preprint arXiv:2204.10376.

[Bassily et al., 2014] Bassily, R., Smith, A., and Thakurta, A. (2014).
Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds.
In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, Philadelphia, PA, USA. IEEE.

[Chang and Shokri, 2020] Chang, H. and Shokri, R. (2020).
On the privacy risks of algorithmic fairness.
arXiv preprint arXiv:2011.03731.

[Chaudhuri et al., 2011] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011).
Differentially Private Empirical Risk Minimization.
Journal of Machine Learning Research, 12(29):1069–1109.

[Cummings et al., 2019] Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019).
On the compatibility of privacy and fairness.
In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315.

[Farrand et al., 2020] Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020).
Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy.
arXiv preprint arXiv:2009.06389.

[Jagielski et al., 2019] Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019).
Differentially private fair learning.
In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

[Kilbertus et al., 2018] Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. (2018).
Blind justice: Fairness with encrypted sensitive attributes.
In *International Conference on Machine Learning*, pages 2630–2639. PMLR.

[Lohaus et al., 2020] Lohaus, M., Perrot, M., and Von Luxburg, U. (2020).
Too relaxed to be fair.
In *International Conference on Machine Learning*, pages 6360–6369. PMLR.

[Mozannar et al., 2020] Mozannar, H., Ohannessian, M., and Srebro, N. (2020).
Fair learning with private demographic data.
In *International Conference on Machine Learning*, pages 7066–7075. PMLR.

[Pujol et al., 2020] Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020).
Fair decision making using privacy-protected data.
In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 189–199, New York, NY, USA. Association for Computing Machinery.

[Sanyal et al., 2022] Sanyal, A., Hu, Y., and Yang, F. (2022).
How unfair is private learning?
arXiv preprint arXiv:2206.03985.

[Tran et al., 2021] Tran, C., Dinh, M., and Fioretto, F. (2021).
Differentially private empirical risk minimization under the fairness lens.
Advances in Neural Information Processing Systems, 34:27555–27565.

[Tran et al., 2020] Tran, C., Fioretto, F., and Van Hentenryck, P. (2020).
Differentially private and fair deep learning: A lagrangian dual approach.
arXiv preprint arXiv:2009.12562.

[Uniyal et al., 2021] Uniyal, A., Naidu, R., Kotti, S., Singh, S., Kenfack, P. J., Mireshghallah, F., and Trask, A. (2021).
Dp-sgd vs pate: Which has less disparate impact on model accuracy?
arXiv preprint arXiv:2106.12576.

[Xu et al., 2020] Xu, D., Du, W., and Wu, X. (2020).
Removing disparate impact of differentially private stochastic gradient descent on model accuracy.
arXiv preprint arXiv:2003.03699.

[Xu et al., 2019] Xu, D., Yuan, S., and Wu, X. (2019).
Achieving differential privacy and fairness in logistic regression.
In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599.

RELATED WORK

- Empirical evidence that privacy can exacerbate unfairness [Bagdasaryan et al., 2019] [Pujol et al., 2020, Farrand et al., 2020, Uniyal et al., 2021], and that enforcing fairness can lead to more privacy leakage for the unprivileged group [Chang and Shokri, 2020]
- Approaches to learn models that are both fair and privacy-preserving have limited guarantees [Kilbertus et al., 2018, Xu et al., 2019, Xu et al., 2020, Tran et al., 2020] and/or require stochastic decisions [Jagielski et al., 2019, Mozannar et al., 2020]
- Incompatibility results [Sanyal et al., 2022, Cummings et al., 2019, Agarwal, 2020] consider unrealistic cases that are hardly encountered in practice
- [Tran et al., 2021] analyze the impact of privacy on fairness in ERM, but only in terms of loss-based fairness and via loose Taylor approximations

GENERAL FORM FOR GROUP FAIRNESS

- Assume that the data D can be partitioned into K disjoint groups denoted by $D_1, \ldots, D_K$ (based on the sensitive attribute and possibly the label)
- Our results hold for any fairness measure that, for each group $k = 1, \ldots, K$, can be written as

$$F_k(h) = C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \Pr[H(X) = Y \mid D_{k'}]$$

- See the paper for the derivation of the 4 classic group fairness measures from this general formula