Learning Fair Scoring Functions

Robin Vogel[1,2] Aurélien Bellet[3] Stephan Clémençon[2]

[1] IDEMIA, [2] Télécom Paris, [3] Inria

24/03/2020

Outline

Fairness in algorithmic decisions

Algorithmic decisions are increasingly used in many domains:

Banking (*e.g.* loans) Recruting (*e.g.*, hiring)

Insurance (*e.g.* cars) Judiciary (*e.g.*, bail)

Recently, the fairness of algorithms has gathered lots of attention.

e.g. May 2016: The COMPAS system assesses the likelihood of recidivism of a defendant for U.S. courts.

While algorithms are usually designed for the interest of some user, fair algorithms suggests confronting those to the law.

"Predictive models are really just opinions embedded in math."

Cathy O'Neil.

Fairness in classification

Binary classification: $(X, Y) \sim P$ and $(X, Y) \in \mathcal{X} \times \{-1, 1\}$,
learn a classifier $g : \mathcal{X} \to \{-1, 1\}$ from data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$.

Fairness: Sensitive information $Z \in \{0, 1\}$, a Z_i for each (X_i, Y_i).
e.g. gender, ethnicity, ...

Fairness without ground truth: Parity in ...

· Treatment: $g(X, Z) = g(X)$ almost surely.
i.e. the decision does not depend on the sensitive attribute.

· Impact: $\mathbb{P}\{g(X) = +1 | Z = 0\} = \mathbb{P}\{g(X) = +1 | Z = 1\}$.

Fairness with ground truth: Parity in ...

· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$,

· FPR: $\mathbb{P}\{g(X) = 1 \mid Z = 0, Y = -1\} = \mathbb{P}\{g(X) = 1 \mid Z = 1, Y = -1\}$,

· TPR, ...

Related work

Fairness of ML has gathered lots of attention recently.

In binary classification:

· A flexible approach for relaxed constraints [Zafar et al., 2019a],

· ERM guarantees [Donini et al., 2018],

Other notable works:

· Textbook (WIP) on fairness in ML [Barocas et al., 2019],

· "Adversarially fair" representations [Madras et al., 2018].

Fairness in ranking became only recently a research topic, mostly tackled by the information retrieval (IR) community.

Some authors:

· modify a fixed score to induce a notion of fairness [Zehlike et al., 2017, Biega et al., 2018],

· introduce fairness in exposure over several rankings [Singh and Joachims, 2018, Singh and Joachims, 2019],

· use a notion of fairness based on the AUC [Borkan et al., 2019, Beutel et al., 2019].

Bipartite ranking (1/2)

Scoring: $(X, Y) \sim P$ and $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$,
learn a score $s : \mathcal{X} \to \mathbb{R}$ from data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$.

Objective: Order new elements $X'_1, \ldots, X'_m$ by relevance,
i.e. by decreasing posterior probability $\eta(x) := \mathbb{P}\{Y = +1 \mid X = x\}$.

Perf. measure: The ROC curve: the true positive rate (TPR) for any false positive rate (FPR) for testing $Y = +1$ with $s(X) > t$.

Introduce the distributions (cdf) of $s(X)|Y = -1$ and $s(X)|Y = +1$ as:

$$H_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = -1\} \quad \text{and} \quad G_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = +1\}.$$

In the context of **fairness**, we denote by:
· $H_s^{(z)}$ the cdf of $s(X)|Y = -1, Z = z$,
· $G_s^{(z)}$ the cdf of $s(X)|Y = +1, Z = z$,
for any $z \in \mathcal{Z}$ with $\mathcal{Z} = \{0, 1\}$.

Bipartite ranking (2/2)

Let $\bar{F} = 1 - F$ and define the pseudo-inverse of F as:

$$F^{-1} : u \mapsto \inf\{t \mid F(t) > u\}.$$

The FPR (resp. TPR) of s at threshold t is equal to $\overline{H}_s(t)$ (resp. $\overline{G}_s(t)$).
Formally, the ROC and AUC write:

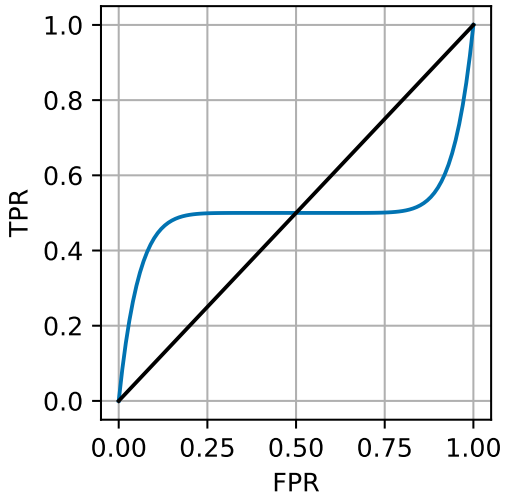
$$\mathrm{ROC}_{H_s,G_s}(\alpha) = \overline{G}_s \circ \overline{H}_s^{-1}(\alpha) \quad \text{and} \quad \mathrm{AUC}_{H_s,G_s} = \int_0^1 \mathrm{ROC}_{H_s,G_s}(\alpha)\, d\alpha.$$

The ROC **measures the difference** between two cdfs in $\mathbb{R}$.

Specifically, given two distributions F, F' on $\mathbb{R}$:

$$\forall \alpha \in [0,1], \quad \mathrm{ROC}_{F,F'}(\alpha) = \alpha \quad \Leftrightarrow \quad F = F'.$$

The AUC is a **scalar summary** of the ROC.



Our contributions

We focus on fairness for bipartite ranking, and provide:

- A **general formulation** for AUC-based fairness constraints,
- **Guarantees** for learning under AUC-based constraints,
- A gradient descent (GD) **method** for learning w/ AUC constraints,
- A new, restrictive type of **constraint**: ROC-based constraints,
- **Guarantees** and a **GD method** for learning with ROC constraints.

Outline

A general AUC constraint (1/2)

Intra-group pairwise and BNSP AUC fairness ([Borkan et al., 2019]):

$$\mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} = \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}},$$
$$\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}},$$

and many many more…

Introduce all relevant distributions as $D(s) = (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})$.
Any known AUC constraint writes as:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)},$$

with $\alpha, \alpha', \beta, \beta' \in [0,1]^4$ and any of those sums to 1.

A general AUC constraint (2/2)

$$\mathrm{AUC}_{\alpha^\top D(s),\beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s),\beta'^\top D(s)}, \tag{1}$$

Theorem 1

The following propositions are equivalent:

1. *Eq.* (1) *is verified when* $X|Y = y, Z = 0$ *and* $X|Y = y, Z = 1$ *have same dist. for any* $y \in \mathcal{Y}$ *and* $\eta(X)$ *is not a.s constant.*
2. $(e_1 + e_2)^\top[(\alpha - \alpha') - (\beta - \beta')] = 0.$
3. *Eq.* (1) *is equivalent to* $\Gamma^\top C(s) = 0$ *where* $\Gamma \in \mathbb{R}^5$ *and* $C(s) \in \mathbb{R}^5$ *are 5 elementary fairness measures.*

Theorem 1 shows that most AUC constraints are summarized with 5 coefficients. The estimation of the $C(s)$ is straightforward.

Learning with AUC constraints

Let $\mathcal{S}$ be a proposal family of scores. With an example constraint, integrating the constraint as a penalty gives, where $\lambda > 0$ is fixed:

$$\max_{s \in \mathcal{S}} L_\lambda(s) \quad \text{with} \quad L_\lambda(s) = \mathrm{AUC}_{H_s, G_s} - \lambda \left| \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}} \right|,$$

and its solution is written s^*_λ.

The empirical counterpart $\widehat{L}_\lambda$ of L_λ replaces the AUC's above by estimators, using the empirical counterparts of H_s, $H_s^{(z)}$, G_s, $G_s^{(z)}$ on the sample $\mathcal{D}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$. Its maximizer is written $\widehat{s}_\lambda$.

Theorem 2

Assume that $\mathcal{S}$ is VC-major with VC-dim $V < +\infty$,
and there exists $\epsilon > 0$, $\epsilon \leq \mathbb{P}\{Y = y, Z = z\}$ for any $y \in \mathcal{Y}, z \in \mathcal{Z}$.
Then, for any $\delta > 0$ and $n > 1$, with probability $\geq 1 - \delta$:

$$\epsilon^2 \left[L_\lambda(s^*_\lambda) - L_\lambda(\widehat{s}_\lambda) \right] \leq C \sqrt{\frac{V}{n}} + (8\lambda + 2\epsilon) \sqrt{\frac{\log(13/\delta)}{n-1}} + O(n^{-1}).$$

Optimizing L_λ by gradient descent

$\widehat{L}_\lambda$ is not continuous, to relax it:

We replace $x \mapsto \mathbb{I}\{x \geq 0\}$ by a logistic $\sigma : x \mapsto 1/(1+e^{-x})$.

Introduce a parameter $c \in [-1, +1]$, our relaxed objective writes:

$$\widetilde{L}_\lambda(s) := \widetilde{\mathrm{AUC}}_{H_s, G_s} - \lambda \cdot c \left(\widetilde{\mathrm{AUC}}_{H_s^{(1)}, G_s^{(1)}} - \widetilde{\mathrm{AUC}}_{H_s^{(0)}, G_s^{(0)}} \right).$$

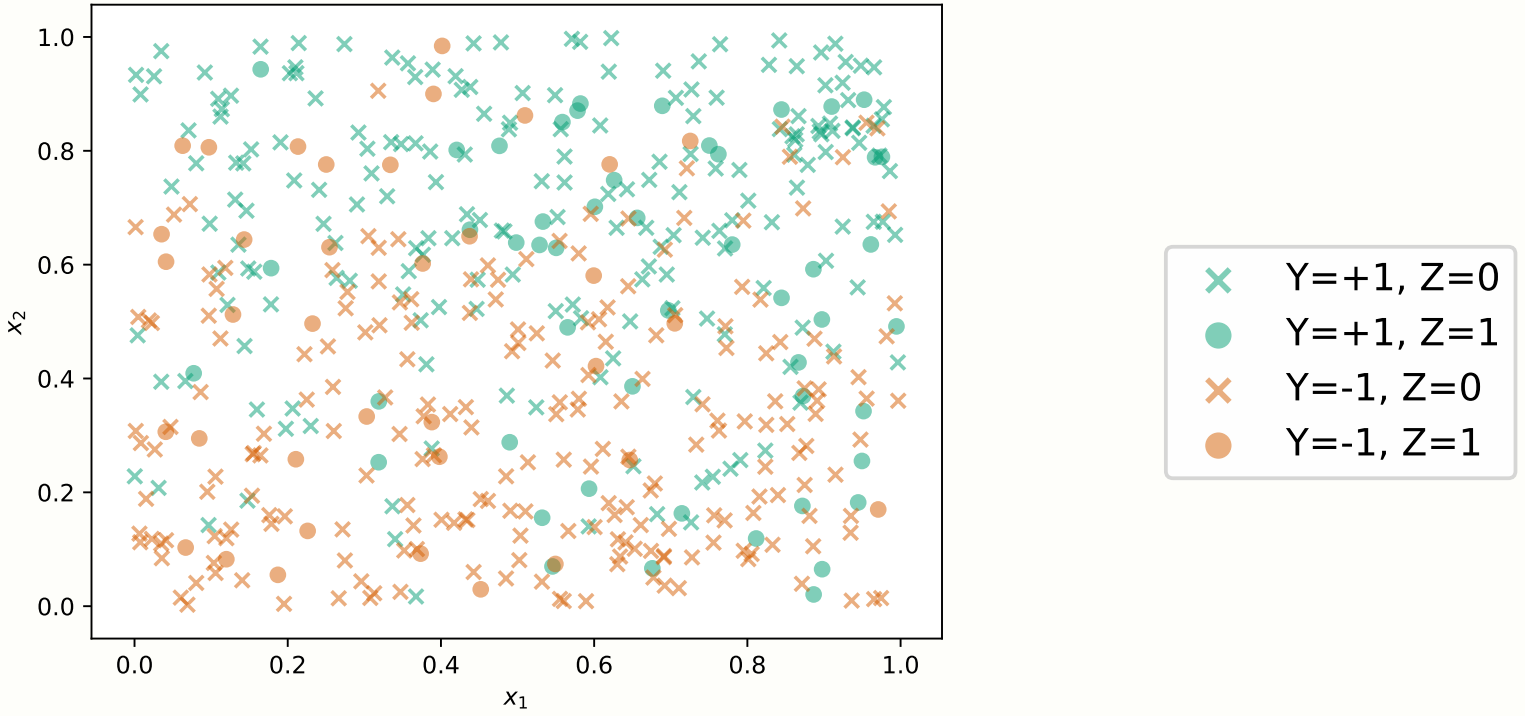
The parameter c changes every n_{adapt} iterations, on the basis of statistics computed on a validation dataset:
$\rightarrow$ if the term in the constraint is positive then $c = \min(c + \Delta c, 1)$,
$\rightarrow$ otherwise $c = \max(c - \Delta c, -1)$,

Since the AUC only evaluate the order of elements, we normalize s using moving means and averages as batch-normalization does.

Toy example: fairness with AUC constraints (1/2)

Set $\mathcal{X} = [0,1]^2$, with $X|Z = z$ uniform, and for $x = (x_1\ x_2)^\top \in \mathcal{X}$,
$\eta^{(0)}(x) = x_1$ and $\eta^{(1)}(x) = x_2$ and $\eta^{(z)}(x) = \mathbb{P}\{Y = +1 | Z = z, X = x\}$.

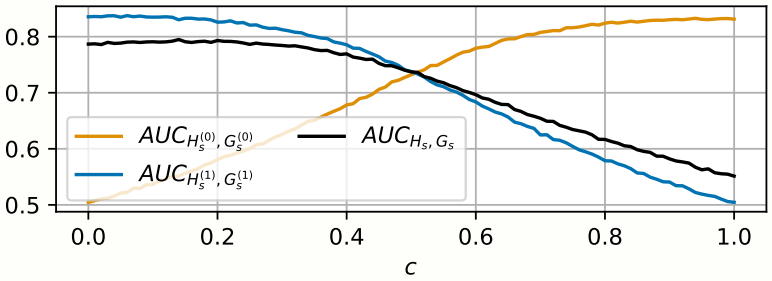
Fix $\mathbb{P}\{Z = 1\} = 17/20$, *i.e.* $Z = 1$ is the majority group.



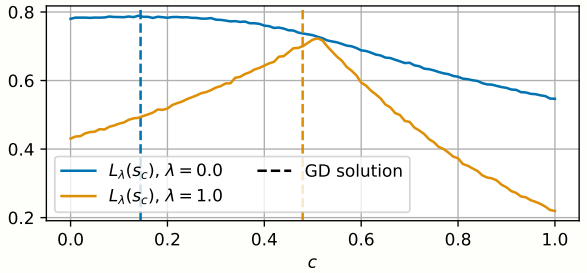
Example sample from P.

Toy example: fairness with AUC constraints (2/2)

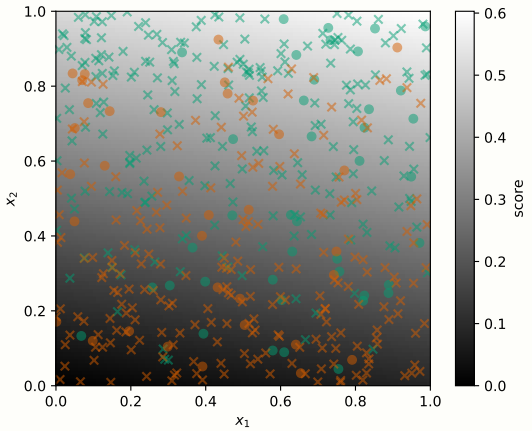
Consider a family of scorers $\{s_c\}_{c\in[0,1]}$, with $s_c(x) = cx_1 + (1-c)x_2$.



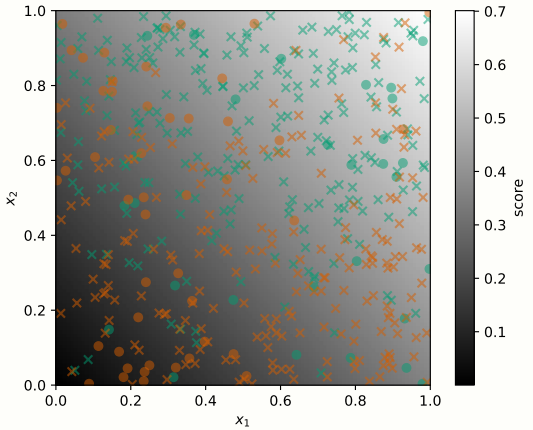
Values of the AUC's for any c.



Solutions s_c with AUC fairness.



Solution s_c with $\lambda = 0$.

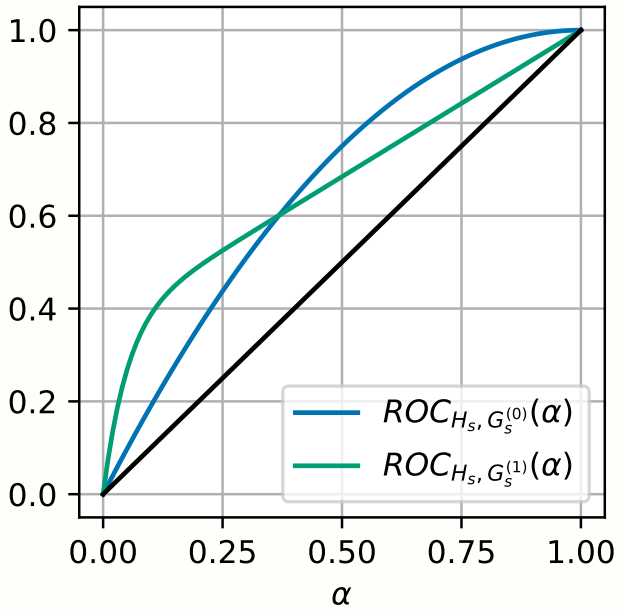


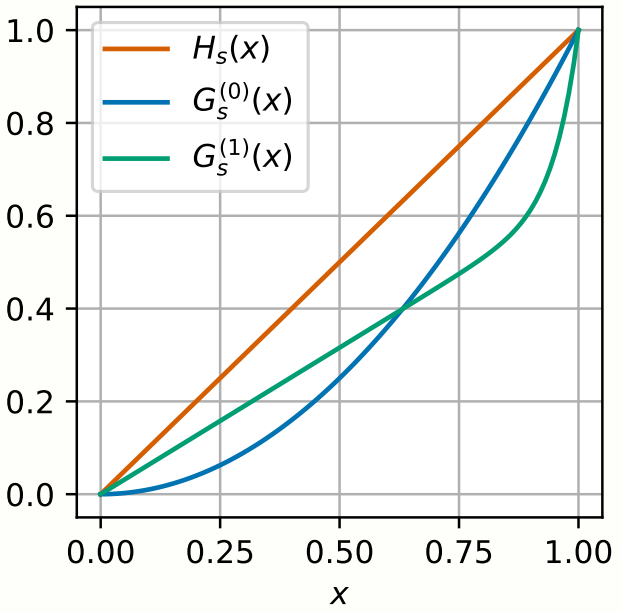
Solution s_c with $\lambda = 1$.

Outline

Limitations of AUC constraints

In the example below, with $s \in [0,1]$, an AUC constraint is verified.
However, $\sup_{t\in[0,1]} |G_s^{(0)}(t) - G_s^{(1)}(t)| \approx 0.10$.





Let h, g, h', g' cdfs on $\mathbb{R}$ s.t. $\mathrm{ROC}_{h,g}$ and $\mathrm{ROC}_{h',g'}$ are continuous.
If $\mathrm{AUC}_{h,g} = \mathrm{AUC}_{h',g'}$, $\exists \alpha \in (0,1)$ s.t. $g \circ h^{-1}(\alpha) = g' \circ h'^{-1}(\alpha)$.

Conclusion: An AUC constraint imposes a "pointwise constraint".

Learning with pointwise constraints

To measure the difference between cdfs for $Z = 0$ and $Z = 1$, let:

$$\Delta_{H,\alpha}(s) = \mathrm{ROC}_{H_s^{(0)},H_s^{(1)}}(\alpha) - \alpha \quad \text{and} \quad \Delta_{G,\alpha}(s) = \mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}(\alpha) - \alpha.$$

We introduce a sum of m_H pointwise constraints for $\Delta_{H,\cdot}$ and m_G for $\Delta_{G,\cdot}$ as a penalization, and maximize L_Λ in $\mathcal{S}$, where:

$$L_\Lambda(s) := \mathrm{AUC}_{H_s,G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} |\Delta_{H,\alpha_H^{(k)}}(s)| - \sum_{k=1}^{m_G} \lambda_G^{(k)} |\Delta_{G,\alpha_G^{(k)}}(s)|$$

which gives the score s_Λ^*.

The empirical counterpart of L_Λ is $\widehat{L}_\Lambda$, its maximizer is $\widehat{s}_\Lambda$.

Theorem 3

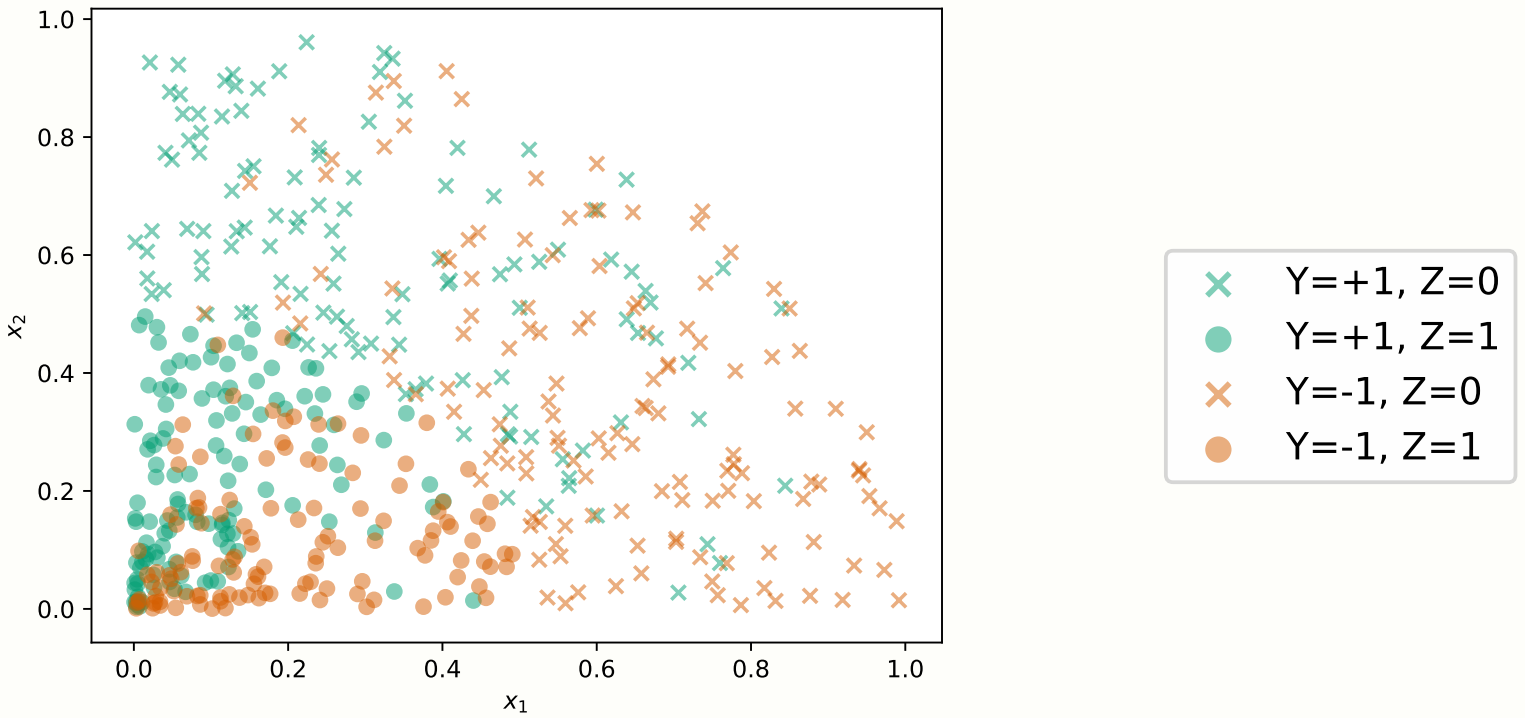
Assume that $\exists M, \kappa > 0$ *s.t.* $M \leq D_k'(s) \leq M \cdot \kappa$ *for all* $k \in [\![1, 4]\!], s \in \mathcal{S}$.
Under the assumptions of Theorem 2,

$$\epsilon^2 \cdot [L_\Lambda(s_\Lambda^*) - L_\Lambda(\widehat{s}_\Lambda)] \leq C_{\lambda,\epsilon,\kappa} \sqrt{\frac{V}{n}} + C'_{\Lambda,\epsilon,\kappa} \sqrt{\frac{\log(19/\delta)}{n-1}} + O(n^{-1}).$$

Toy example: fairness with ROC constraints (1/2)

Set $\mathcal{X} = [0,1]^2$ and $\eta^{(0)}(x) = \eta^{(1)}(x) = (2/\pi) \cdot \arctan(x_2/x_1)$, and:

$$\mu^{(0)}(x) = \frac{16}{\pi}\mathbb{I}\left\{x^2 + y^2 \leq \frac{1}{2}\right\} \quad \text{and} \quad \mu^{(1)}(x) = \frac{16}{3\pi}\mathbb{I}\left\{\frac{1}{2} \leq x^2 + y^2 \leq 1\right\}$$

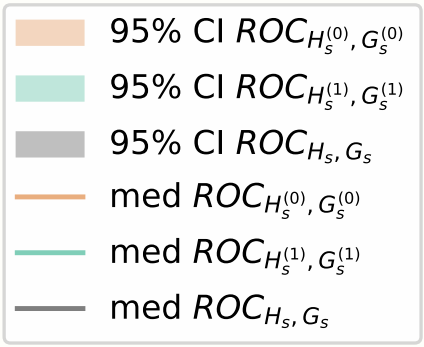


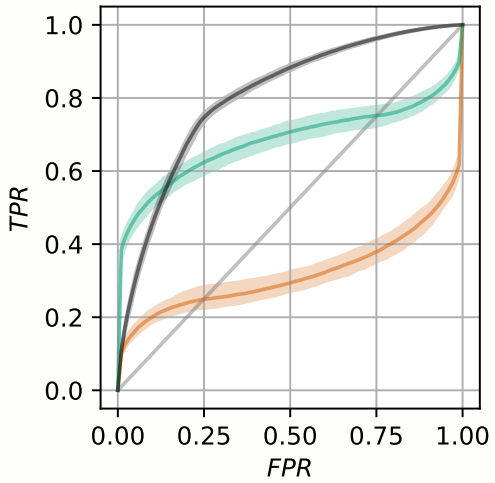
Example sample from P.

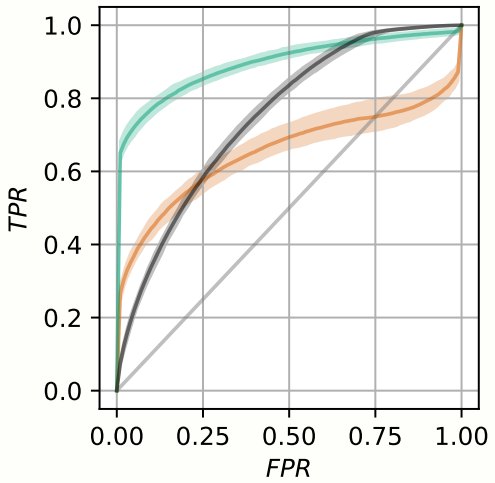
Toy example: fairness with ROC constraints (2/2)

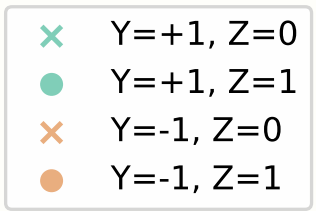
We chose for $\mathcal{S}$ a family of linear scores.

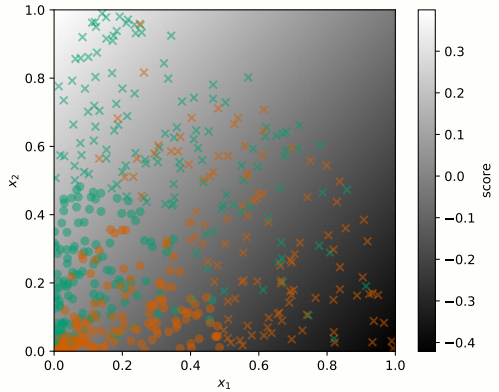
The constraint we impose is $\Delta_{H,3/4} = 0$.

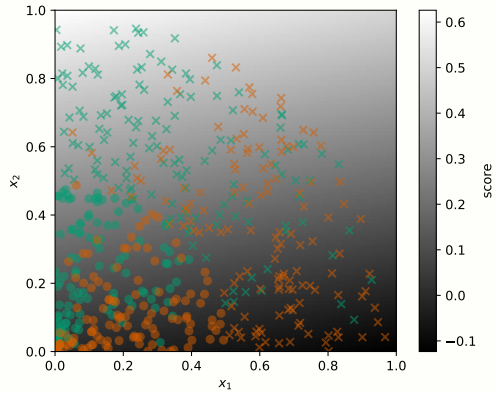












$\lambda = 0$ $\lambda = 1$

Outline

Datasets and parameters

We use three datasets from the fairness literature:

- *Adult Income Dataset*, featured *e.g.* in [Donini et al., 2018],
  Prediction: salary $\geq \$50K$ / sensitive group: gender.
- *Compas Dataset*, featured *e.g.* in [Donini et al., 2018],
  Prediction: recidivist or not / sensitive group: ethnicity.
- *Bank Marketing*, featured in [Zafar et al., 2019b],
  Prediction: client subscription / sensitive group: gender.

AUC-based constraints:

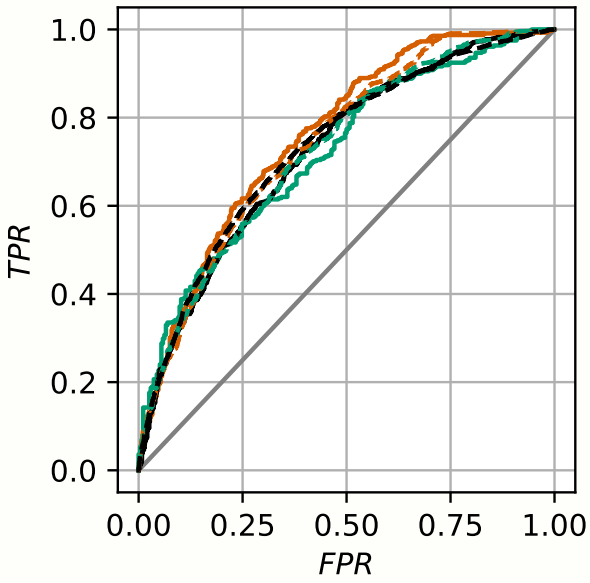
Different constraints are used, depending on the dataset.

ROC-based constraints:

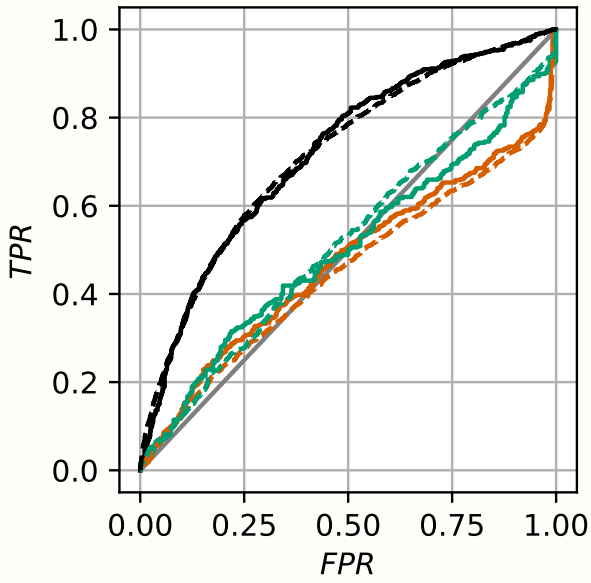
To align the dist. of low FPR's between $Z = 0$ and $Z = 1$,
we penalize high $|\Delta_{H,1/8}(s)|$ and $|\Delta_{H,1/4}(s)|$.

Results - Compas

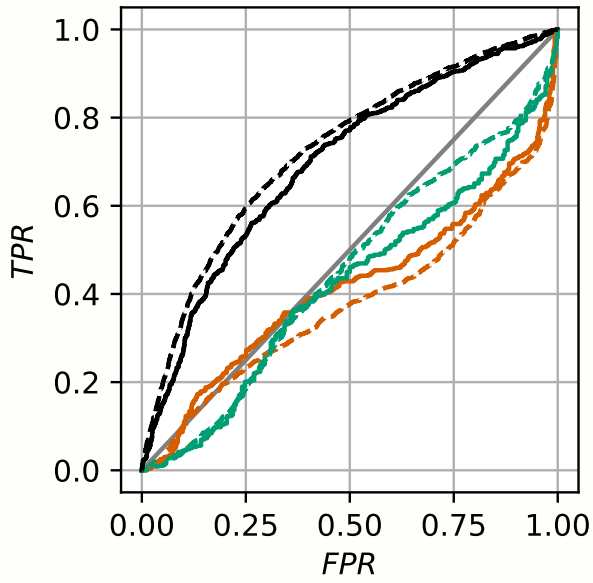
The AUC fairness sufficed to impose equality of low FPR's.



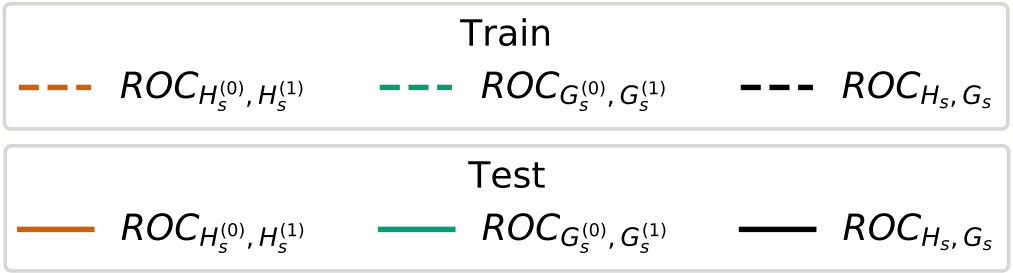
No constraint



AUC fairness

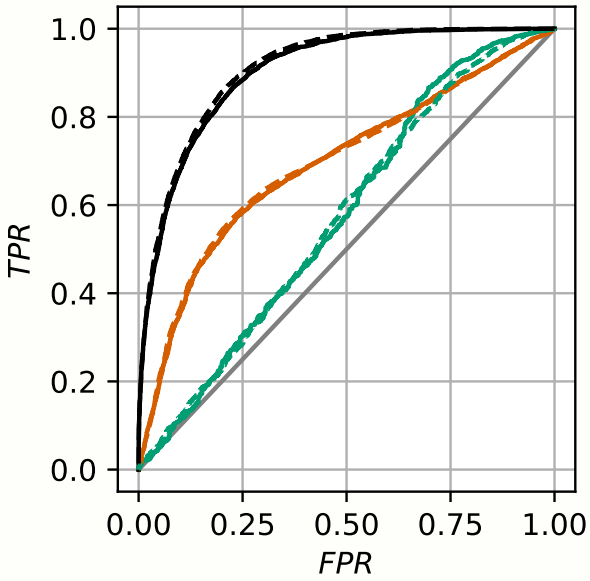


ROC fairness

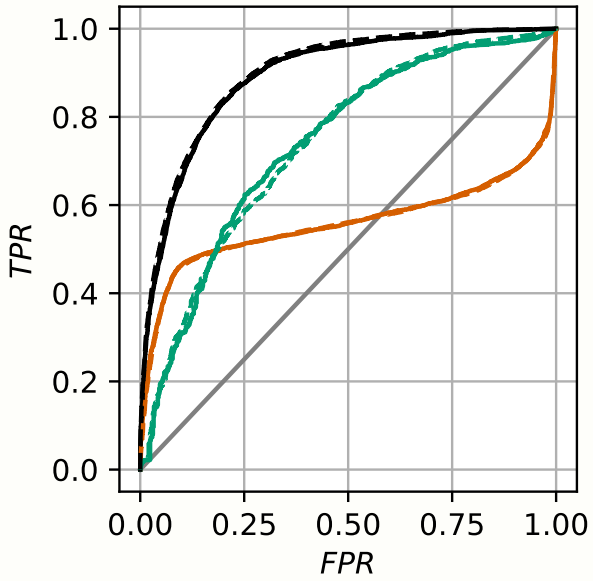


Results - Adult

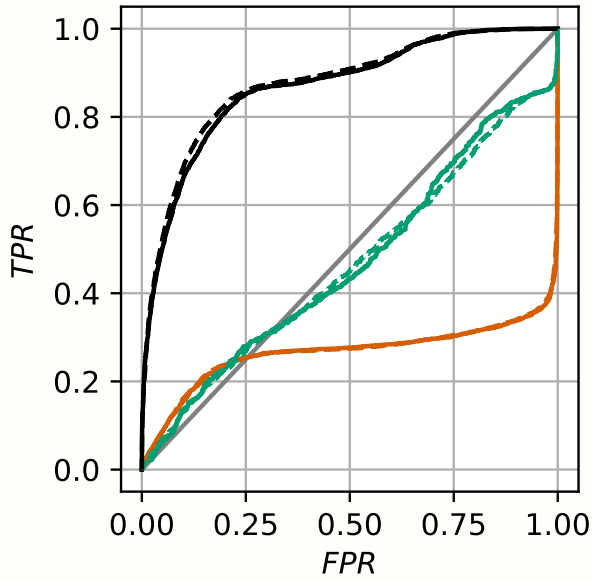
AUC fairness changed the distribution of the negative scores,
but was far from equality of low FPR's, while ROC fairness was not.



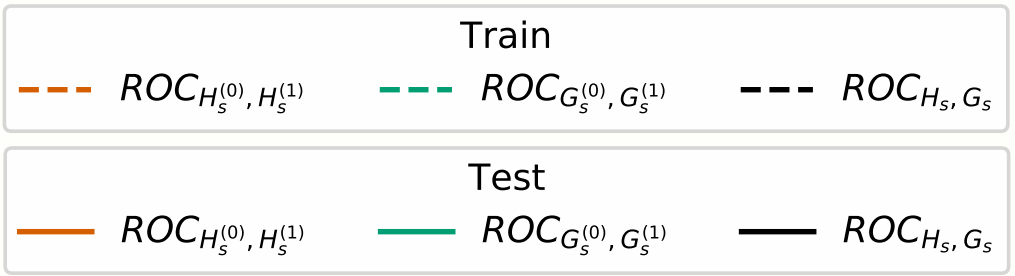
No constraint



AUC fairness

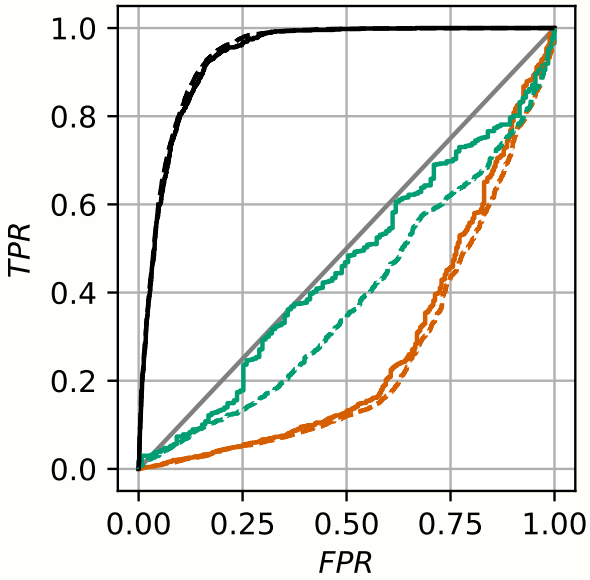


ROC fairness

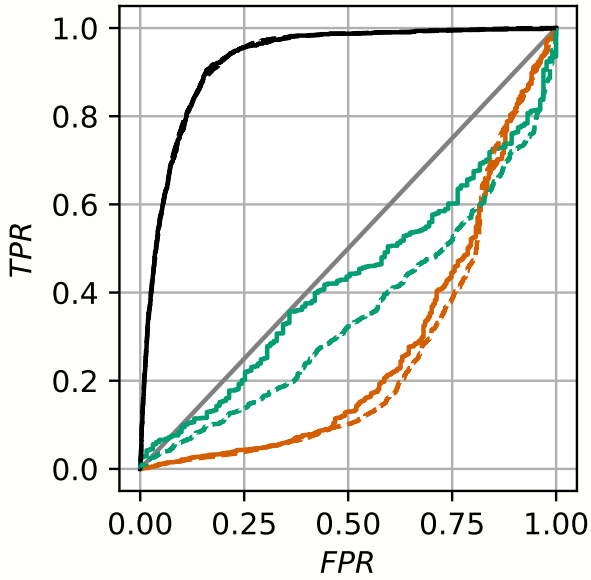


Results - Bank

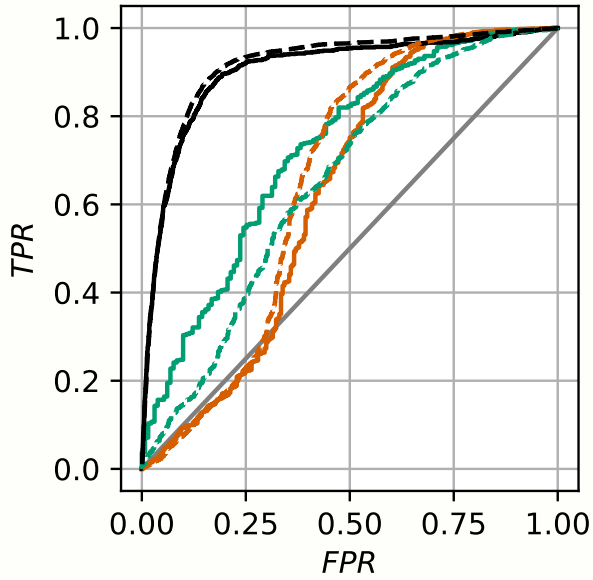
AUC fairness induced slight changes of $\mathrm{ROC}_{F_s^{(0)}, F_s^{(1)}}$ with $F \in \{H, G\}$, while ROC fairness was again very effective.



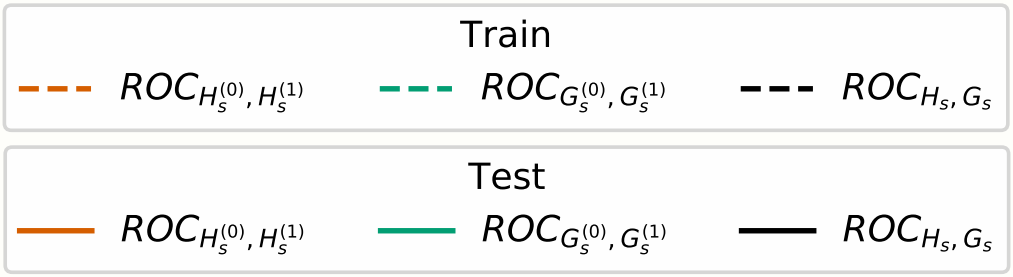
No constraint



AUC fairness



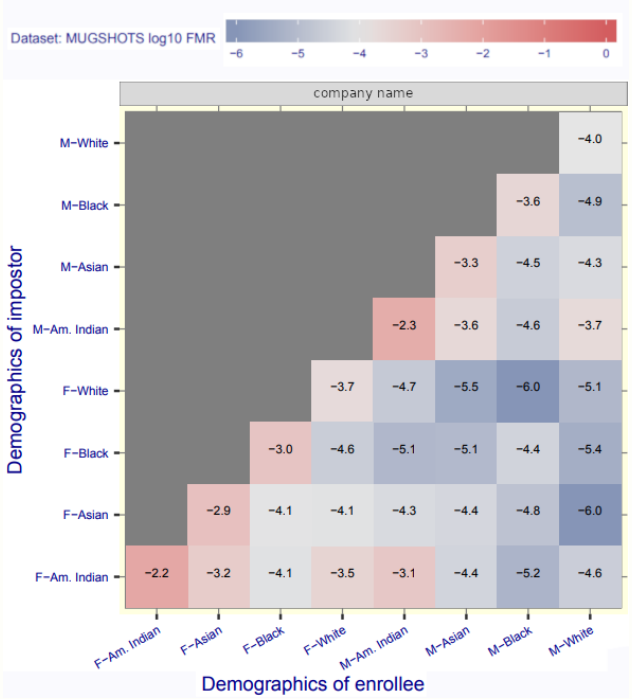
ROC fairness



Outline

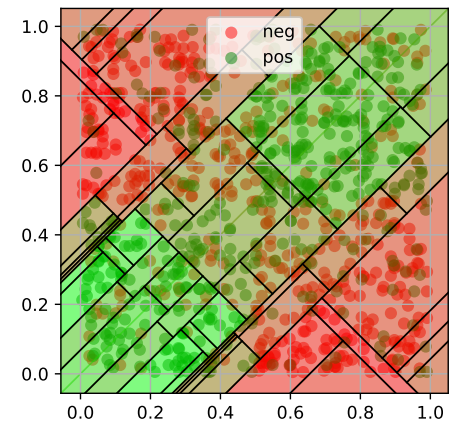
Conclusion

Extension:
This work extends to **similarity ranking**, *i.e.* ranking pairs of items by similarity, see [Vogel et al., 2018].
The NIST currently investigates fairness of facial recognition algorithms in terms of ethnicity and gender.
The r.h.s. figure is from their report.



FRVT: FPR by ethnicity at fixed *t*.

Future work:
Fair constraints for ROC optimization, based on recursive partitioning, see [Clémençon et al., 2010].



Thank you !

References I


Barocas, S., Hardt, M., and Narayanan, A. (2019).
Fairness and Machine Learning.
fairmlbook.org.
`http://www.fairmlbook.org`.


Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. (2019).
Fairness in recommendation ranking through pairwise comparisons.
In *KDD*.


Biega, A. J., Gummadi, K. P., and Weikum, G. (2018).
Equity of attention: Amortizing individual fairness in rankings.
In *SIGIR*.


Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019).
Nuanced metrics for measuring unintended bias with real data for text classification.
arXiv:1903.04561.


Clémençon, S., Depecker, M., and Vayatis, N. (2010).
Adaptive partitioning schemes for bipartite ranking.
Machine Learning.


Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018).
Empirical risk minimization under fairness constraints.
In *NeurIPS*.

References II


Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. (2018).
Learning adversarially fair and transferable representations.
ICML, abs/1802.06309.


Singh, A. and Joachims, T. (2018).
Fairness of exposure in rankings.
In *KDD*.


Singh, A. and Joachims, T. (2019).
Policy learning for fairness in ranking.
In *NeurIPS*.


Vogel, R., Bellet, A., and Clémençon, S. (2018).
A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization.
In *ICML*. PMLR.


Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019a).
Fairness constraints: A flexible approach for fair classification.
Journal of Machine Learning Research, 20(75):1–42.


Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019b).
Fairness constraints: A flexible approach for fair classification.
Journal of Machine Learning Research, 20(75):1–42.

References III

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017).
FA*IR: A Fair Top-k Ranking Algorithm.
In *CIKM*.