Learning Fair Scoring Functions

Work published at AISTATS 2021

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

[1] University of Edinburgh, [2] Télécom Paris, [3] Inria

18/11/2021

Outline

Generalities on Fair ML

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.

05/2016: The COMPAS system predicts recidivism likelihood for US courts.

Algorithms are designed for the interest of some party,
fairness in ML suggests confronting those to the law.

"Predictive models are really just opinions embedded in math." *C. O'Neil.*

Fairness Definitions in Binary Classification

A lot of recent works considered fairness in binary classification, with two sensitive groups.
[Donini et al., 2018, Menon and Williamson, 2018, Zafar et al., 2019]

They add a **sensitive variable** $Z \in \{0, 1\}$ to the usual binary classification model (X, Y), and learn $g : \mathcal{X} \to \{-1, +1\}$ from:

$$\mathcal{D}_n = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\} \subset \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}.$$

Many definitions of fairness exist, and apply to specific use-cases.
· Treatment: $g(X, Z) = g(X)$ a.s.
· Impact: $\mathbb{P}\{g(X) = +1 \mid Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Z = 1\}$
· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$
· FPR: $\mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 1\}$

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:

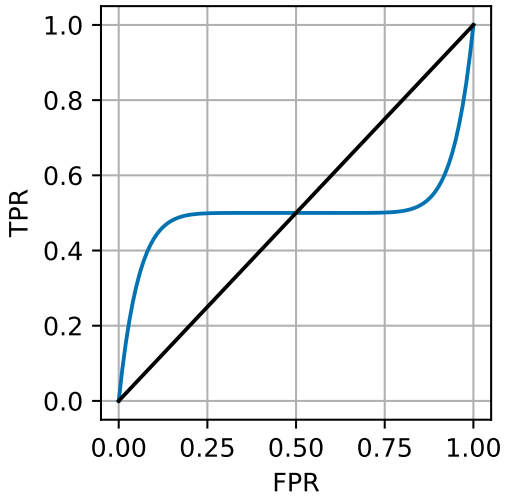
$$\mathrm{ROC}_{H_s,G_s}(\alpha) = \overline{G}_s \circ \overline{H}_s^{-1}(\alpha) \quad \text{and} \quad \mathrm{AUC}_{H_s,G_s} = \int_0^1 \mathrm{ROC}_{H_s,G_s}(\alpha)\, d\alpha.$$

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

Fairness Definitions in Bipartite Ranking

Fair ranking is a recent topic, mostly tackled by the IR community.
Authors:

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

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

· use fairness def° based on AUC's [Borkan et al., 2019, Beutel et al., 2019].

For any $z \in \{0, 1\}$, let: $H_s^{(z)}(t) := \mathbb{P}\{s(X) \leq t \mid Y = -1, Z = z\}$,
$G_s^{(z)}(t) := \mathbb{P}\{s(X) \leq t \mid Y = +1, Z = z\}$.

BNSP AUC ([Borkan et al., 2019]): $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$.

Many similar def° of fairness based on the AUC were proposed.

Contribution:
We proposed: 1) an unified framework for learning with AUC constraints, as well as 2) generalization garantees with an AUC fairness penalization.

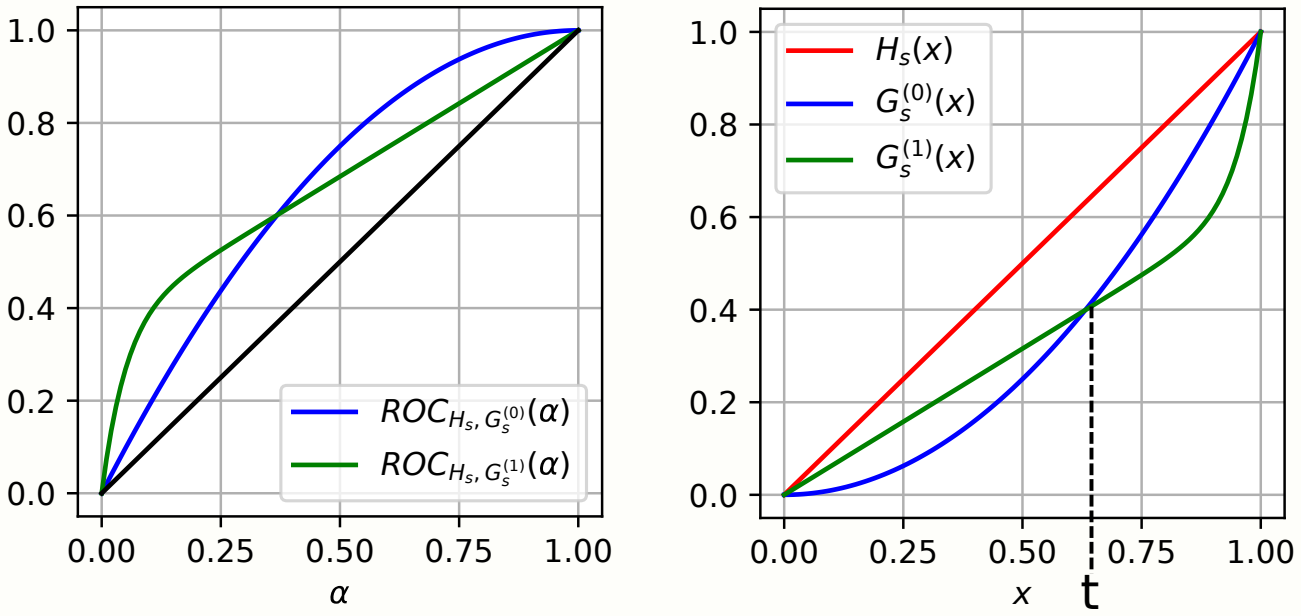
1) detailed in appendix, 2) consequence of U-statistics theory.

Outline

Limitations of AUC Constraints

Below, with $s \in [0, 1]$, we have $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$.

However, $\sup_{t \in [0,1]} |G_s^{(0)}(t) - G_s^{(1)}(t)| \approx 0.10$.



There exists an **unknown threshold** t, for which the induced classifier $g_{s,t}(x) = \mathrm{sign}(s(x) - t)$ is fair in FNR.

We propose fairness contraints specified on points of ROC's.

Learning with Pointwise ROC Constraints

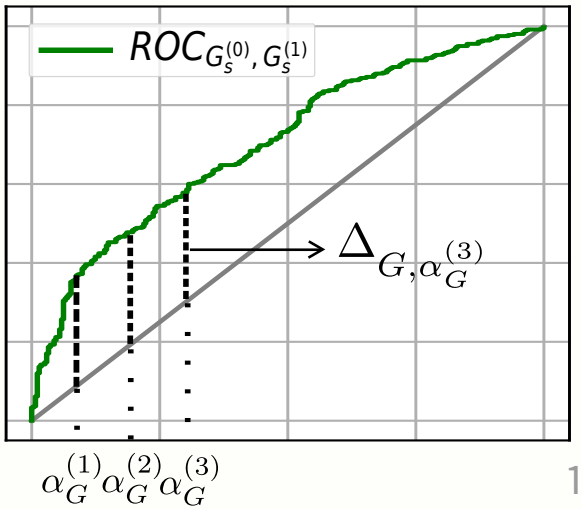
To measure the difference between cdfs for $Z = 0$ and $Z = 1$, let:

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

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

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

We prove finite-sample generalization bounds in $O(n^{-1/2})$ for maximizing L_Λ.
See appendix.



Outline

Experimental Settings

Here, we report on two datasets from the fairness literature:

· *Compas Dataset*, featured *e.g.* in [Donini et al., 2018],
Prediction: recidivist or not / sensitive group: ethnicity.

· *Adult Income Dataset*, featured *e.g.* in [Donini et al., 2018],
Prediction: salary $\geq \$50K$ / sensitive group: gender.

AUC-based constraints:

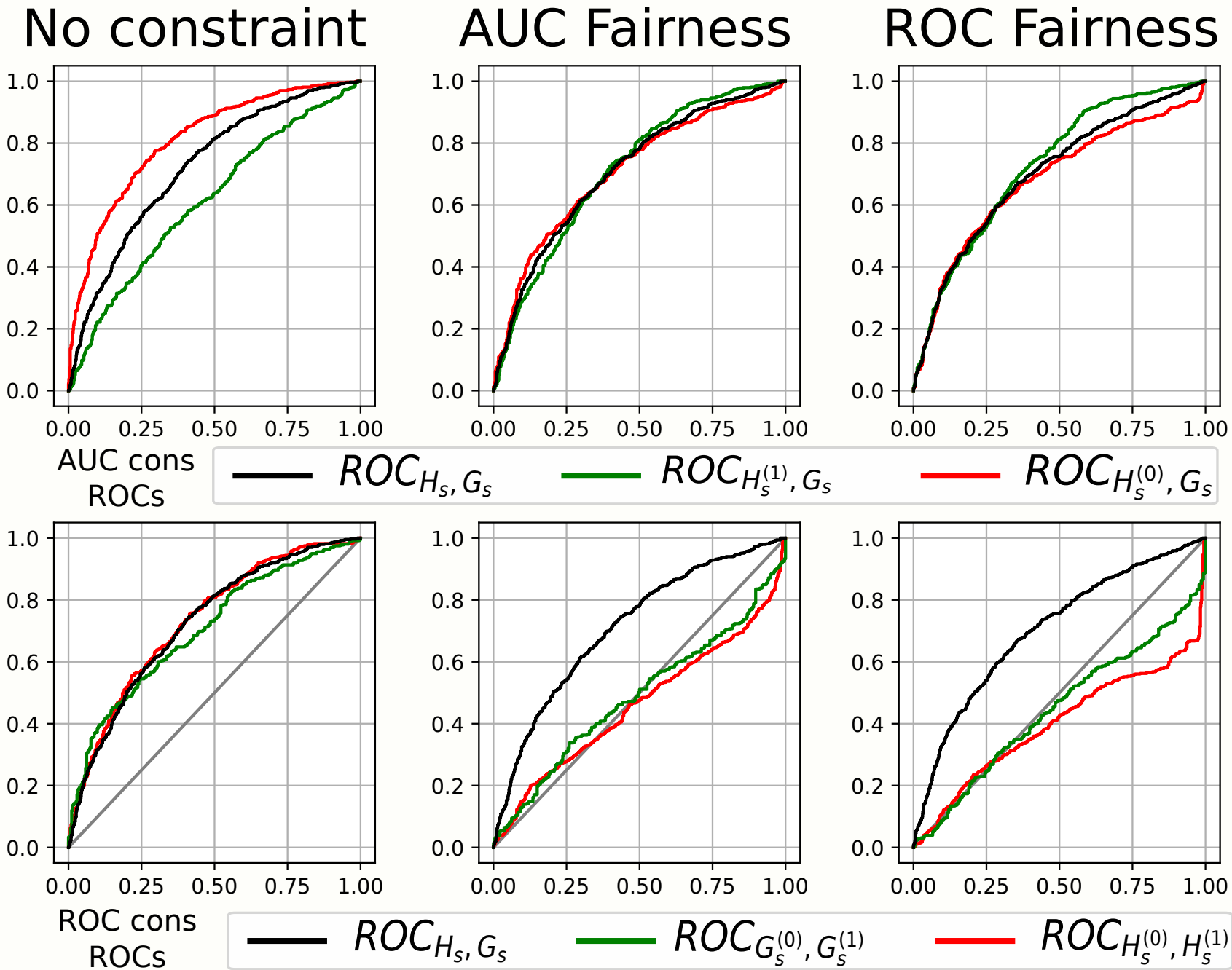
Different constraints are used, depending on the dataset.

ROC-based constraints:

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

Experimental Results - Compas

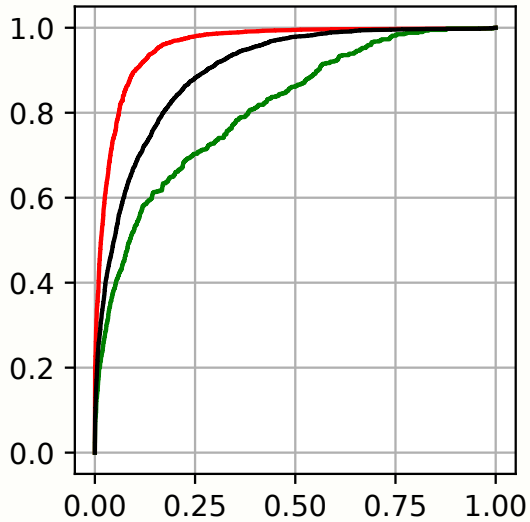
0: caucasian
1: ethnic minority



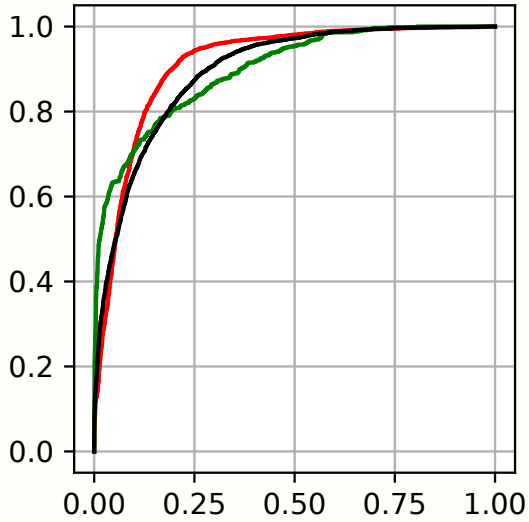
Experimental Results - Adult

0: woman
1: man

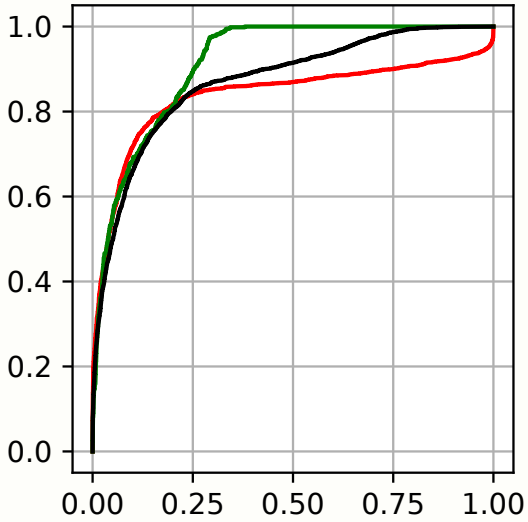
No constraint



AUC Fairness

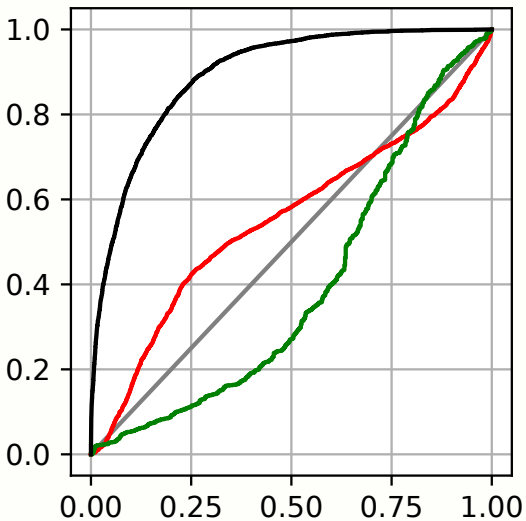


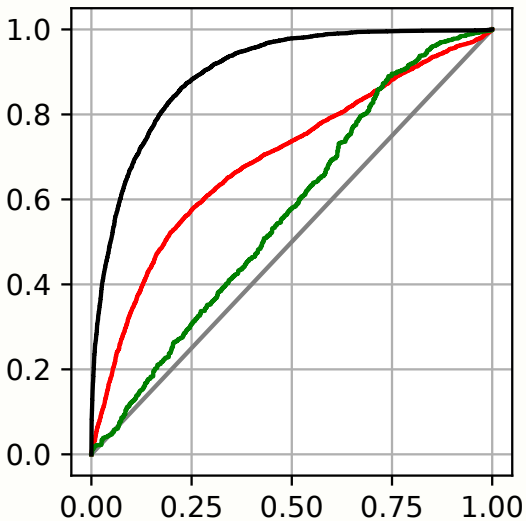
ROC Fairness

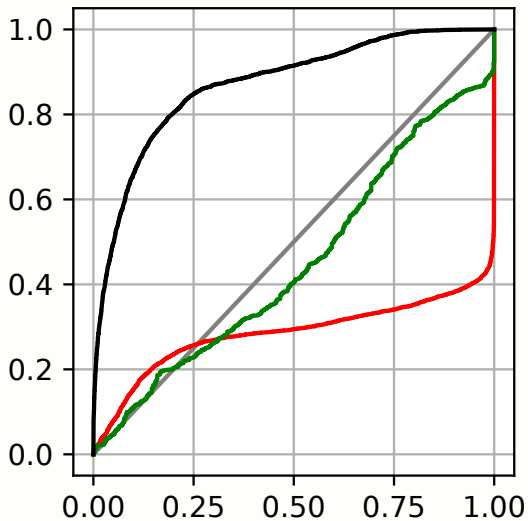


AUC cons ROCs

ROC_{H_s, G_s} $ROC_{H_s^{(1)}, G_s^{(0)}}$ $ROC_{H_s^{(0)}, G_s^{(1)}}$







ROC cons ROCs

ROC_{H_s, G_s} $ROC_{G_s^{(0)}, G_s^{(1)}}$ $ROC_{H_s^{(0)}, H_s^{(1)}}$

Outline

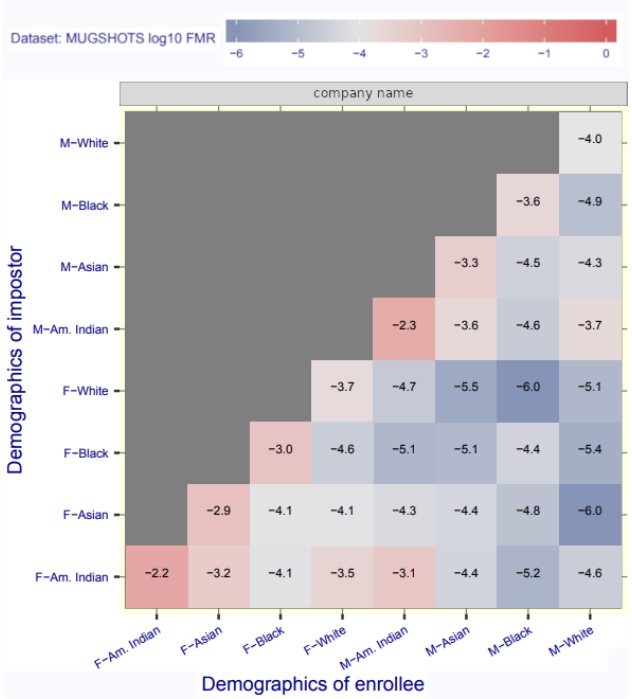
Conclusion

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.

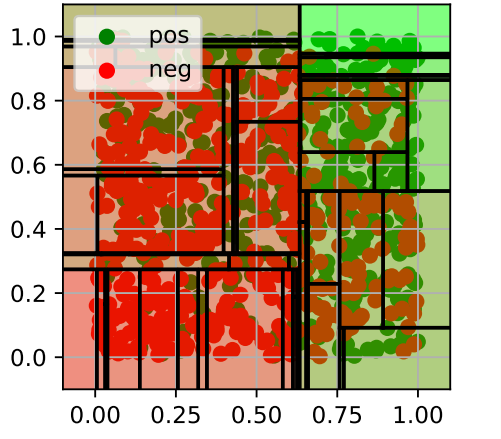
The r.h.s. figure is from their report.



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


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*.


Hsieh, F. and Turnbull, B. W. (1996).
Nonparametric and semiparametric estimation of the receiver operating characteristic curve.
The Annals of Statistics, 24(1):25–40.

References II


Menon, A. K. and Williamson, R. C. (2018).
The cost of fairness in binary classification.
In *Conference on Fairness, Accountability and Transparency, FAT 2018*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR.


Shorack, G. and Wellner, J. a. (1989).
Empirical Processes with applications to Statistics.
SIAM.


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. (2019).
Fairness constraints: A flexible approach for fair classification.
Journal of Machine Learning Research, 20(75):1–42.


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*.

General AUC constraint details

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:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)}, \tag{2}$$

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

Theorem 1

The following propositions are equivalent:

1. *Eq.* (2) *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.* (2) *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.

Sketch of proof of ROC fairness guarantees

This result is based on a control of the terms, for both $F \in \{F, G\}$:

$$\sup_{s,\alpha \in \mathcal{S} \times [0,1]} |\widehat{\Delta}_{F,\alpha}(s) - \Delta_{F,\alpha}(s)|,$$

which is the uniform deviation of an empirical ROC.

From [Hsieh and Turnbull, 1996], we have:

$$\widehat{\Delta}_{F,\alpha}(s) - \Delta_{F,\alpha}(s) = \left[F_s^{(1)} \circ F_s^{(0)-1} - F_s^{(1)} \circ \widehat{F}_s^{(0)-1}\right](1-\alpha) \quad (3)$$

$$+ \left[F_s^{(1)} \circ \widehat{F}_s^{(0)-1} - \widehat{F}_s^{(1)} \circ \widehat{F}_s^{(0)-1}\right](1-\alpha). \quad (4)$$

We can bound Eq. (4) by: $\sup_{(s,t) \in \mathcal{S} \times \mathbb{R}} |\widehat{F}_s^{(1)}(t) - F_s^{(1)}(t)|$.

Since the derivative of $F_s^{(1)}$ is bounded, using the mean value theorem, we can bound Eq. (3) by *almost* a quantile process.

The equality of the uniform deviation of a standard and a quantile process [Shorack and Wellner, 1989] implies the result.