Wasserstein-based fairness interpretability framework for machine learning models

Alexey Miroshnikov
Konstandinos Kotsiopoulos
Ryan Franks
Arjun Ravi Kannan

Emerging Capabilities & Data Science Research Group, Discover Financial Services

Boston University, CISE seminar, October 2021
Boston, Massachusetts

Overview

- Introduction
- Fairness/bias for classifiers
- Fairness/bias for regressors
- Model bias metrics
- ML interpretability
- Fairness interpretability

Introduction

- Predictive ML models, and strategies that rely on such models, are subject to laws and regulations that ensure fairness (e.g. ECOA, EEOA).
- Examples of protected attributes: race, gender, age, ethnicity, national origin, marital status, etc.
- Tradeoff between accuracy and bias

Main steps in ML fairness

1. Fairness assessment (or bias measurement)
2. Bias mitigation

Fairness for classifier

Notation

Data (X, G, Y)

- $X \in \mathbb{R}^n$, predictors
- $G \in \{0,1\}$ (e.g. male/female)
- $Y \in \{0,1\}$, response variable

Models

- $f(X) = \widehat{\mathbb{P}}(Y = 1|X)$, trained classification score
- $Y_t = 1_{\{f(X)>t\}}$, a classifier for a given threshold $t \in \mathbb{R}$
- $\hat{Y}$, a classifier

Labels

- Non-protected class: $G = 0$
- Favorable outcome: $Y = 0$

Fairness for classifier

- ML bias can be viewed as an ability to differentiate between subpopulations at the level of data or outcomes (*Dwork et al 2012*)

Statistical parity *(Feldman et al, 2015)*

$$\mathbb{P}(\hat{Y}=0|G=0)=\mathbb{P}(\hat{Y}=0|G=1)$$

Equalized odds *(Hardt et al, 2015)*

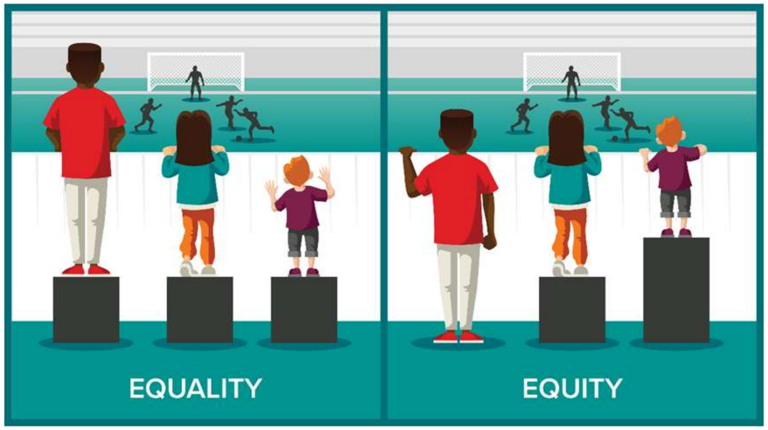
$$\mathbb{P}(\hat{Y}=0|Y=y,G=0)=\mathbb{P}(\hat{Y}=0|Y=y,G=1), y\in\{0,1\}$$

Equal opportunity *(Hardt et al, 2015)*

$$\mathbb{P}(\hat{Y}=0|Y=0,G=0)=\mathbb{P}(\hat{Y}=0|Y=0,G=1)$$

Geometric parity for $\hat{Y}_{t_*}$ *(Miroshnikov et al, 2021a)*

$$F_0^{[-1]}(p_*)=F_1^{[-1]}(p_*),\ \ p_*=F_0(t_*)=\mathbb{P}(f(X)\leq t_*|Y=0)$$



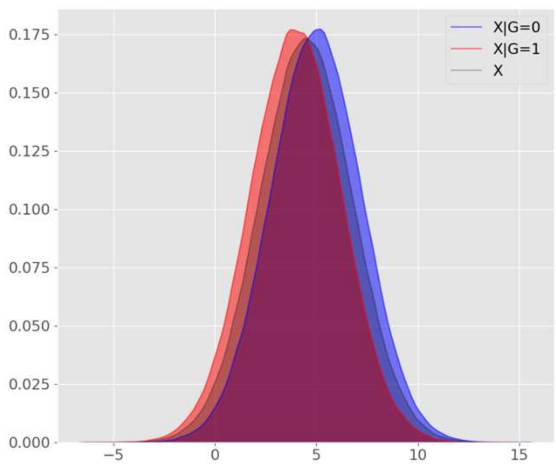
Fairness in classifiers

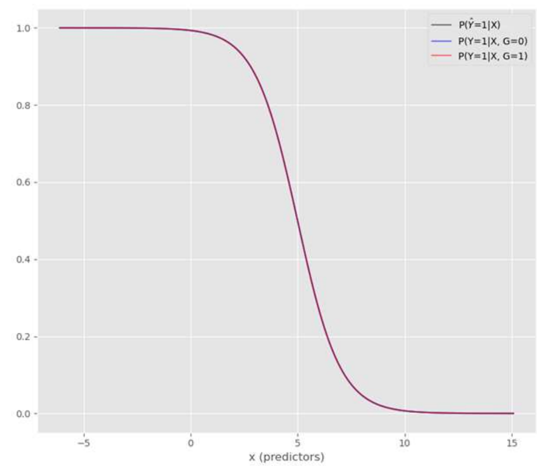
Statistical parity classifier bias

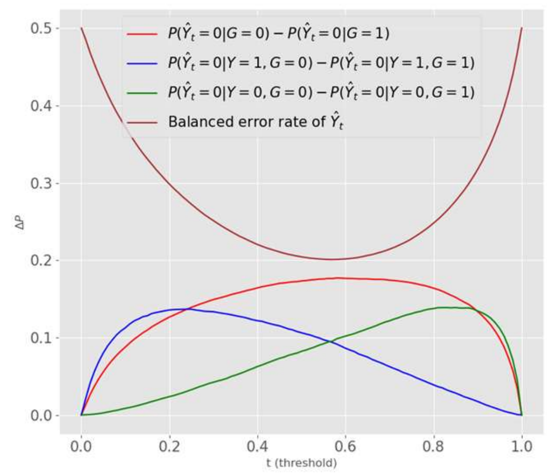
$$bias(Y_t|X,G) = |\mathbb{P}(Y_t = 0|G = 0) - \mathbb{P}(Y_t = 0|G = 1)|$$

Example (proxy predictor)

- $X \sim N\left(5 - G, \sqrt{5}\right)$, $\mathbb{P}(G = 0) = \mathbb{P}(G = 1) = 0.5$
- $Y \sim Bernoulli(f(X)), f(x) = logistic(5 - x)$







Fairness in classifiers

Some approaches for bias mitigation of classifiers:

- Maximization with fairness constraints

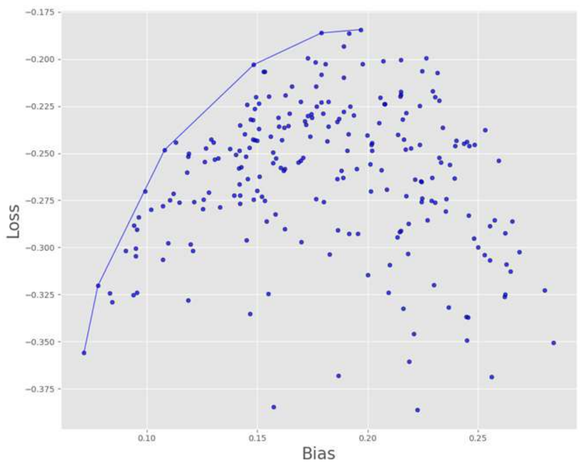
$$Y^*(X,G) \; or \; Y^*(X) = \max_{fairness(Y^*|G)} \mathcal{L}\left(Y^*, X^{(train)}\right),$$ or mini-max approach

Dwork et al (2012), Woodworth et al (2017), Zhang et al (2018), and many others.

- Post-corrective methods (Hardt et al, 2015)
  - Design randomized (equalized odds) optimal classifier $\tilde{Y}(X,G;f) \in \mathcal{P}(\{0,1\})$ given the trained score f.

- Fair dataset construction. Feldman et al, 2015

- Pareto efficient frontier. Schmidt and Stephens (2019), Perrone et al (2020).



Motivation

- Explicit use of the protected attribute G is not allowed by ECOA neither in training nor prediction
- Typical bias measurements test fairness of a classifier Y_t, not the regressor score $f(X)$
- Mitigation procedures often focus on the construction of a fair classifier $Y^*(X,G)$, not a fair model $f^*(X,G)$
- Fair ML hyperparameter search might be computationally expensive due to retraining
- Determining the main drivers (predictors) for the model bias

Acceptable form of bias mitigation

1. Given the (regressor) model f assess the bias across subpopulation distribution of $f(X)|G=k,\ k \in \{0,1\}$
2. Determine the main drivers for the bias $X_{i_1}, X_2, \dots X_{i_n} = X_I$
3. Construct a post-processed model $\tilde{f}(X; f, X_I)$ that does not rely on G

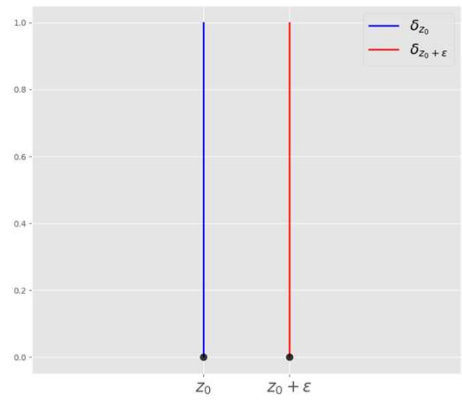
Model bias metrics for regressors

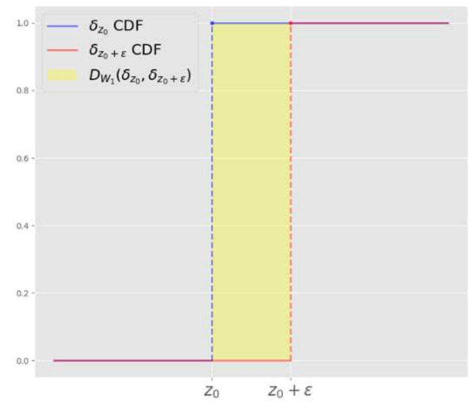
- At an algorithmic level, the bias can be viewed as an ability to differentiate between two subpopulations at the level of data or outcomes.

- Bias metrics requirements:

1. Must keep track of the geometry of the model distribution $P_{f(X)}$ (values control)
2. Must be consistent with a wide class of classifier fairness criteria
3. Must keep track of the sign of the bias across subpopulations
4. Must be meaningful (interpretable)

- An ability to differentiate vs independence:





Model bias metrics

Potential candidates

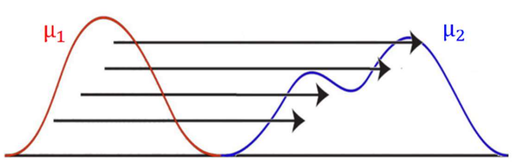
μ_1, μ_2 probability measures on a metric space $\mathcal{Z}$ equipped with a metric $d(z_1, z_2)$.

- Randomized binary classifier (RBC) based bias [Dwork et al (2012)]

$M_z : \mathcal{Z} \to \mathcal{P}(\{0,1\})$, randomized classifier.

$$Bias_{d,D_{TV}}(\mu_1, \mu_2) = \sup_{M \in Lip_1(\mathcal{Z},d,D_{TV})} \{ \ \mathbb{E}_{z\sim\mu_1}[M_z(0)] - \mathbb{E}_{x\sim\mu_2}[M_z(0)] \ \}$$

- Wasserstein metric W_q (optimal transport cost of μ_1 to μ_2 and vice verse)



$$W_q(\mu_1, \mu_2; d)^q = \inf_{\pi \in \mathcal{P}(\mathcal{Z}^2)} \{ \ \mathbb{E}_{(z_1,z_2)\sim\pi} [d(z_1, z_2)]^q, \ \text{(transport plan) } \pi \text{ with marginals } \mu_1, \mu_2 \ \}$$

- In our application μ_1, μ_2 are $P_{f(X)|G=k}$, $k = 0,1$.
- What about statistical distance such KS or mutual information between $f(X)$ and G?

Model bias metrics

Facts

- (Dwork et al 2012): if μ_1, μ_2 have discrete supports and $d \leq 1$

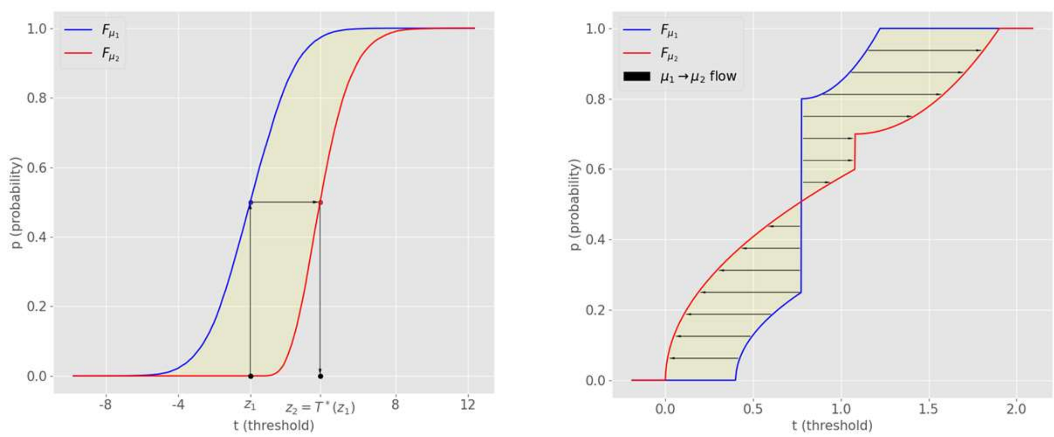
$$Bias_{d,D_{TV}}(\mu_1, \mu_2) = W_1(\mu_1, \mu_2; d)$$

- (Miroshnikov et al 2021a): for any μ_1, μ_2 with support in $B_L(z_*)$ and $d(\mathrm{z}_1, \mathrm{z}_2) = \|z_1 - z_2\|$

$$Bias_{d,D_{TV}}(\mu_1, \mu_2) = \frac{1}{L} W_1(\mu_1 \circ T^{-1}, \mu_2 \circ T^{-1}; d),\ T,\ \text{affine transformation}$$

- μ_1, μ_2 on $\mathcal{B}(\mathbb{R})$, with $d(z_1, z_2) = |z_1 - z_2|$, there exists order preserving optimal transport plan π^*

$$\mathrm{W}_1(\mu_1, \mu_2) = \int |x_1 - x_2|\, d\pi^* = \int \left| F_{\mu_1}^{[-1]}(p) - F_{\mu_2}^{[-1]}(p) \right| dp = [\text{Shorack}, 1956] = \int \left| F_{\mu_1}(t) - F_{\mu_2}(t) \right| dt$$



Model bias metrics

Facts (Miroshnikov et al, 2021a)

- W_q scales under linear transformations of μ_k $(d = \|\cdot\|)$, but $Bias_{D,TV} \in [0,1]$ saturates.

- Given predictors X ,model f, and $G \in \{0,1\}$

  (model bias) $Bias_{W_1}(f|X,G) = W_1(f(X)|G=0, f(X)|G=1)$

- Connection with statistical parity:

$$Bias_{W_1}(f|X,G) = \int bias(Y_t|X,G)dt$$

- Connection with generic parity: $\mathcal{A} = \{A_1, \dots, A_M\}$, $\mathbb{P}(Y_t = 1|G=0, A_m) = \mathbb{P}(Y_t = 1|G=1, A_m)$, $A_m \in \mathcal{A}$

$$Bias_{W_1,\mathcal{A}}(f|X,G) = \sum w_m W_1(f(X)|\{G=0, A_m\}, f(X)|\{G=1, A_m\}) = \int bias_{\mathcal{A}}(Y_t|X,G)dt$$

Model bias metrics

Assumption

Model $f(X) \in \mathbb{R}$ has a favorable direction (for a risk score the direction is ←)

Definition

Positive/negative model bias $Bias^{\pm}_{W_1}(f|X,G)$ is the transport effort (under π^*) of $\mathrm{P}_{f(X)|G=0}$ in favorable/non-favorable directions

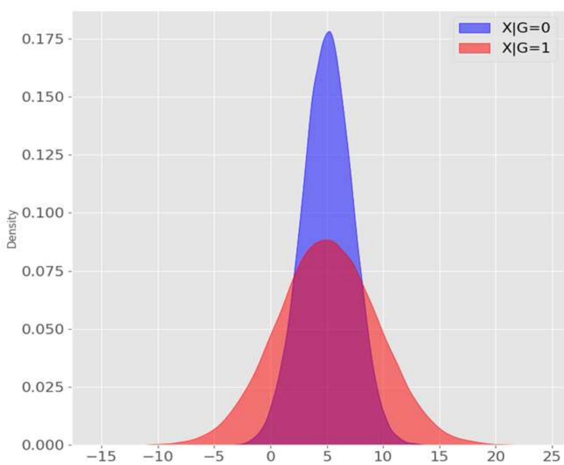
Example

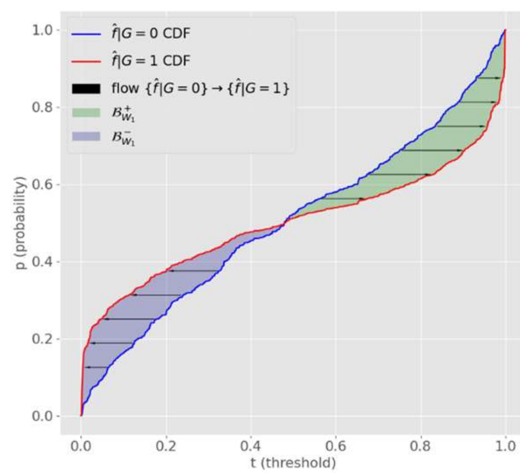
$X \sim \mathcal{N}(\mu, (1+G)\sqrt{\mu})$

$Y \sim Bernoulli\big(f(X)\big)$

$f(X) = \sigma(\mu - X)$

$\zeta_f = -1$





Fairness interpretability objectives

Objective

- Determine the main drivers for the model biases $Bias^{\pm}_{W_1}(f|X,G)$

Main idea

- Combine ML interpretability methods and transport approach

ML Interpretability

Having a complex model structure comes at the expense of interpretability.

Interpretability approaches

- Self-explainable models
- Post-hoc explanations

Post-hoc explainers (examples)

- $E_i^{ME}(X;f) = \mathbb{E}\left[f\left(x_i, X_{-\{i\}}\right)\right]|_{x_i=X_i}$, marginal expectation (ME), [PDP, Freidman, 2001]
- $E_i^{CE}(X;f) = \mathbb{E}[f(X)|X_i]$, conditional expectation (CE)

ML Interpretability

Post-hoc explainers (game theory)

- Players: $N = \{1,2,\ldots,n\}$ (features become player)
- Game: set function $v(S)$, $S \subset N$, $v(N)$ = total payoff
- Game value: $h[N,v] = (h_1[v], h_2[v], \ldots h_n[v]) \in \mathbb{R}^n$

Shapley value (Shapley, 1953)

$$\varphi_i[v] = \sum_{S \subset N} \frac{(s-1)!(n-s)!}{n!} \big(v(S) - v(S \backslash \{i\})\big), \; i \in N$$

φ is efficient: $\sum_i \varphi_i[v] = v(N)$, linear, symmetric.

Probabilistic games

- $v^{CE}(S; X, f) = \mathbb{E}[f(X_S, X_{-S}) | X_S]$, conditional game explores model predictions
- $v^{ME}(S; X, f) = \mathbb{E}[f(x_S, X_{-S})]|_{x_S = X_S}$, marginal game explores the model

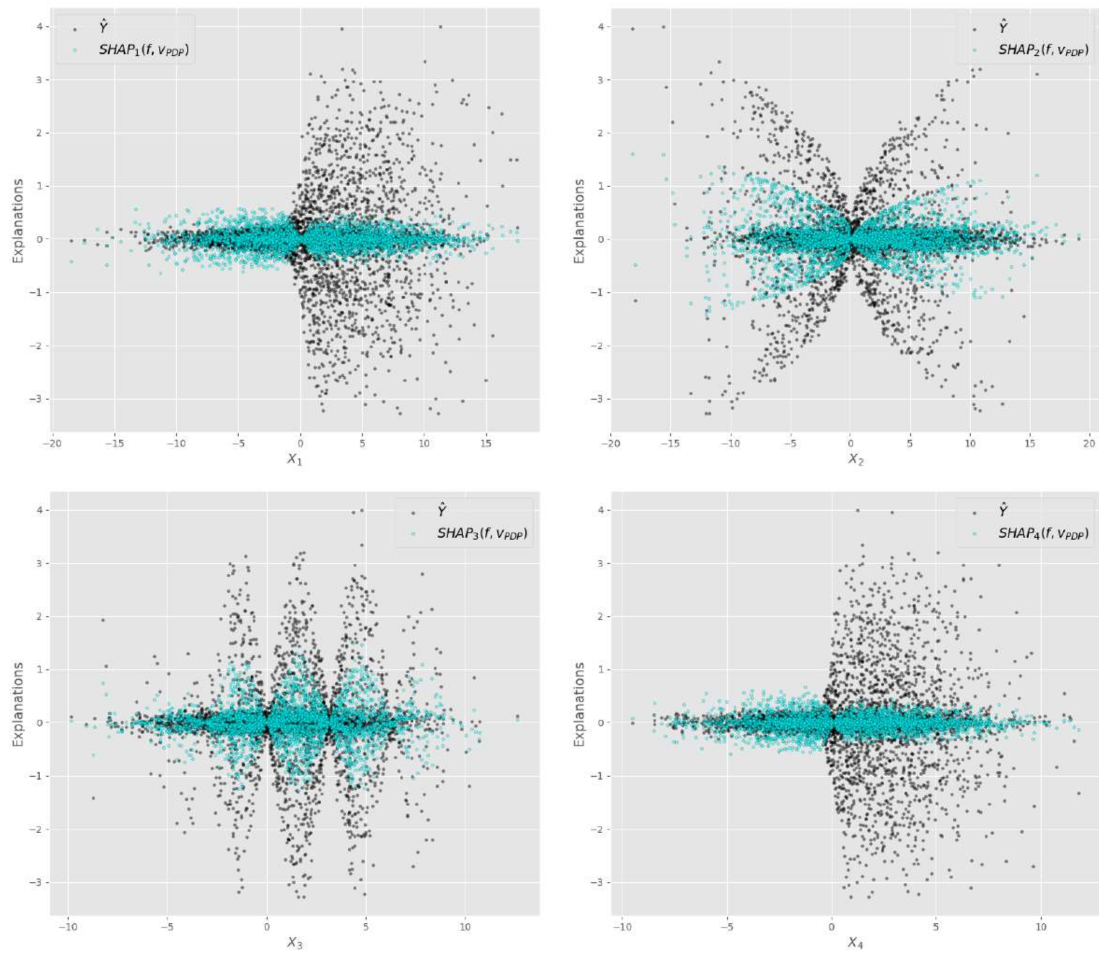
ML Interpretability

Fun Example (Marginal Shapley $h[v] = \varphi[v]$)

$$Y = \prod_{i=1}^{4} f_i(X_i) + \epsilon = f(X) + \epsilon$$

$$f_1(X_1) = logistic(2X_1), \quad f_2(X) = \text{sgn}(X_2)\sqrt{|X_2|},$$
$$f_3(X_3) = \sin(X_3), \quad f_4(X_4) = logistic(5X_4).$$

$$(X_1, X_2) \sim \mathcal{N}((1,1), \Sigma_1), \quad \Sigma_1 = \begin{bmatrix} 26 & -10 \\ -10 & 26 \end{bmatrix}$$
$$(X_3, X_4) \sim \mathcal{N}((1,1), \Sigma_2), \quad \Sigma_2 = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$



Fairness Interpretability

Definition (basic bias explanations)

- Given an explainer $E_i(X; f)$ of predictor X_i, the bias explanation is defined via the transport cost

$$\beta_i(f|X, G) = W_1(E_i(X)|G = 0, E_i(X)|G = 1)$$

- Positive and negative bias explanations $\beta^{\pm}$ are defined as transport effort in favorable and non-favorable directions.

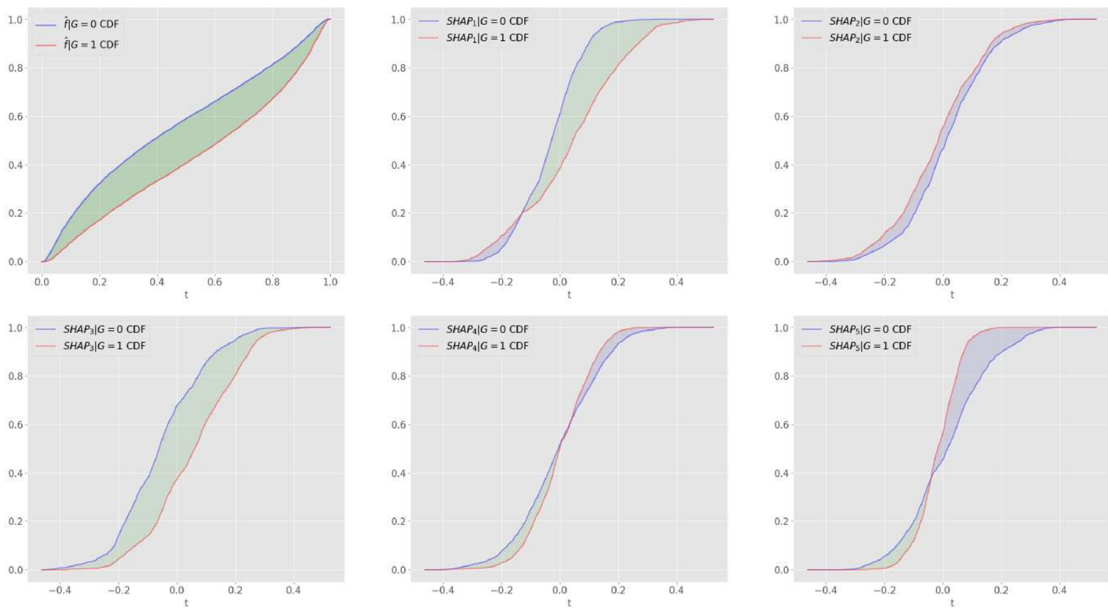
Notes

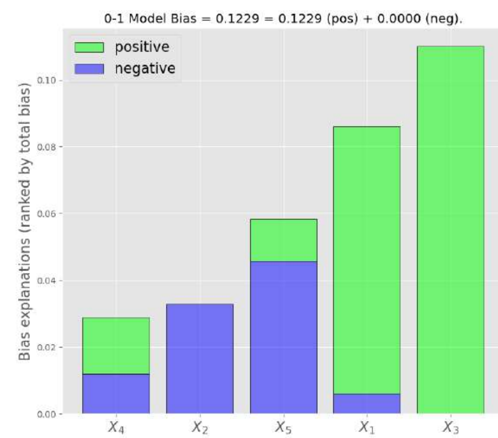
- Type of ML explainers matters (marginal vs conditional)
- Some ML explainers isolate the effect of each predictor and some not (local vs global)

Fairness Interpretability

Example: bias explanations based on marginal Shapley values

$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$
$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$
$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$
$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$
$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$
$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$
$Y \sim Bernoulli\big(f(X)\big), f(X) = \sigma(\sum X_i - 24.5)$

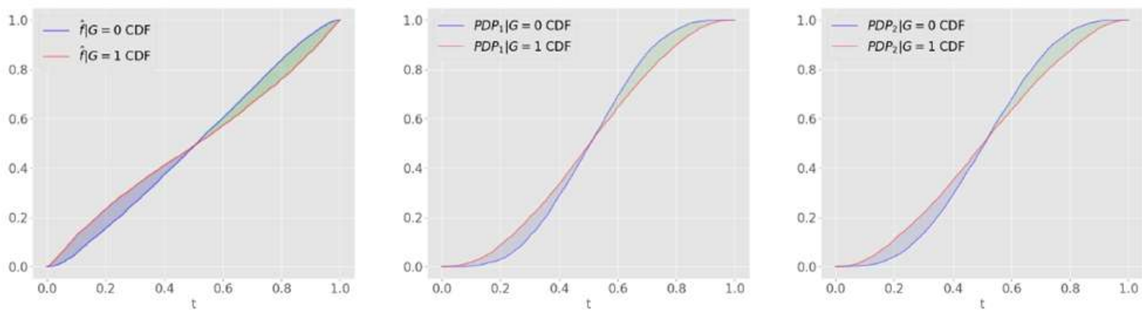




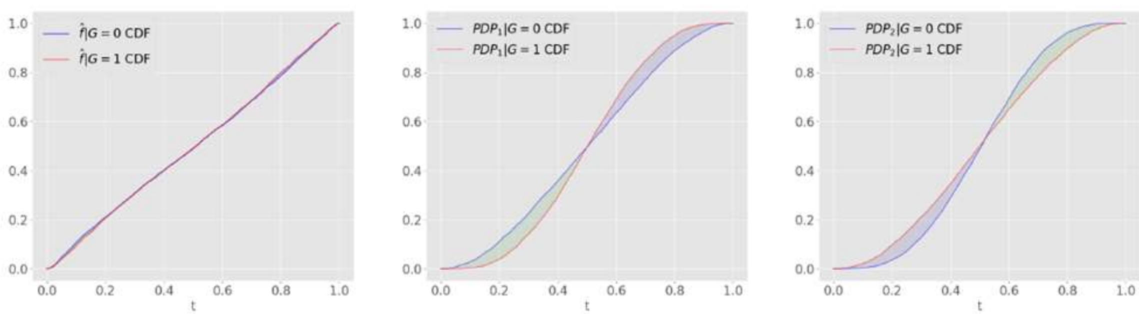
Fairness Interpretability

Example (offsetting)

$X_1 \sim \mathcal{N}(\mu, 1+G), X_2 \sim \mathcal{N}(\mu, 1+G)$
$Y \sim Bernoulli(f(X)), f(X) = \sigma(2\mu - X_1 - X_2)$



$X_1 \sim \mathcal{N}(\mu, 2-G), X_2 \sim \mathcal{N}(\mu, 1+G)$
$Y \sim Bernoulli(f(X)), f(X) = \sigma(2\mu - X_1 - X_2)$



Notes
- Bias explanations are the same
- Bias predictor interactions

Fairness Interpretability

- Basic bias explanations are not additive
- Cannot handle bias interactions when mixed bias predictors are present or predictors interact
- No tracking of how mass is transported

Fairness Interpretability

- Basic bias explanations are not additive
- Cannot handle bias interactions when mixed bias predictors are present or predictors interact
- No tracking of how mass is transported

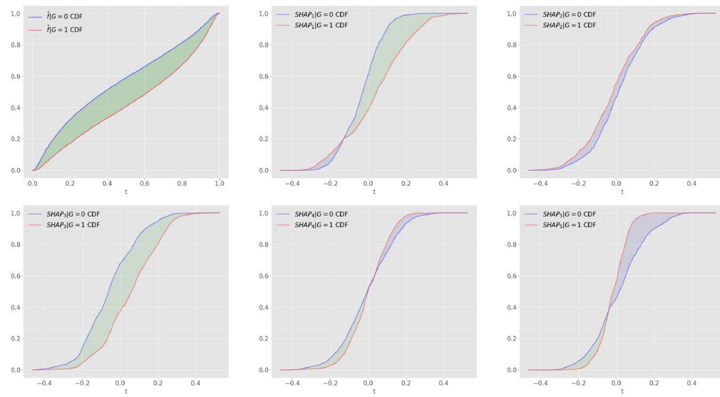
Game theoretical approach

- Consider an ML explainer $E_S(X;f)$ of predictor X_S, $S \subset \{1,2,\ldots n\}$
- Predictors $\{X_i\}_{i\in N}$ are players that push/pull explainer subpopulation distributions apart when joining a coalition $S \subset N$
- A game $v^{bias}(S) = W_1(E_S(X)|G=0, E_S(X)|G=1)$
- Shapley bias explanations $\varphi^{bias}(f|X,G) = \varphi[v^{bias}]$

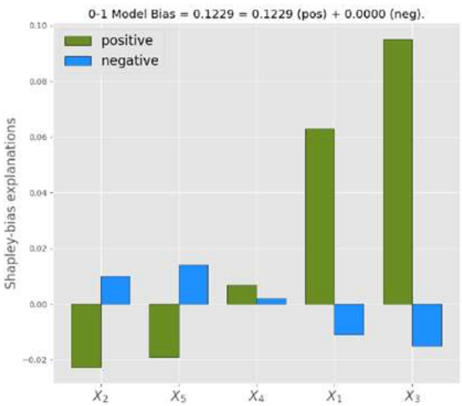
Fairness Interpretability

Example (marginal Shapley-bias explanations)

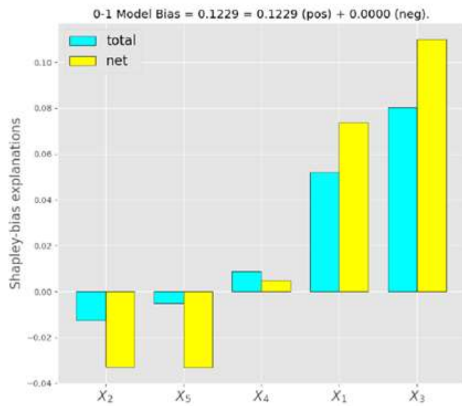
$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$
$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$
$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$
$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$
$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$
$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$
$Y \sim Bernoulli\big(f(X)\big), f(X) = \sigma(\sum X_i - 24.5)$



$\varphi[v^{bias\pm}(\cdot, \varphi[v^{ME}])]$



$\varphi[v^{bias}(\cdot, \varphi[v^{ME}])]$



On stability of bias explanations

- Conditional bias explanations are consistent with the data; computational complexity might be infeasible under dependencies in X.
- Marginal bias explanations are consistent with the structure of the model $f(x)$, complexity $O(2^n)$

Lemma (stability [Miroshnikov et al 2021a])

The conditional and marginal Shapley-bias explanations have the following properties:

i. $|\,\varphi_i^{bias\pm}(f|G, \varphi_S[v^{CE}]) - \varphi_i^{bias\pm}(f|g, \varphi_S[v^{CE}])\,| \leq C\|f-g\|_{L^2(P_X)}$

ii. $|\varphi_i^{bias\pm}(f|G, \varphi_S[v^{ME}]) - \varphi_i^{bias\pm}(f|g, \varphi_S[v^{ME}])| \leq C\|f-g\|_{L^2(\tilde{P}_X)}, \tilde{P}_X = \frac{1}{2^n}\sum_{S\subset N} P_{X_S} \otimes P_{X_{-S}}$

On stability of bias explanations

- Conditional bias explanations are consistent with the data; computational complexity might be infeasible under dependencies in X
- Marginal bias explanations are consistent with the structure of the model $f(x)$, complexity $O(2^n)$

Lemma (stability [Miroshnikov et al 2021a])

The conditional and marginal Shapley-bias explanations have the following properties:

i. $|\varphi_i^{bias\pm}(f|G,\varphi_S[v^{CE}]) - \varphi_i^{bias\pm}(f|g,\varphi_S[v^{CE}])| \leq C\|f-g\|_{L^2(P_X)}$

ii. $|\varphi_i^{bias\pm}(f|G,\varphi_S[v^{ME}]) - \varphi_i^{bias\pm}(f|g,\varphi_S[v^{ME}])| \leq C\|f-g\|_{L^2(\tilde{P}_X)}, \tilde{P}_X = \frac{1}{2^n}\sum_{S\subset N} P_{X_S} \otimes P_{X_{-S}}$

Notes (Miroshnikov et al, 2021b, arXiv:2102.10878) :

- For marginal Shapley-bias explanations continuity in $L^2(P_X)$ in general breaks down under dependencies in X
- Marginal and conditional points of view can be unified via grouping and stability in $L^2(P_X)$ is guaranteed
- Complexity can be reduced via quotient games and recursive approach

Acknowledgements

- Steve Dickerson (SVP, Chief Data Science Officer, Decision Management, Discover)
- Raghu Kulkarni (VP, Data Science, Discover)
- Melanie Wiwczaroski (Sr. Director, Enterprise Fair Banking, Discover)
- Patrick Haggerty (Director & Senior Counsel, Discover)
- Kate Prochaska (Sr. Counsel & Director, Regulatory Policy, Discover)
- Markos Katsoulakis (Full Professor, University of Massachusetts Amherst)
- Robin Young (Full Professor, University of Massachusetts Amherst)
- Matthias Steinrücken (Assistant Professor, University of Chicago)