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

North Carolina State University, Mathematics Department, Numerical Analysis Seminar, March 2022

Overview

- Introduction
- Classifier fairness
- Regressor fairness
- ML interpretability
- Fairness interpretability
- Bias mitigation with regulatory constraints

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.

Setup

Data (X, G, Y)

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

Models

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

Labels

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

Classifier fairness

- 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)*

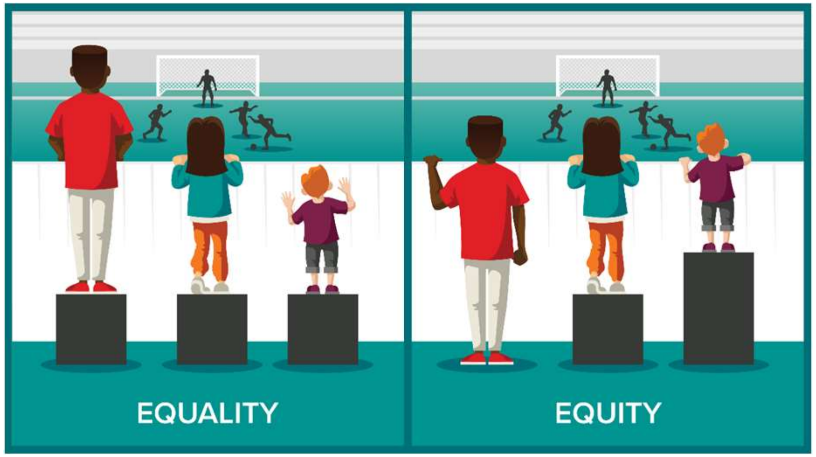
$$\mathbb{P}(\hat{Y} = 0 | G = 0) = \mathbb{P}(\hat{Y} = 0 | G = 1)$$

  - 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)*

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



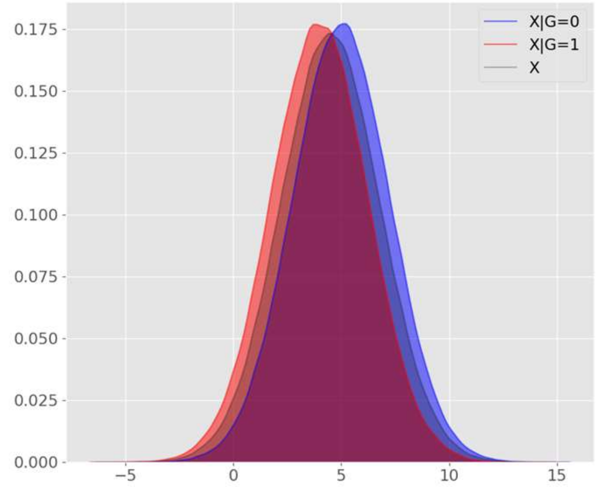
Classifiers fairness

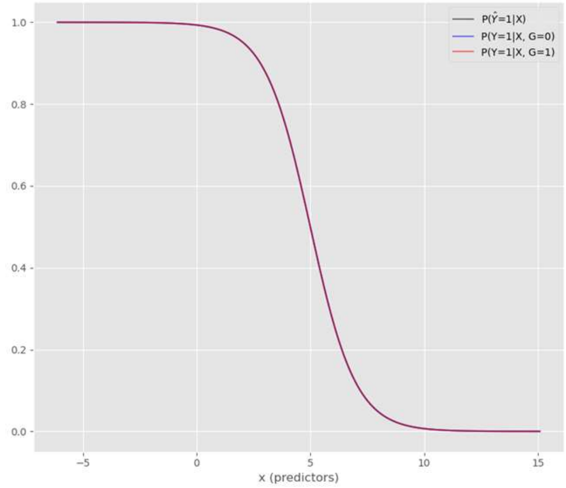
Statistical parity classifier bias

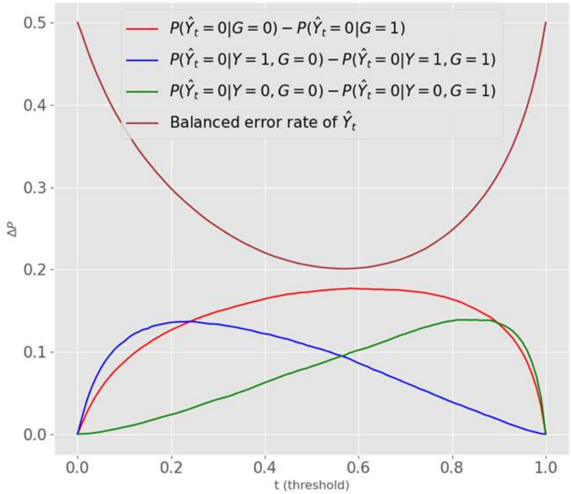
$$bias(Y_t|X,G) = |\mathbb{P}(Y_t = 0|G = 0) - \mathbb{P}(Y_t = 0|G = 1)|$$

Example (proxy predictor)

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







Fairness with awareness

Selected approaches for bias reduction in classifiers with access to protected attributes

- Maximization with fairness constraints

$$Y^*(X,G) \ \ or \ \ Y^*(X) = \max_{fairness(Y^*|G)} \mathbb{E}\left[\mathcal{L}\left(Y^*, X^{(train)}\right)\right]$$

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

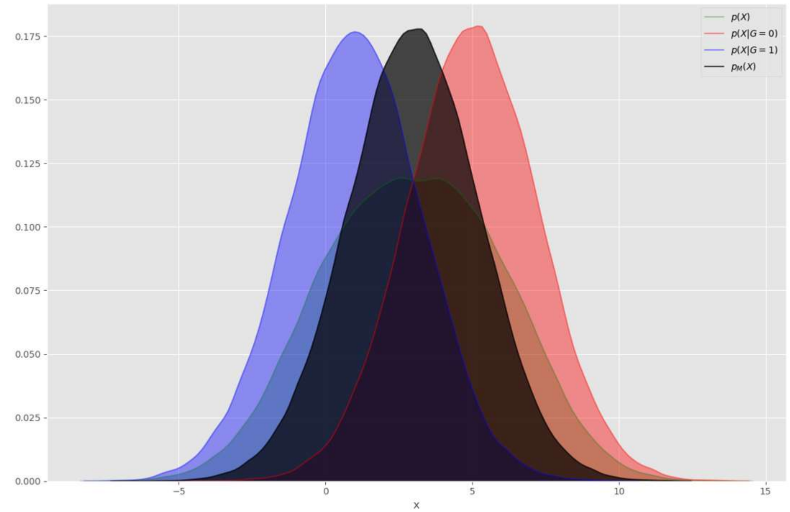
- Post-corrective methods (Hardt et al, 2015)
    - Study of equalized odds, equal opportunity, statistical parity
    - Construction of fair randomized classifier $\tilde{Y}(X,G;f) \in \mathcal{P}(\{0,1\})$ via post-processing

- Dataset repairment via optimal transport. Feldman et al (2015), Gordaliza et al. (2019).

Fairness with awareness

Fair dataset construction. Feldman et al (2015)

- Geometric repair: $X_i|G = k$ moving towards Wasserstein barycenter $\tilde{X}_i$.
- Training a classifier on repaired dataset $\tilde{X}(X, G, \lambda),\ \lambda \in [0,1]$
- Fairer predictors imply fairer classifier
- Useful when Y is not available



Random repair. Gordaliza et al (2019)

- Controlling the statistical parity bias via geometric repair is difficult
- Control must be via TV-distance:

$$bias^C(\hat{Y}|X, G) \leq d_{TV}(P_{X|G=0}, P_{X|G=1})$$

- Random repair picks at random, $Bernoulli(\lambda)$, between samples of $P_{X|G=k}$ and the barycenter of subpopulations:

$$bias^C(\hat{Y}|\tilde{X}_\lambda, G) \leq d_{TV}(P_{\tilde{X}|G=0}, P_{\tilde{X}|G=1}) = 1 - \lambda$$

Motivation

Comments

- Bias measurements test fairness of predictors X or a classifier $\hat{Y}$, not the regressor $f(X)$
- Mitigation procedures focus on the construction of fair classifiers $\hat{Y}^*(X,G)$, not a fair regressor.

Regulatory constraints. Fairness without awareness.

- G is typically not collected.
- Training with access to G is not allowed.
- Models (including post-processed ones) $f(X,G)$ that require access to G are not allowed.

Proxy models of G for validation

- Certain proxy models $\tilde{G}$ for G are allowed for validation by compliance office. $\tilde{G}$ is prohibited to share outside of it.
- Postprocessing is possible by compliance but the model $\tilde{f}(X)$ must rely on X only. No leakage of $(X,\tilde{G})$ is allowed.

Objectives of our work

Given a trained model regressor or classification score $f(X)$:

1. Measurement. Evaluate regressor bias.
2. Bias Interpretability. Quantify the contribution of each predictor to that bias.
3. Mitigation. Produce family of post-processed models $\{\tilde{f}_\alpha(X; f)\}$ using a proxy model $\tilde{G}$.

Regressor bias

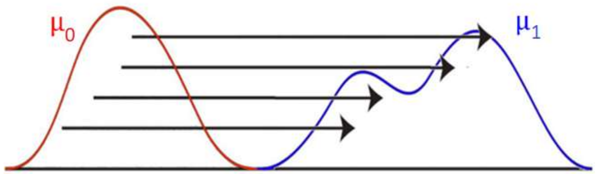
Model (regressor) bias

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

- Wasserstein metric W_1 (optimal transport cost)

$$W_1(\mu_0,\mu_1) = \inf_{\pi \in \mathcal{P}(\mathcal{Z}^2)} \{ \int |z_1 - z_2| \, \pi(dz_1, dz_2), \ \pi \text{ with marginals } \mu_0, \mu_1 \}$$

- $\mu_k = P_{f(X)|G=k}, \mathrm{k} \in \{0,1\}$



Note: The main focus is on the bias in the output (model), not the input (predictors).

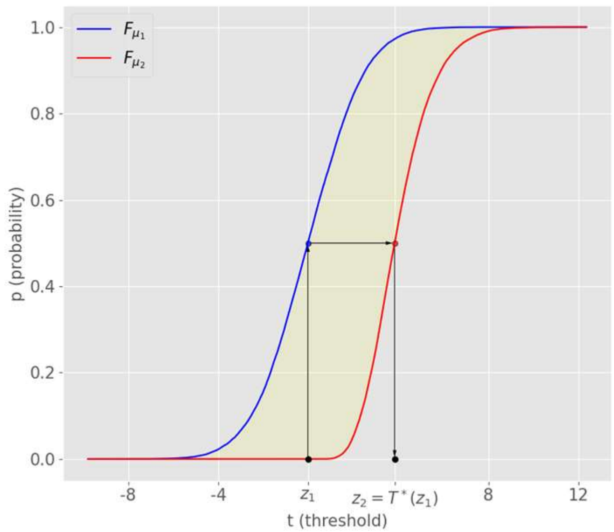
Model bias metrics

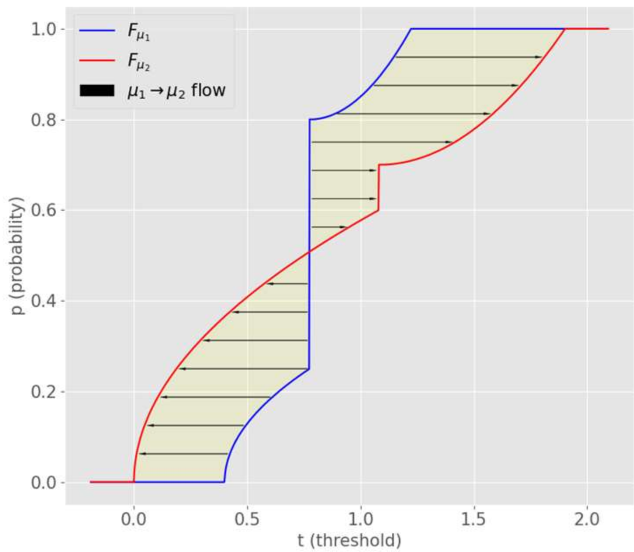
Basic properties (one-dimension)

- μ_1, μ_2 on $\mathcal{B}(\mathbb{R})$, there exists order preserving optimal transport plan π^* such that

$$W_1(\mu_1, \mu_2) = \int |x_1 - x_2|\, d\pi^* = \int \left|F_{\mu_1}^{[-1]}(p) - F_{\mu_2}^{[-1]}(p)\right| dp = \int \left|F_{\mu_1}(t) - F_{\mu_2}(t)\right| dt$$

- Transport map vs transport plan:





Positive and negative flows

Need to understand whether the model favors majority class or minority one.

Assumption: Model $f(X) \in \mathbb{R}$ has a favorable direction $\varsigma_f = \pm 1$.

Definition: $Bias^{\pm}_{W_1}(f|X,G)$ is the cost of transporting $\mathrm{P}_{f(X)|G=0}$ in favorable/non-favorable directions.

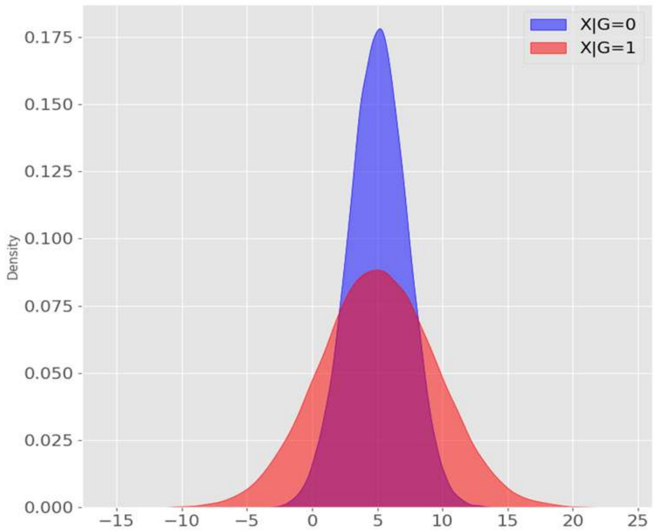
Example:

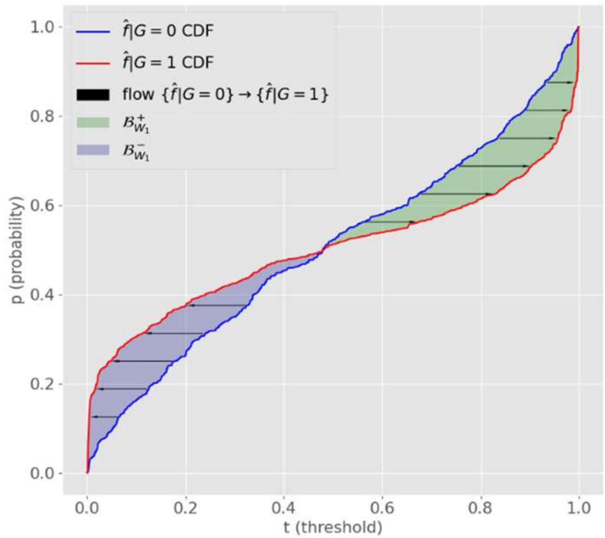
$X \sim \mathcal{N}(\mu, (1+G)\sqrt{\mu})$

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

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

$\zeta_f = -1$





Model bias metrics

Facts [Miroshnikov et al, 2021a]

- Integrated statistical parity bias:
  - $Bias_{W_1}(f|X,G) = \int bias(Y_t|X,G)dt$
  - $Bias^{\pm}_{W_1}(f|X,G) = \int_{\mathcal{T}_{\pm}} bias(Y_t|X,G)\, dt$

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

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

Input-output bias relationship

- Bias in predictors propagates through the model:

$$Bias_{W_1}(f|X,G) \leq [f]_{Lip} W_1(X|G=0, X|G=1)$$

- Fairness of predictors is sufficient for model fairness, but not necessary:

$$X_1 \sim N(\sqrt{\tau} \cdot G, 1),\ X_2 \sim N(1,1),\ \ Y = \frac{1}{\tau} X_1 + X_2$$

Here $Bias(Y|X,G) \to 0$ and $Bias(X|G) \to \infty$ as $\tau \to \infty$.

- We would like to understand how each predictor contributes to the model bias $Bias_{W_1}(f|X,G)$.

Model explanations

To design fairness interpretability we first review model explanations.

Basic post-hoc model explainers.

Given f and $X \in \mathbb{R}^n$, the contribution of X_i to $f(X)$ can be quantified by

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

Note: Marginal explains $x \to f(x)$ and the conditional $X(\omega) \to f(X(\omega))$.

Model explanations

Post-hoc explainers (game-theoretical)

- Players: $N = \{1,2,\dots,n\}$ (features become player)
- Game: set function $v(S)$, $S \subset N$, $v(N) =$ total payoff
- Shapley value [Shapley, 1953]

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

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

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 $\beta_i^{\pm}$ are defined as transport effort in favorable and non-favorable directions: $\beta_i^{\pm} = \int_{\mathcal{P}_{i\pm}} \left| F_{E_i|G=0}^{[-1]}(p) - F_{E_i|G=1}^{[-1]}(p) \right| dp$

Notes

- Type of ML explainers matters (marginal vs conditional)
- β_+ quantifies the positive contribution (increase in positive flow and decrease in negative)

Fairness Interpretability

Example: basic bias explanations based on marginal Shapley model explainer

$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$

$X_1 \sim \mathcal{N}(\mu - a_1(1 - G), 0.5 + G)$

$X_2 \sim \mathcal{N}(\mu - a_2(1 - G), 1)$

$X_3 \sim \mathcal{N}(\mu - a_3(1 - G), 1)$

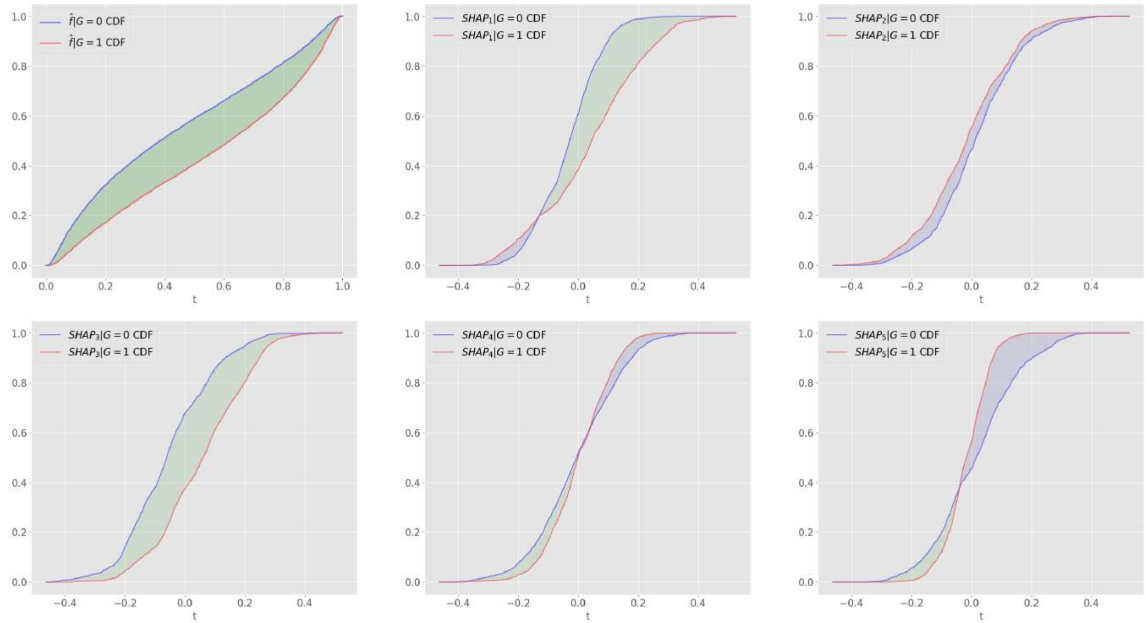
$X_4 \sim \mathcal{N}(\mu - a_4(1 - G), 1 - 0.5G)$

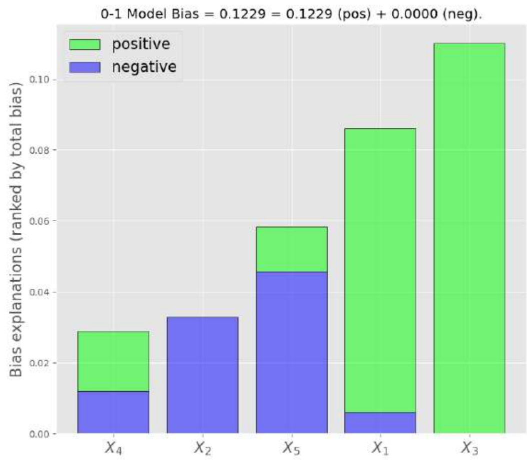
$X_5 \sim \mathcal{N}(\mu - a_5(1 - G), 1 - 0.75G)$

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

$f(X) = \sigma(\sum X_i - 24.5)$

$\varsigma_f = -1$

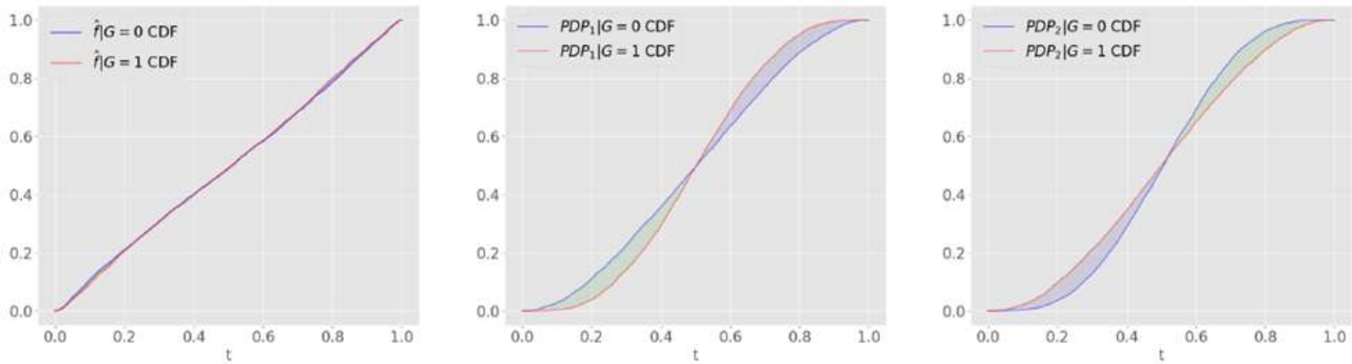




Fairness Interpretability

Example (bias offsetting)

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



Fairness Interpretability

- Basic bias explanations are not additive.
- Do not explain the direct contribution to the negative and positive model bias.

Game theoretical approach

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

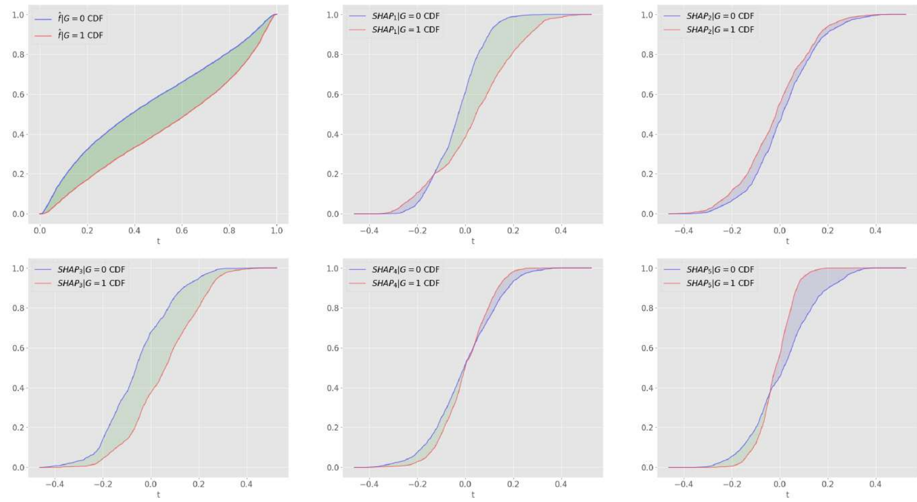
$$Bias^{\pm}_{W_1}(f|X,G) = \sum_i \varphi^{bias\pm}(f|X,G)$$

Note: explanations are signed and additive

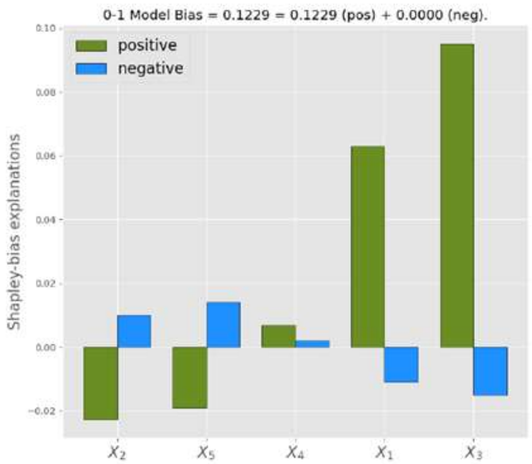
Fairness Interpretability

Example (marginal Shapley-bias explanations)

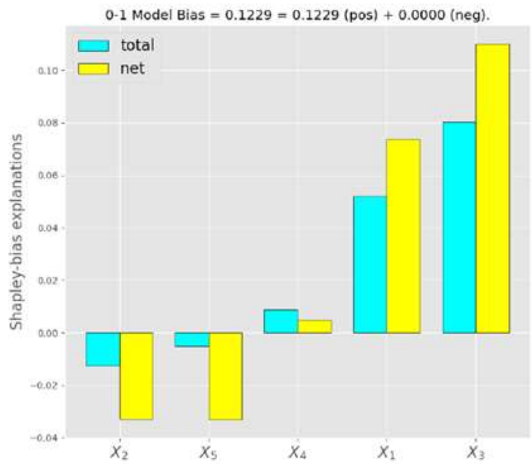
$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$
$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$
$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$
$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$
$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$
$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$
$Y \sim Bernoulli(f(X)),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



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



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



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

Bias mitigation

Superposition [Miroshnikov et al, 2021c]

$$Bias_{W_1}(f|X,G) = \sum\bar{\beta}_i^{++} + \sum\bar{\beta}_i^{-+} - \sum\bar{\beta}_i^{+-} - \sum\bar{\beta}_i^{--} \geq 0$$

with $\bar{\beta}_i^{\pm+} = \max(\varphi_i[v^{bias\pm}], 0)$, $\bar{\beta}_i^{\pm-} = \max(-\varphi_i[v^{bias\pm}], 0)$

Special case (typical one)

Let f be positively-biased model, that is, $Bias_{W_1}^{+}(f|X,G) > 0$, $Bias_{W_1}^{-}(f|X,G) = 0$. Then

$$Bias_{W_1}(f|X,G) = \sum\bar{\beta}_i^{+} - \sum\bar{\beta}_i^{-} \geq 0$$

where $\bar{\beta}_i^{+} = \bar{\beta}_i^{++} + \bar{\beta}_i^{--}$, $\bar{\beta}_i^{-} = \bar{\beta}_i^{-+} + \bar{\beta}_i^{+-}$.

Note: This expression is the key to the bias mitigation procedure.

Bias mitigation

The relationship $Bias_{W_1}(f|X,G) = \sum\bar{\beta}_i^+ - \sum\bar{\beta}_i^- \geq 0$ is the key for bias mitigation via postprocessing:

1. Predictors with insignificant bias explanations are not relevant. This reduces dimensionality.

2. Adjusting the model so that $\bar{\beta}_i^+ \downarrow$ and $\bar{\beta}_i^- \uparrow$ should lead to model bias decrease.

Question: How to construct a postprocessed model $\tilde{f}(X;f)$ that does not rely on (X,G)?

Bias mitigation

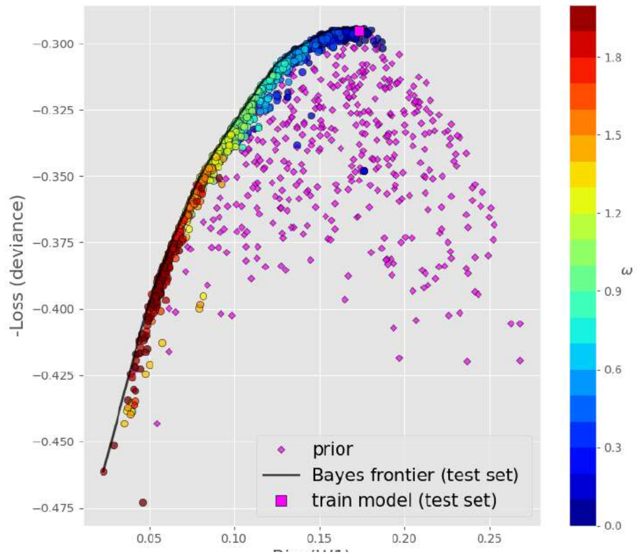
Efficient frontier via rebalancing [Miroshnikov et al 2021c]

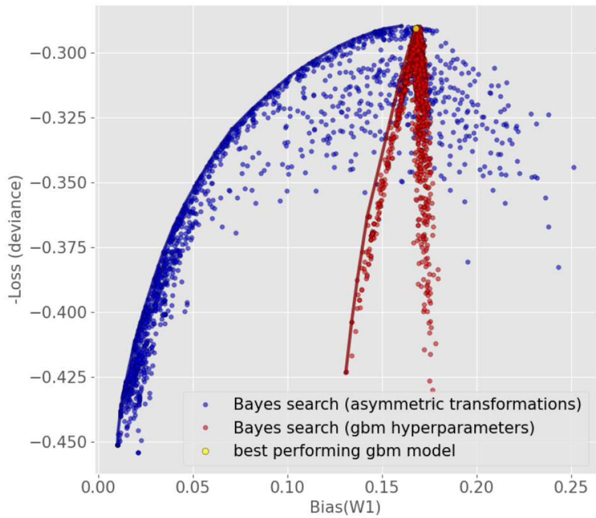
- $M = \{i_1, i_2, \dots i_m\}$ most bias impactful predictors
- $\mathcal{F} = \{\tilde{f} : \tilde{f} = \mathcal{C}[f(T(X_M; \alpha), X_{-M})], \alpha \in A \subset \mathbb{R}^{mk}\}$
- $T(\cdot; \alpha)$ adjusts each predictor appropriately (scaling)
- $\mathcal{C}[\cdot]$ calibrates the distribution
- Efficient frontier is recovered by solving:

$$\alpha_*(\omega) = argmin_{\tilde{f}} \{\mathbb{E}[L(Y, \tilde{f})] + \omega \cdot Bias_{W_1}(\tilde{f} | X, G)\}$$

Strategies for choosing M

1. Given m_*: $N_{\pm} = \{i: m_*\text{-highest } \beta_i^{\pm}\}$. Set $M = N_+ \cup N_-$.
2. Given m_*: $M = \{i: m_*\text{-highest } \beta_i\}$. Set $M = N_+ \cup N_-$.

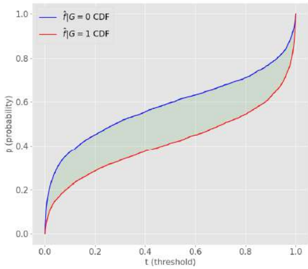


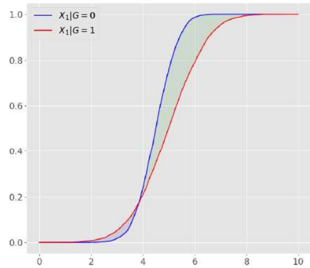


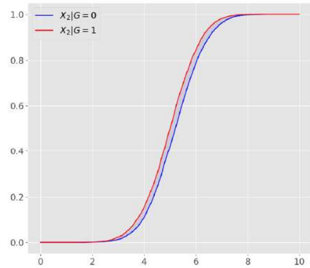
Bias mitigation

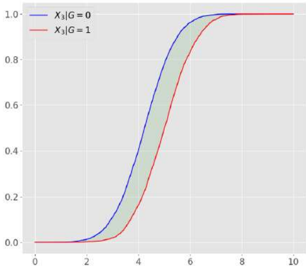
Example

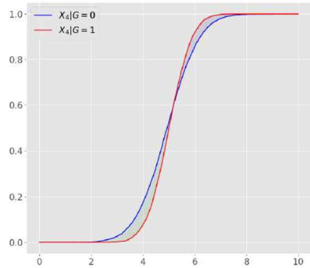
$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$
$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$
$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$
$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$
$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$
$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$
$Y \sim Bernoulli(f(X)), f(X) = \sigma\left(2(\sum X_i - 24.5)\right)$

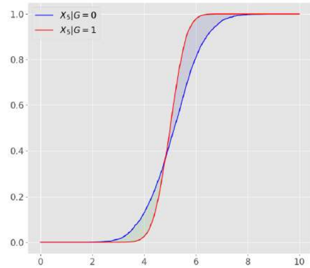

(a) Subpopulation distributions


(b) X_1 CDFs


(c) X_2 CDFs

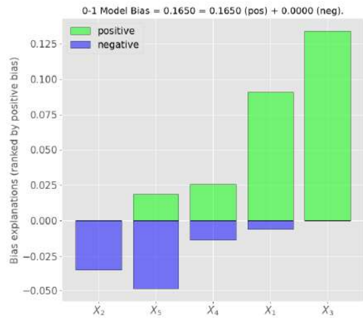

(d) X_3 CDFs


(e) X_4 CDFs

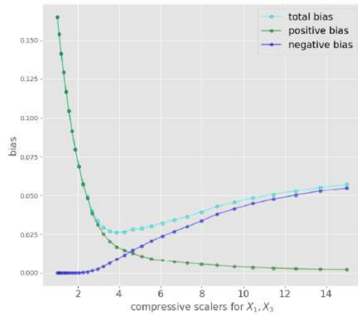

(f) X_5 CDFs

Effect of compression:

- Compressing X_1, X_3 via a compressive map $T(x_i; x_i^*)$
- Set $\tilde{f} = f(T(X_1; x_1^*), X_2, T(X_3; x_3^*), X_4, X_5)$, $x_i^* = \mathbb{E}[X_i]$



(g) Bias explanations



(h) Change in bias

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)
- Melinda Milenkovich (VP & Assistant General 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)

References

- C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R.S. Zemel, Fairness through awareness. In Proc. ACM ITCS, 214-226, (2012).
- M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In Proc. 21st ACM SIGKDD, 259-268, (2015).
- J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics, Vol. 29, No. 5, 1189-1232, (2001).
- P. Gordaliza, E. D. Barrio, G. Fabrice, J.-M. Loubes Obtaining Fairness using Optimal Transport Theory, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:2357-2365, (2019).
- P. Hall, B. Cox, S. Dickerson, A. Ravi Kannan, R. Kulkarni, N. Schmidt, A United States Fair Lending Perspective on Machine Learning. Front. Artif. Intell. 4:695301. doi: 10.3389/frai.2021.695301 (2021).
- M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 3315-3323, (2015).
- S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, 31st Conference on Neural Information Processing Systems, (2017).
- A. Miroshnikov, K. Kotsiopoulos, R. Franks, A. Ravi Kannan, Wasserstein-based fairness interpretability framework for machine learning models, *arXiv preprint* (2021a), arXiv:2011.03156.
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Mutual information-based group explainers with coalition structure for machine learning model explanations, *arXiv preprint* (2021b), arXiv:2102.10878.
- A. Miroshnikov, K. Kotsiopoulos, R. Franks, A. Ravi Kannan, Model-agnostic bias mitigation methods with regressor distribution control for Wasserstein-based fairness metrics, *arXiv preprint* (2021), arXiv:2111.11259
- A. Müller, Integral probability metrics and their generating classes of functions. Advances in Applied Probability,29(2):429–443, (1997).
- V. Perrone, M. Donini, K. Kenthapadi, Cedric Archambeau Fair Bayesian optimization, ICML AutoML Workshop. 2020
- L. S. Shapley, A value for n-person games, Annals of Mathematics Studies, No. 28, 307-317 (1953).
- E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst., 41, 3, 647-665, (2014).
- Schmidt, N., Curtis, J., Siskin, B., and Stocks, C. Methods for Mitigation of Algorithmic Bias Discrimination, Proxy Discrimination, and Disparate Impact. U.S. Provisional Patent 63/153,692, (2021).
- B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning nondiscriminatory predictors. In Proc. of Conference on Learning Theory, p. 1920–1953, (2017).
- B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating Unwanted Biases with Adversarial Learning. In Proc. of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340).