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

LANL, ML reading seminar, January 2022

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)*

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

Equalized odds *(Hardt et al, 2015)*

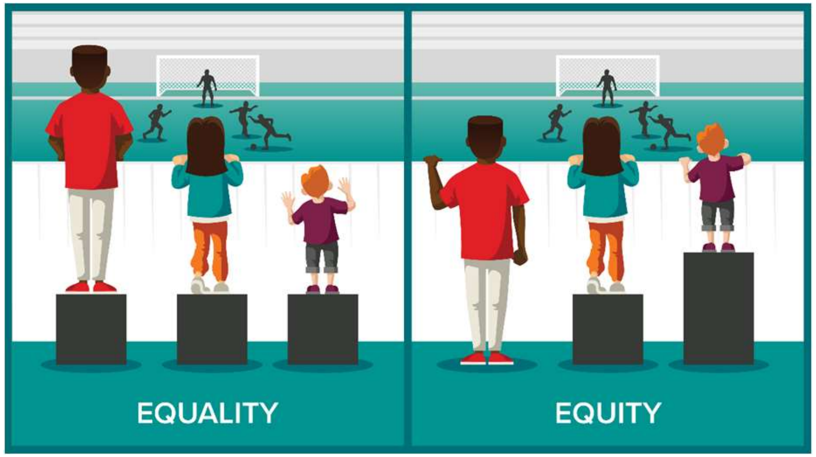
$$\mathbb{P}\big(\hat{Y} = 0 \big| Y = y, G = 0\big) = \mathbb{P}\big(\hat{Y} = 0 \big| Y = y, G = 1\big), y \in \{0,1\}$$

Equal opportunity *(Hardt et al, 2015)*

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

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

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



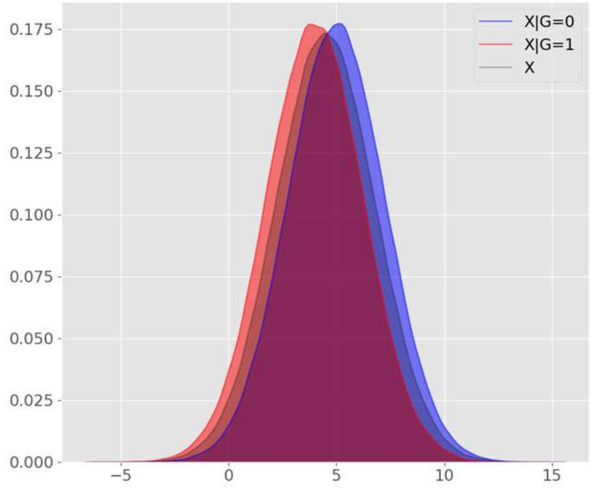
Fairness in classifiers

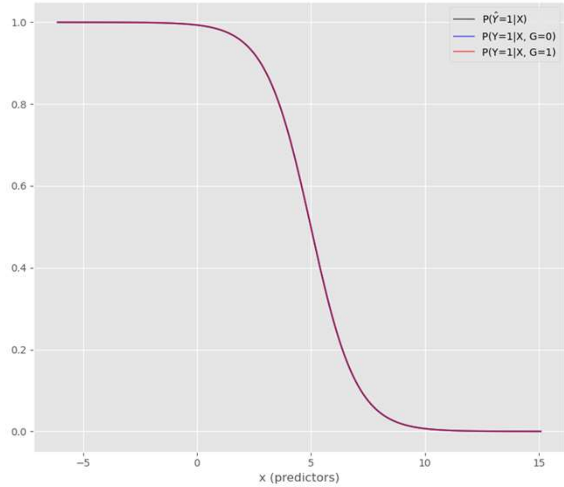
Statistical parity classifier bias

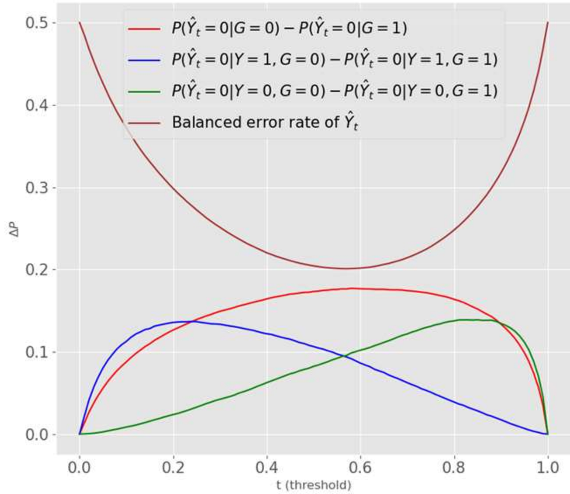
$$bias(Y_t|X,G) = |\mathbb{P}(Y_t = 0|G=0) - \mathbb{P}(Y_t = 0|G=1)|$$

Example (proxy predictor)

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







Fairness in classifiers

Approaches for bias mitigation

- Maximization with fairness constraints

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

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

- Post-corrective methods (Hardt et al, 2015)
  - 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

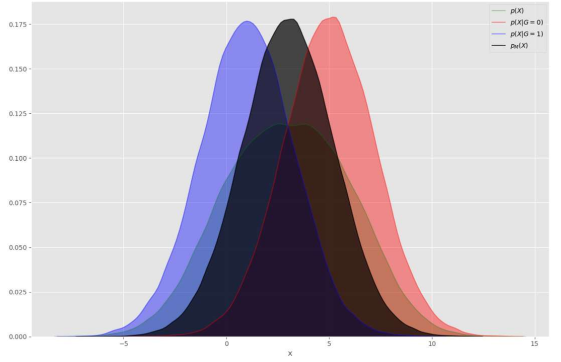
Fairness in classifiers

Approaches for bias mitigation

Fair dataset construction. Feldman et al (2015), Gordaliza et al. (2019)

Construction of partially fair $\tilde{X} = \tilde{X}(X, G; \lambda)$ using optimal transport, $\lambda \in [0,1]$

- Training a classifier on $\tilde{X}$
- Fairer predictors imply fairer classifier
- Useful when Y is not available

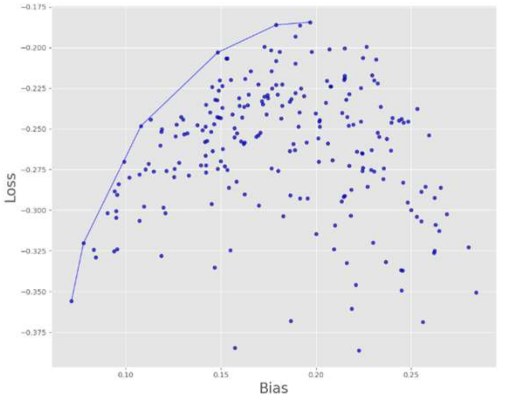


Classifier bias bounds:

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

For random repair Gordaliza et al. able to control $P_{\tilde{X}|G=0}, P_{\tilde{X}|G=1}$

- Pareto efficient frontier. Schmidt and Stephens (2019), Perrone et al (2020).
  - Trained models $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$
  - Construction of the frontier over $\mathcal{F}$



Motivation

Issues

- Typical bias measurements test fairness of a classifier Y_t, not the regressor $f(X)$
- 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

Regulatory constraints

- Explicit use of the protected attribute G is not allowed by ECOA neither in **training** nor **prediction**:
  - Training with fairness constrains or repairing predictors and then training is not allowed
  - Postprocessing models in the form $f(X, G)$ are not allowed
  - Using information on (X, G), for instance $\mathbb{P}(G|X)$, in prediction, is not allowed

Motivation

Stages of model development

1. Model training with access to X
2. Fairness assessment on (X, G) and (appropriate) post-processing
3. Prediction with access to X

Note: Post-processed model must depend on X and should not allow one to learn (X, G), including $\mathbb{P}(G|X)$.

Acceptable form of bias mitigation under regulatory settings

1. Given the regressor 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_M , $M = \{i_1, i_2, \dots, i_m\}$
3. Construct a post-processed model $\tilde{f}(X; f, X_M)$ that does not rely on G and does not leak information on (X, G)

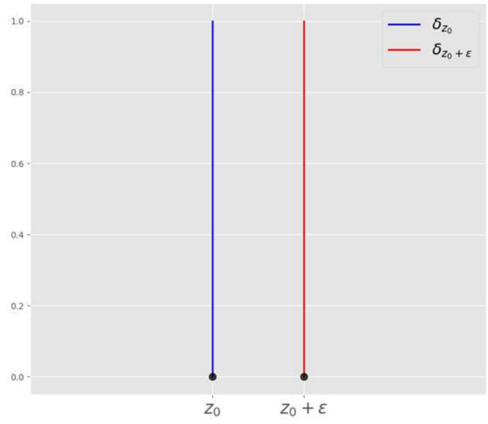
Model bias metrics for regressors

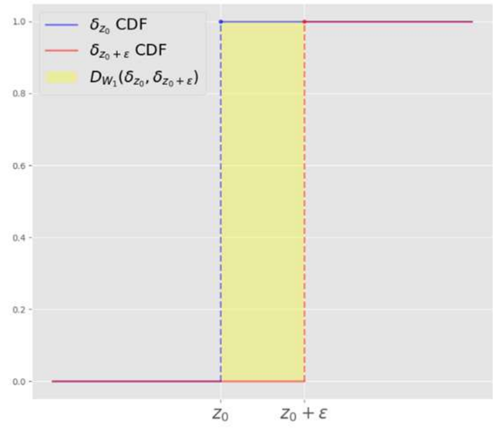
At an algorithmic level, the bias can be viewed as an ability to differentiate between two subpopulations at the level of data or outcomes.

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 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

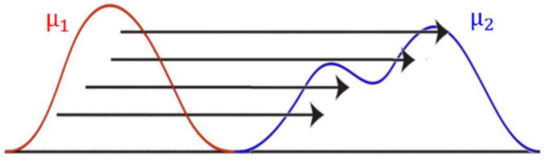
μ_1, μ_2 probability measures on a metric space $\mathcal{Z}$ equipped with a metric $d(z_1, z_2)$.

- 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 $P_{f(X)|G=k}$, $k = 0,1$.

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

Model bias metrics

Facts:

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

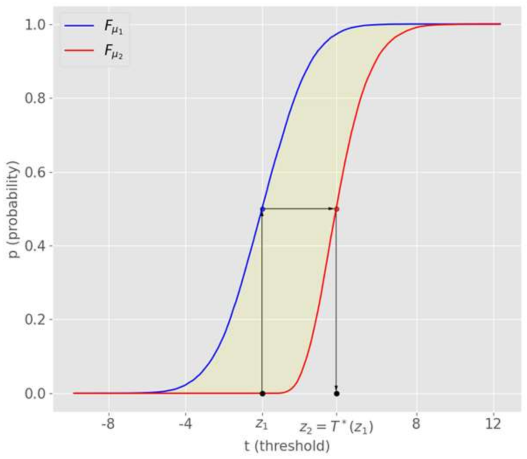
$$Bias_{d,D_{TV}}(\mu_1,\mu_2) = W_1(\mu_1,\mu_2; d)$$

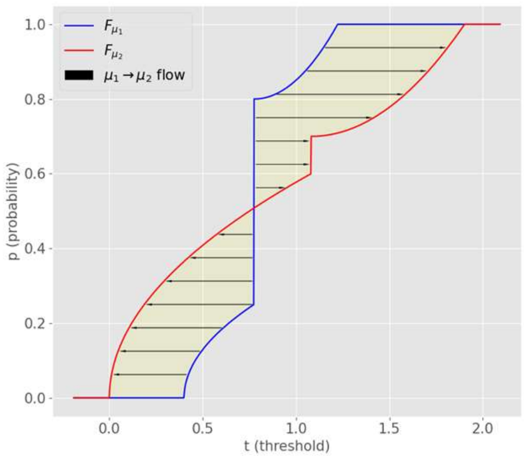
- (Miroshnikov et al 2021a): for any μ_1, μ_2 with support in $B_L(z_*)$ and $d(z_1, z_2) = \|z_1 - z_2\|$

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

- 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^*

$$W_1(\mu_1,\mu_2) = \int |x_1 - x_2| \, d\pi^* = \int \left| F_{\mu_1}^{[-1]}(p) - F_{\mu_2}^{[-1]}(p) \right| dp = [\text{Shorack, 1956}] = \int \left| F_{\mu_1}(t) - F_{\mu_2}(t) \right| dt$$





Model bias metrics

Model bias definition

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

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

Facts [Miroshnikov et al, 2021a]

- Connection with statistical parity:

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

- 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 $\leftarrow$)

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

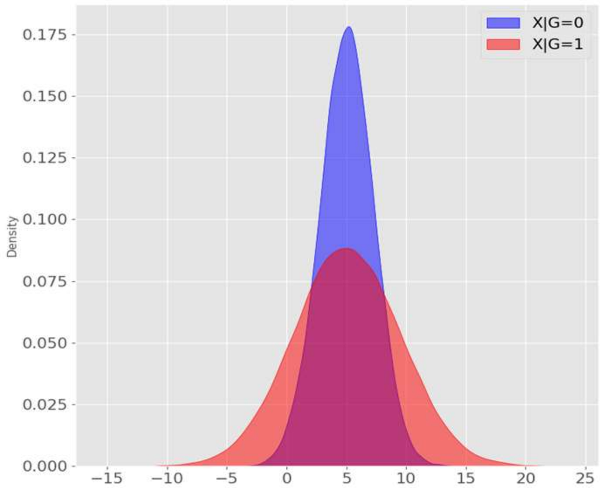
Example

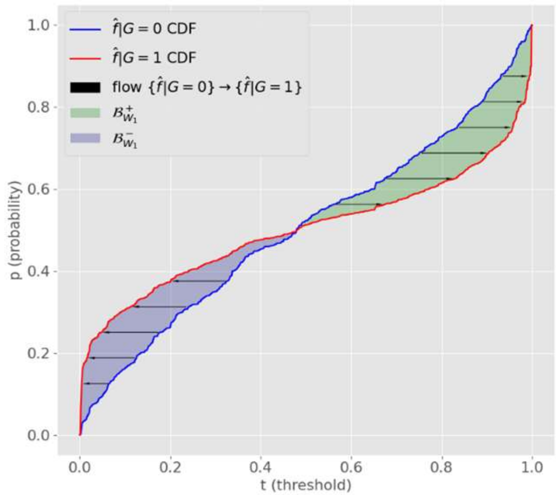
$X \sim \mathcal{N}(\mu, (1+G)\sqrt{\mu})$

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

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

$\zeta_f = -1$





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)

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

ML Interpretability

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]

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

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

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.

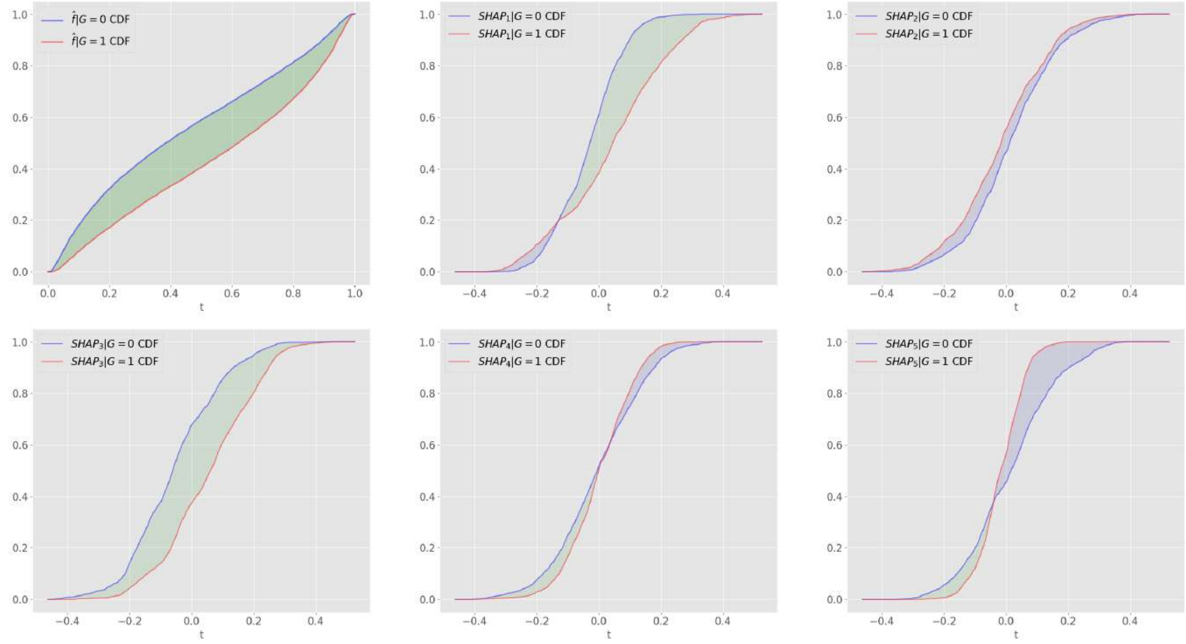
Notes

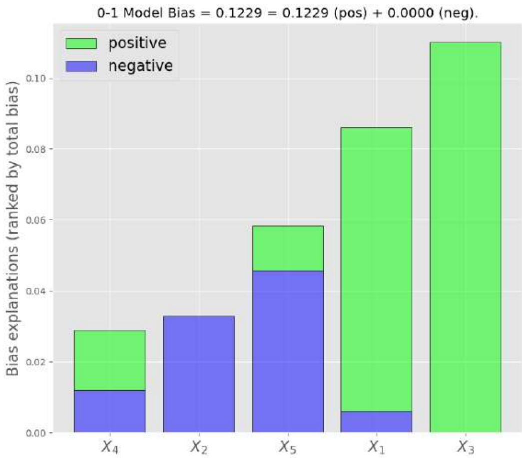
- Type of ML explainers matters (marginal vs conditional)
- Some ML explainers isolate the effect of each predictor and some not (local vs global)

Fairness Interpretability

Example: bias explanations based on marginal Shapley values

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

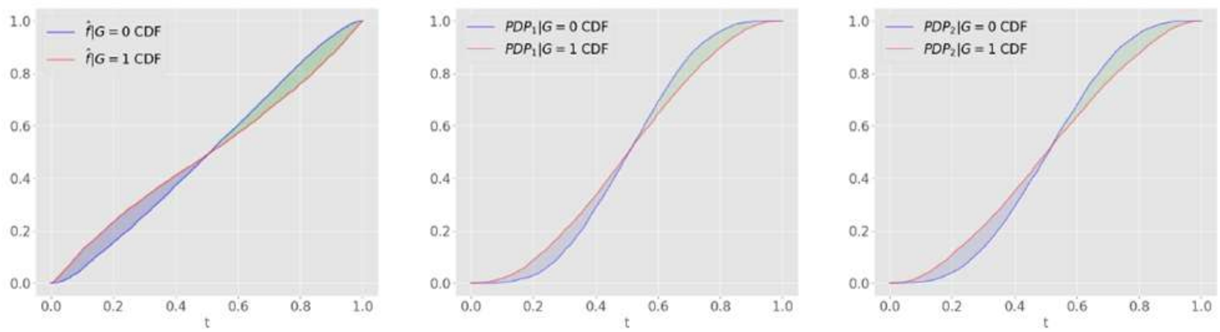




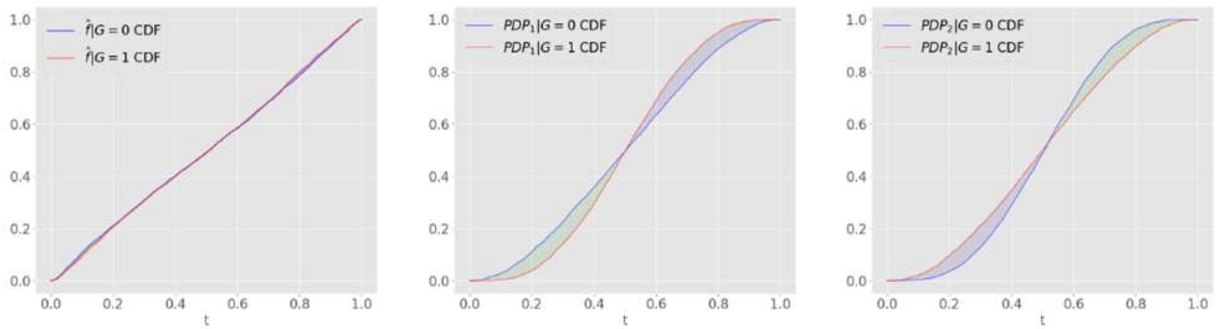
Fairness Interpretability

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

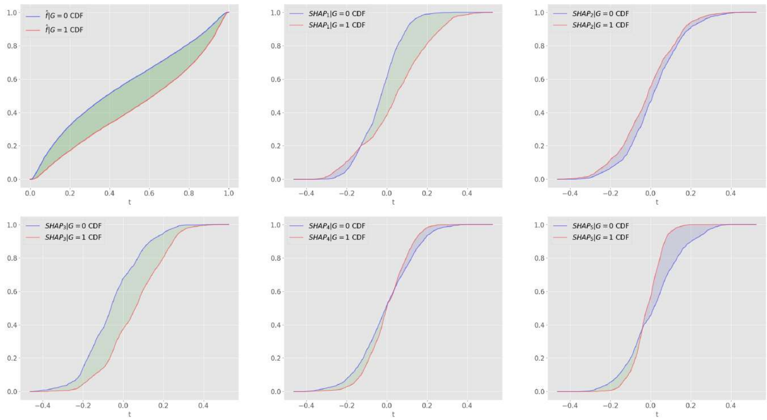
- Consider an ML explainer $E_S(X; f)$ of predictor X_S, $S \subset \{1,2, \ldots n\}$
- Predictors $\{X_i\}_{i \in N}$ are players that push/pull explainer subpopulation distributions apart when joining a coalition $S \subset N$
- A game $v^{bias}(S) = Bias_{W_1}(E_S(X)|G) = W_1(E_S(X)|G = 0, E_S(X)|G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

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

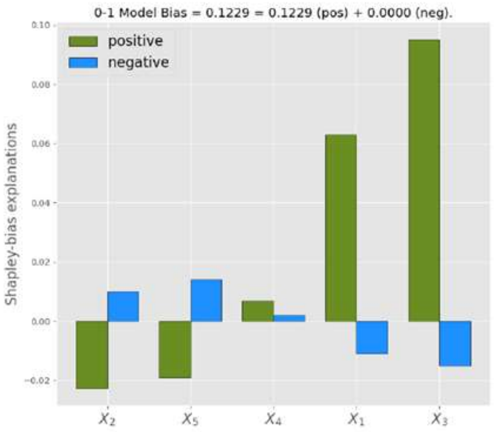
Fairness Interpretability

Example (marginal Shapley-bias explanations)

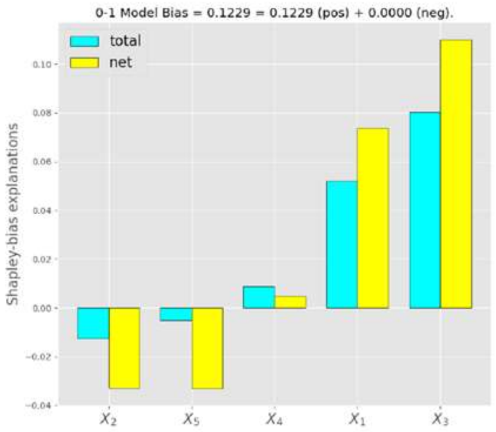
$$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$$
$$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$$
$$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$$
$$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$$
$$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$$
$$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$$
$$Y \sim Bernoulli\big(f(X)\big),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



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



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



Application

Superposition [Miroshnikov et al, 2021c]

Set

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

Then

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

Special case

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

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

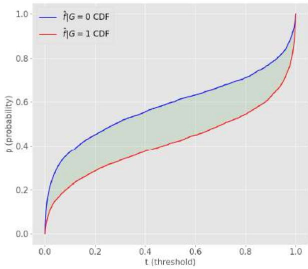
where $\bar{\beta}_i^{+} = \bar{\beta}_i^{++} + \bar{\beta}_i^{--}$, $\bar{\beta}_i^{-} = \bar{\beta}_i^{-+} + \bar{\beta}_i^{+-}$ (positive and negative Shapley-bias explanations)

Note: for non-additive bias explanations the above relationship is true provided $f = \sum_i f_i$

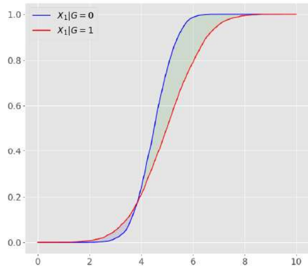
Application

Example

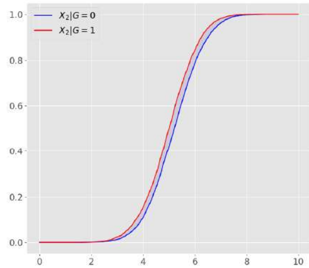
$$\mu = 5, a = \frac{1}{20}(10, -4, 16, 1, -3)$$
$$X_1 \sim \mathcal{N}(\mu - a_1(1-G), 0.5 + G)$$
$$X_2 \sim \mathcal{N}(\mu - a_2(1-G), 1)$$
$$X_3 \sim \mathcal{N}(\mu - a_3(1-G), 1)$$
$$X_4 \sim \mathcal{N}(\mu - a_4(1-G), 1 - 0.5G)$$
$$X_5 \sim \mathcal{N}(\mu - a_5(1-G), 1 - 0.75G)$$
$$Y \sim Bernoulli\big(f(X)\big),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



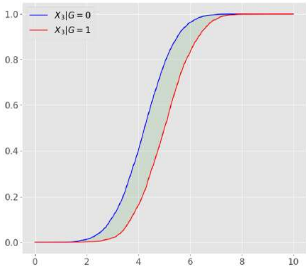
(a) Subpopulation distributions



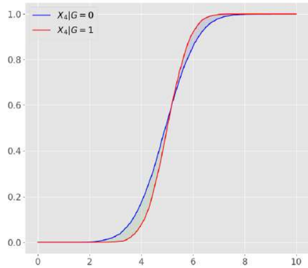
(b) X_1 CDFs



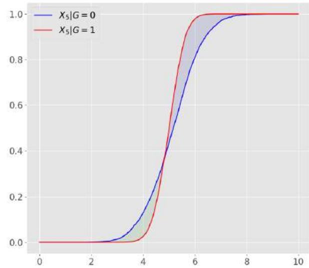
(c) X_2 CDFs



(d) X_3 CDFs



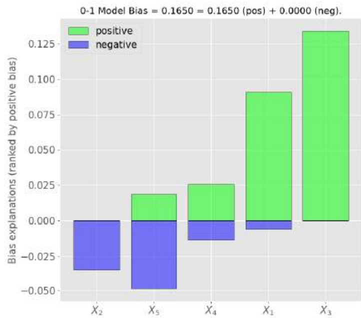
(e) X_4 CDFs



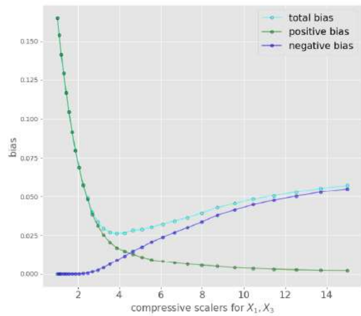
(f) X_5 CDFs

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

Application

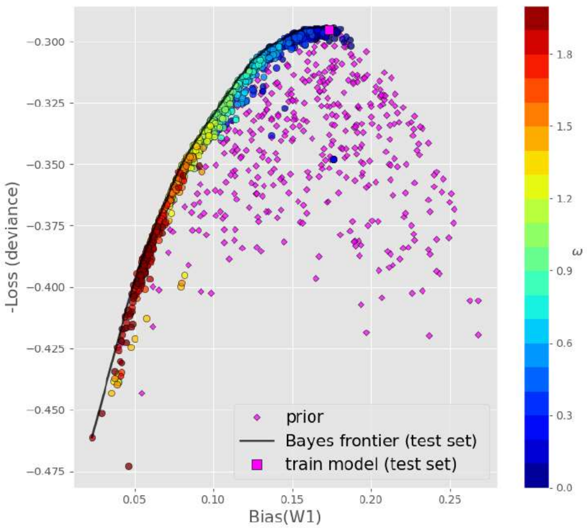
Efficient frontier via rebalancing [Miroshnikov et al 2021c]

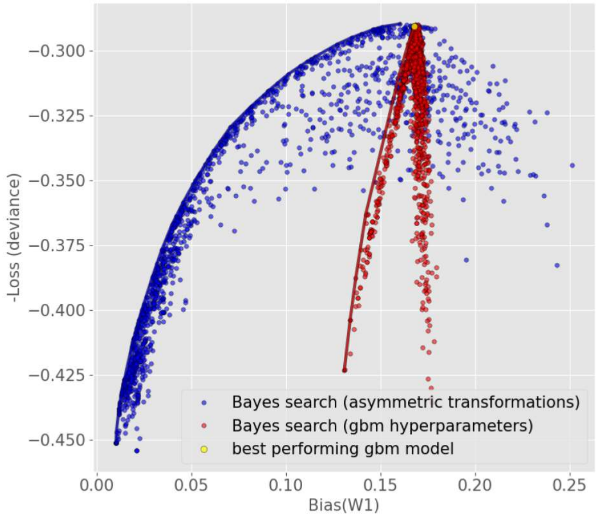
- $\mathcal{F} = \{\tilde{f}: \tilde{f} = \mathcal{C}[f(T(X_M;\alpha), X_{-M})], \alpha \in A \subset \mathbb{R}^{mk}\}$
- $T(\cdot;\alpha)$ adjusts each predictor appropriately
- $\mathcal{C}[\cdot]$ calibrates the distribution
- Efficient frontier is recovered by solving:

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

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 $M = N_+ \cup N_-$.





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)
- 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).