Machine Learning (2022) 111:3307–3357
https://doi.org/10.1007/s10994-022-06213-9

Wasserstein-based fairness interpretability framework for machine learning models

Alexey Miroshnikov[1] · Konstandinos Kotsiopoulos[1] · Ryan Franks[1] · Arjun Ravi Kannan[1]

Received: 17 April 2021 / Revised: 7 March 2022 / Accepted: 12 June 2022 /
Published online: 21 July 2022

Abstract
The objective of this article is to introduce a fairness interpretability framework for measuring and explaining the bias in classification and regression models at the level of a distribution. In our work, we measure the model bias across sub-population distributions in the model output using the Wasserstein metric. To properly quantify the contributions of predictors, we take into account favorability of both the model and predictors with respect to the non-protected class. The quantification is accomplished by the use of transport theory, which gives rise to the decomposition of the model bias and bias explanations to positive and negative contributions. To gain more insight into the role of favorability and allow for additivity of bias explanations, we adapt techniques from cooperative game theory.

Keywords Optimal transport · ML fairness · ML interpretability · Cooperative game

Mathematics subject classification 49Q22 · 91A12 · 68T01

1 Introduction

Contemporary machine learning (ML) techniques surpass traditional statistical methods in terms of their higher predictive power and their capability of processing a larger number of attributes. However, these novel ML algorithms generate models that have a

Editor: Scott Sanner.

✉ Alexey Miroshnikov
alexeymiroshnikov@discover.com

Konstandinos Kotsiopoulos
kostaskotsiopoulos@discover.com

Ryan Franks
ryanfranks@discover.com

Arjun Ravi Kannan
arjunravikannan@discover.com

[1] Emerging Capabilities, Discover Financial Services, Riverwoods, IL 60015, USA

complex structure which makes it difficult for their outputs to be interpreted with high precision. Another important issue is that a highly accurate predictive model might lack fairness by generating outputs that may result in discriminatory outcomes for protected subgroups. Thus, it is imperative to design predictive systems that are not only accurate but also achieve the desired fairness level.

When used in certain contexts, predictive models, and strategies that rely on such models, are subject to laws and regulations that ensure fairness. For instance, a hiring process in the United States (US) must comply with the Equal Employment Opportunity Act (EEOA 1972). Similarly, financial institutions (FI) in the US that are in the business of extending credit to applicants are subject to the Equal Credit Opportunity Act (ECOA 1974), the Fair Housing Act (FHA 1968), and other fair lending laws. These laws often specify protected attributes that FIs must consider when maintaining fairness in lending decisions.

Examples of protected attributes include race, gender, age, ethnicity, national origin, marital status, and others. Under the ECOA, for example, it is unlawful for a creditor to discriminate against an applicant for a loan on the basis of race, gender or age. Even though direct usage of protected attributes in building a model is often prohibited by law (e.g. overt discrimination), some otherwise benign attributes can serve as "proxies" because they may share dependencies with a protected attribute. For this reason, it is crucial for data scientists to conduct a fairness review of their trained models in consultation with compliance professionals in order to evaluate the predictive modeling system for potential unfairness. In this paper, we develop a fairness interpretability framework to aid in this important task.

At an algorithmic level, bias can be viewed as an ability to differentiate between two subpopulations at the level of data or outcomes. Regardless of its definition, if bias is present in data when training an ML model, the ability to differentiate between subgroups might potentially lead to discriminatory outcomes. For this reason, the model bias can be viewed as a measure of unfairness and hence its measurement is central to the model fairness assessment.

There is a comprehensive body of research on ML fairness that discusses bias measurements and mitigation methodologies. Kamiran et al. (2009) introduced a classification scheme for learning unbiased models by modifying the biased data sets without direct knowledge of the protected attribute. Kamishima et al. (2012) proposed a regularization approach for discriminative probabilistic models. Zemel et al. (2013) designed an optimization problem that incorporates fairness constraints. Feldman et al. (2015) proposed a geometric repair scheme to remove disparate impact in classifiers by making data sets unbiased. Hardt et al. (2015) indtroduced post-processing techniques removing discrimination in classifiers based on equalized odds and equal opportunity fairness criteria. Woodworth et al. (2017) designed a framework for nearly-optimal learning predictors with equalized odds fairness constraint. Zhang et al. (2018) proposed to use adversarial learning to mitigate bias, and Jiang (2020) suggested a bias correction technique via re-weighting the data.

The work of Dwork et al. (2012) studies Lipschitz randomized classifiers and their statistical parity bias. It establishes a bound on that bias by a transport-like distance between the input subpopulation distributions. The bound aids in constructing an optimal Lipschitz classifier with control over the statistical parity bias by transporting one of the subpopulation input datasets into the other. The work of Gordaliza et al. (2019) establishes a similar bound for non-randomized classifiers by the total variance distance between input subpopulation distributions. Guided by the bound and utilizing optimal transport theory, their method focuses on repairing input datasets in a way that allows for control of the total variance distance, and hence the statistical parity bias.

Though the bounds in the aforementioned works are of theoretical and practical importance, they provide little information on how each component of the input contributes to the bias in the output. The main reason for that is that the bias from the inputs propagates through the model structure in a non-trivial way. For this reason, in our work, we focus on designing a fairness interpretability framework that evaluates how each predictor contributes to the model bias, incorporating the predictor's favorability with respect to protected (or minority) class into the framework. The construction is carried out by employing optimal transport theory and game-theoretic techniques.

Another issue regarding the ML fairness literature is that it mainly focuses on classifiers. Specifically, given the data (X, G, Y), where $X \in \mathbb{R}^n$ are predictors, $G \in \{0, 1\}$ is a protected attribute and $Y \in \{0, 1\}$ is a binary output variable, with favorable outcome $Y = 1$, the bias measurements are often based on fairness criteria such as statistical parity, which reads $\mathbb{P}(\hat{Y} = 1|G = 0) = \mathbb{P}(\hat{Y} = 1|G = 1)$, or alternative criteria such as equalized odds and equal opportunity (Feldman et al. 2015; Hardt et al. 2015).

Many models in the financial industry, however, are regressors $f = \mathbb{E}[Y|X]$. In turn, classification models are usually obtained by thresholding the regressor, $Y_t(X) = \mathbb{1}_{\{f(X)>t\}}$, but the thresholds are in general not chosen during the model development stage. Thus, data scientists select the classification score $f(X) = \widehat{\mathbb{P}}(Y = 1|X)$ based on the overall performance across all thresholds. The same is true for fairness assessment, which is conducted at the level of the whole classification score. The main reason for this is that the strategies and decision-making procedures in FIs may rely on the classification score or its distribution, not a single classifier with a fixed threshold. This motivates us to measure and explain the bias exclusively in the regressor model.

Our interpretability framework in principle can be applied to a wide range of predictive ML systems. For instance, it can provide insight into predictor attributions for models that appear in economics, social sciences, medicine, and other fields.

Another application of the framework is for bias mitigation under regulatory constraints. In FIs, bias mitigation methodologies that require explicit consideration of protected class status in the training or prediction stages are not acceptable in view of ECOA. Consequently, bias mitigation methods such as those in Dwork et al. (2012); Feldman et al. (2015); Gordaliza et al. (2019) are not feasible. However, a probabilistic proxy model for a protected attribute G such as the Bayesian Improved Surname and Geocoding (BISG) is allowed to be used for fairness assessment and subsequent post-processing[1] (Elliot et al. 2009; Hall et al. 2021); for an alternative proxy model, see Chen et al. (2019). This setup allows for the use of our framework in the following regulatory-compliant fashion:

(S1) Given a model f and the proxy protected attribute $\tilde{G}$, perform a fairness assessment by measuring the bias across the subpopulation distributions $f(X)|\tilde{G} = k, k \in \{0, 1\}$.

(S2) If the model bias exceeds a certain threshold, determine the main drivers for the bias, that is, determine the list of predictors $X_{i_1}, X_{i_2}, \dots, X_{i_r}$ contributing the most to that bias.

(S3) Mitigate the bias by constructing a post-processed model $\tilde{f}(X;f)$ utilizing the information on the most biased predictors $\{X_{i_1}, X_{i_2}, \dots, X_{i_r}\}$ and without the direct use of $\tilde{G}$ or any information on the joint distribution $(X, \tilde{G})$.

[1] Compliance departments employ the proxy model for compliance purposes only.

In this article, the interpretability framework we develop addresses steps (S1) and (S2). The post-processing methods (S3) are investigated in our companion paper Miroshnikov et al. (2021b). In what follows, we provide a summary of the key ideas and main results.

Problem setup We consider the joint distribution (X, G, Y), where $X \in \mathbb{R}^n$ are predictors, $G \in \{0, 1\}$ is the protected attribute, with the non-protected class $G = 0$, and Y is either a response variable with values in $\mathbb{R}$ (not necessarily a continuous random variable) or binary one with values in $\{0, 1\}$. We denote a trained model by $f(x) = \widehat{\mathbb{E}}[Y|X = x]$, assumed to be trained on (X, Y) without access to G. We assume that there is a predetermined favorable model direction, denoted by ↑ and ↓; if the favorable direction is ↑ then the relationship $f(x) > f(y)$ favors the input x, and if it is ↓ the input y. In the case of binary $Y \in \{0, 1\}$, the favorable direction ↑ is equivalent to $Y = 1$ being a favorable outcome, and ↓ to $Y = 0$. To simplify the exposition, the main text focuses on the case of a binary protected attribute G. However, the framework and all of the results in the article have a natural extension to the multi-labeled case.

Key components of the framework

- Motivated by optimal transport theory, we focus on the bias measurement in the model output via the Wasserstein metric W_1

$$\mathrm{Bias}_{W_1}(f|G) = \inf_{\pi \in \mathscr{P}(\mathbb{R}^2)} \left\{ \int_{\mathbb{R}^2} |x_1 - x_2| \, d\pi(x_1, x_2), \quad \text{with marginals } P_{f(X)|G=0}, P_{f(X)|G=1} \right\},$$

  which measures the minimal cost of transporting one distribution into another; see Santambrogio (2015). More importantly, we introduce the model bias decomposition into the sum of the positive and negative model biases, $\mathrm{Bias}^{\pm}_{W_1}(f|G)$, which measure the transport effort for moving points of the unprotected subpopulation distribution $f(X)|G = 0$ in the non-favorable and favorable directions, respectively. This allows us to obtain a more informed perspective on the predictor's impact; see Sects. 3.2 and 3.3.
- We establish the connection of the model bias with that of a classifier. We show that the positive and negative model bias can be viewed as the integrated statistical parity bias over the family of classifiers induced by the regressor. This integral relationship is then used to construct an extended family of transport metrics for regressor bias. Via integration, these metrics incorporate generic group parity fairness criteria for classifiers induced by the given regressor. Furthermore, we prove a more general version of (Dwork et al. 2012, Theorem 3.3) that establishes the connection between the Wasserstein-based bias and the randomized classifier-based bias; see Sects. 3.3 and 3.4.
- We introduce bias predictor attributions called *bias explanations* in order to understand how predictors contribute to the model bias. The bias explanation β_i of predictor X_i is computed as the cost of transporting the distribution of $E_i|G = 0$ to that of $E_i|G = 1$, where $E_i(X;f)$ quantifies the contribution of X_i to the model value. The transport theory gives rise to the decomposition $\beta_i = \beta_i^+ + \beta_i^-$ into the sum of positive and negative model bias explanations. Roughly speaking, β_i^+ quantifies the combined predictor contribution to the increase of the positive model bias and decrease in the negative model bias, and vice versa for β_i^-; see Sect. 4.3.
- The bias explanations are in general not additive, even if the predictor explanations are. To construct additive bias explanations and to better capture the interactions at the distribution level, we employ a cooperative game theory approach motivated by the ideas of

Štrumbelj and Kononenko (2010). We design a cooperative bias game v^{bias} which evaluates the bias in the model attributed to coalitions X_S, $S \subset \{1, \dots, n\}$, and define bias explanations via the Shapley value $\varphi[v^{bias}]$, which yields additivity. Similar approach is applied to construct additive positive and negative bias explanations; see Sect. 4.5.

- We choose to design the bias explanations based upon model explainers E_i that are either conditional or marginal expectations, or game-theoretic explainers in the form of the Shapley value $\varphi[v]$ where v is either a conditional game v^{CE} or a marginal game v^{ME}. For each $v \in \{v^{CE}, v^{ME}\}$ we perform the stability analysis of non-additive and additive bias explanations. By adapting the grouping techniques from Miroshnikov et al. (2021a), we reduce the complexity of game-theoretic bias explanations and unite marginal and conditional approaches; see Sects. 4.4, 4.5 and 4.6.

Structure of the paper. In Sect. 2, we introduce the requisite notation and fairness criteria for classifiers, and discuss ML fairness literature related to our work. In Sect. 3, we introduce the Wasserstein-based regressor bias and investigate its properties. In addition, we discuss a wide class of transport metrics that could be used for fairness assessment. In Sect. 4, we provide a theoretical characterization of the bias explanations and investigate their properties. In Sect. 5 we discuss some regulatory aspects of bias mitigation, and present an application of the framework to a UCI dataset. In Appendix A, we discuss the Kantorovich transport problem. In Appendix B, we state and prove auxiliary lemmas.

2 Preliminaries

2.1 Notation and hypotheses

We consider the joint distribution (X, G, Y), where $X = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ are the predictors, $G \in \{0, 1, \dots, K-1\}$ is the protected attribute and Y is either a response variable with values in $\mathbb{R}$ (not necessarily a continuous random variable) or a binary one with values in $\{0, 1\}$. We encode the non-protected class as $G = 0$ and assume that all random variables are defined on the common probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Ω is a sample space, $\mathbb{P}$ a probability measure, and $\mathscr{F}$ a σ-algebra of sets.

The true model and a trained one, which is assumed to be trained without access to G, are denoted by

$$f(X) = \mathbb{E}[Y|X] \quad \text{and} \quad \hat{f}(X) = \widehat{\mathbb{E}}[Y|X],$$

respectively. In the case of binary Y they read $f(X) = \mathbb{P}(Y = 1|X)$ and $\hat{f}(X) = \widehat{\mathbb{P}}(Y = 1|X)$. We denote a classifier based on the trained model by

$$\widehat{Y}_t = \widehat{Y}_t(X;\hat{f}) = \mathbb{1}_{\{\hat{f}(X)>t\}}, \quad t \in \mathbb{R}.$$

The subpopulation cumulative distribution function (CDF) of $\hat{f}(X)|G = k$ is denoted by

$$F_k(t) = F_{\hat{f}(X)|G=k}(t) = \mathbb{P}(\hat{f}(X) \le t|G = k)$$

and the corresponding generalized inverse (or quantile function) $F_k^{[-1]}$ is defined by:

$$F_k^{[-1]}(p) = F_{\hat{f}(X)|G=k}^{[-1]}(p) = \inf_{x \in \mathbb{R}} \{p \le F_k(x)\}.$$

We assume that there is a predetermined *favorable model direction*, denoted by either ↑ or ↓. If the favorable direction is ↑ then the relationship $f(x) > f(z)$ favors the input x, and if it is ↓ the input z. The sign of the favorable direction of f is denoted by ς_f and satisfies

$$\varsigma_f = \begin{cases} 1, & \text{if the favorable direction of} f \text{is} \uparrow \\ -1, & \text{if the favorable direction of} f \text{is} \downarrow \end{cases}.$$

In the case of binary Y, the favorable direction ↑ is equivalent to $Y = 1$ being a favorable outcome, and ↓ to $Y = 0$; see Sect. 2.4.

In what follows we first develop the framework in the context of the binary protected attribute $G \in \{0, 1\}$ and then extend it to the case of the multi-labeled protected attribute; see Sect. 3.4.

2.2 Fairness criteria for classifiers

When undesired biases concerning demographic groups (or protected attributes) are in the training data, well-trained models will reflect those biases. There have been numerous articles devoted to ML systems that lead to fair decisions. In these works, various measurements for fairness have been suggested. In what follows, we describe several well-known definitions which help measure fairness of classifiers.

Definition 1 Suppose that Y is binary with values in $\{0, 1\}$ and $Y = 1$ is the favorable outcome. Let $\widehat{Y}$ be a classifier.

- $\widehat{Y}$ satisfies statistical parity (Feldman et al. 2015) if

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

- $\widehat{Y}$ satisfies equalized odds (Hardt et al. 2015) if

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

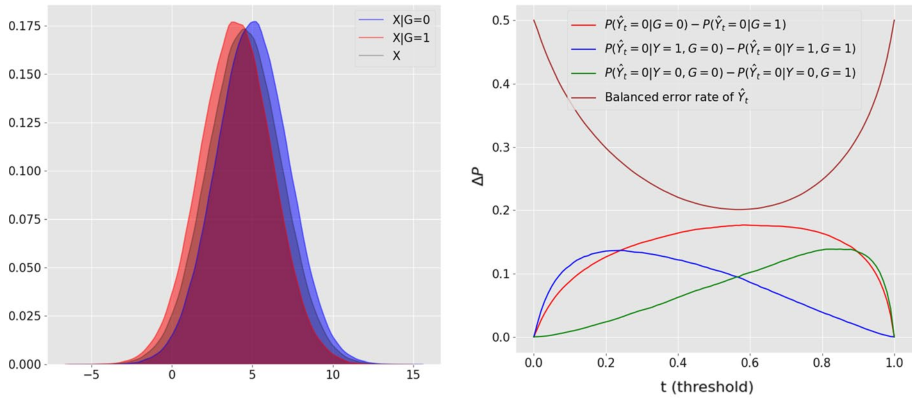
- $\widehat{Y}$ satisfies equal opportunity (Hardt et al. 2015) if

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

- The balanced error rate (BER) of $\widehat{Y}$ (Feldman et al. 2015) is given by

$$BER(\widehat{Y}, G) = \tfrac{1}{2}(\mathbb{P}(\widehat{Y} = 1|G = 0) + \mathbb{P}(\widehat{Y} = 0|G = 1)).$$

The statistical parity requires that the proportions of people in the favorable class $\widehat{Y} = 1$ within each group $G = k, k \in \{0, 1\}$ are the same. The equalized odds constraint requires the classifier to have the same misclassification error rates for each class of the protected attribute G and the label Y. Equal opportunity constraint requires the misclassification rates to be the same for each class $G = k$ only for the individuals labeled as $Y = 1$. The BER is the average class-conditioned error rate of $\widehat{Y}$.



(a) Subpopulation distributions of X.

(b) Fairness measurements.

Fig. 1 Predictor distributions and fairness for the model (M1), $\varsigma_f = -1$

2.3 Group classifier fairness example

There are numerous reasons why a trained classifier may lead to unfair outcomes. To illustrate, we provide an instructive example that shows how predictors and labels, as well as their relationship with the protected attribute, affect classifier fairness.

Consider a data set (X, Y, G) where the predictor X depends on $G \in \{0, 1\}$, $Y \in \{0, 1\}$ is binary, with favorable outcome $Y = 0$, and the classification score f depends explicitly on X only:

$$\begin{aligned} &X \sim N(\mu - a \cdot G, \sqrt{\mu}), \quad \mu = 5, a = 1 \\ &Y \sim Bernoulli(f(X)), \quad f(X) = \mathbb{P}(Y = 1|X) = logistic(\mu - X). \end{aligned} \tag{M1}$$

The data set is constructed in such a way that the proportions of $Y = 0|G = k$ in the two groups are different: $\mathbb{P}(Y = 0|G = 0) = 0.5$, $\mathbb{P}(Y = 0|G = 1) = 0.36$. The predictor X serves as a good proxy for G, which can be seen in Fig. 1a. The plot depicts the density of X and the conditional densities of X given $G = 0$ and $G = 1$, respectively. The shifted conditional densities clearly show the dependence of X on G. Though the true score $f(X)$ does not depend explicitly on G, a classifier trained on X will learn that the higher the value of X the more likely it is that $Y = 0$. Using the logistic regression model $\hat{f}$ we observe that for any threshold $t \in (0, 1)$ the classifier $\widehat{Y}_t$ satisfies neither the statistical parity, nor the equal opportunity, nor the equalized odds criterion. Furthermore, since both classes of G are equally likely, $BER(\widehat{Y}_t, G) < 0.5$ implies that one can potentially infer G from X; see Fig. 1b. The vertical axis in the plot represents the difference between the probabilities for each of the first three fairness metrics described in Definition 1 as well as the value of the balanced error rate. Notice how only in the trivial cases where $t \in \{0, 1\}$ are all metrics satisfied and the balanced error rate is equal to 0.5, since $\widehat{Y}_0 = 1, \widehat{Y}_1 = 0$ for all X.

2.4 Classifier bias based on statistical parity

In this section we provide a definition for classifier bias based on the statistical parity fairness criterion and establish some basic properties of the classifier bias. In what follows, we suppress the symbol ^ , using it only when it is necessary to differentiate between the true model and the trained one. The same rule applies to classifiers.

Definition 2 Let f be a model, $X \in \mathbb{R}^n$ predictors, $G \in \{0,1\}$ protected attribute, $G=0$ non-protected class, ς_f the sign of the favorable direction, and F_k the CDF of $f(X)|G=k$.

- The signed classifier (or statistical parity) bias for a threshold $t \in \mathbb{R}$ is defined by

$$\begin{aligned}\widetilde{bias}^C_t(f|X,G) &= \big(\mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|G=0) - \mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|G=1)\big)\\ &= \big(F_1(t) - F_0(t)\big)\cdot \varsigma_f.\end{aligned}$$

- The classifier bias at $t \in \mathbb{R}$ is defined by

$$bias^C_t(f|X,G) = |\widetilde{bias}^C_t(f|X,G)|.$$

We say that Y_t favors the non-protected class $G=0$ if the signed bias is positive. Respectively, Y_t favors the protected class $G=1$ if the signed bias is negative.

Remark 1 Suppose that $Y \in \{0,1\}$ is binary and that the favorable direction is ↑, which implies that $\mathbb{1}_{\{\varsigma_f=1\}} = 1$. Then Y_t favors the non-protected class $G=0$ if and only if there is a larger proportion of individuals from class $G=0$ for which $Y_t=1$ compared to the class $G=1$. This, from a statistical parity perspective, describes the outcome $Y=1$ as favorable. Similar remarks apply to the case when the favorable direction is ↓. Thus, the favorable direction is ↑ (↓) is equivalent to the favorable outcome $Y=1$ ($Y=0$).

2.5 Quantile bias and geometric parity

Given a model f and a threshold $t \in \mathbb{R}$, the classifier bias based on statistical parity measures the difference in population sizes corresponding to groups $G=\{0,1\}$ for which $Y_t=0$. This measurement however does not take into account the geometry of the model distribution, that is, the score values themselves.

For example, when measuring the bias in incomes among ‘females’ and ‘males’ one can view the difference of expected incomes in the two groups as ‘bias’. Alternatively, one can measure an income bias by evaluating the absolute difference of the ‘female’ median income and ‘male’ median income, which is often done in various social studies. This motivates us to take into account the geometry of the score distribution when defining bias. For this reason, we propose the notion of the quantile bias which operates on the domain of the score rather than the sample space.

Definition 3 Let f, X, G, ς_f and F_k be as in Definition 2. Let $p \in (0,1)$.

- The signed p-th quantile is defined by

$$\widetilde{bias}_p^Q(f|X,G) = \left(F_0^{[-1]}(p) - F_1^{[-1]}(p)\right) \cdot \varsigma_f$$

- The p-th quantile bias is defined by

$$bias_p^Q(f|X,G) = |\widetilde{bias}_p^Q(f|X,G)|.$$

As a counterpart to statistical parity, we also introduce quantile (geometric) parity.

Definition 4 (*geometric parity*) Let f be a model and $G \in \{0,1\}$ the protected attribute.

- We say that the model f satisfies p-th quantile (or geometric) parity if

$$bias_p^Q(f|X,G) = 0.$$

- Let $t \in \mathbb{R}$. The classifier Y_t satisfies quantile (or geometric) parity if

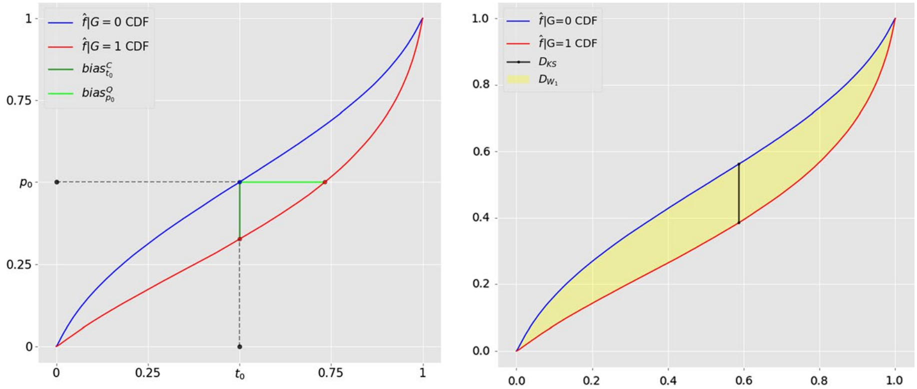
$$bias_{p_0}^Q(f|X,G) = 0, \quad p_0 = F_0(t).$$

Given a score f, the quantile bias measures the difference between subpopulation quantile values. For a given threshold t, the p_0-quantile signed bias, with $p_0 = F_0(t)$, measures by how much the corresponding score values of the protected class $G = 1$ differ from that of $G = 0$ or equivalently by how much the threshold for the protected group should be shifted to achieve the quantile parity (and in some cases statistical parity) between the two populations.

Lemma 1 *Let f be a model, $G \in \{0,1\}$ the protected attribute, and $G = 0$ the non-protected class. Suppose that $t_0 \in \mathbb{R}$ is a point at which the CDFs F_0 and F_1 are continuous and strictly increasing. Then Y_{t_0} satisfies statistical parity if and only if it satisfies geometric parity.*

Proof The result follows from Definitions 2 and 3, and the fact that F_0 and F_1 are locally invertible at t_0. □

To better understand the classifier and quantile biases and their connection, see Fig. 2a. The conditional CDFs of the model scores are plotted given the protected attribute G. The blue line (corresponding to the scores given $G = 0$) is above the red line (scores given $G = 1$) for all values of t. Thus, for a given threshold t_0 we have that $F_0(t_0) - F_1(t_0) > 0$, which means that if the favorable direction is ↑ (↓) then the classifier favors the class $G = 1$ ($G = 0$). In view of the quantile bias, the green horizontal line segment represents the amount we would have to shift the threshold for one of the classes in order to achieve geometric parity. Since the CDFs are shown to be continuous and strictly increasing, the above lemma implies that doing so would achieve statistical parity as well.



(a) Classifier and quantile biases.

(b) Model bias based on W_1 and KS.

Fig. 2 Classifier and quantile bias, and model bias for the model (M1)

2.6 Optimal transport use in ML classifier fairness

2.6.1 Classifier bias mitigation via repaired datasets

Two notable works that utilize optimal transport theory to reduce statistical parity bias are Feldman et al. (2015) and Gordaliza et al. (2019).

The approach in Feldman et al. (2015) seeks to create an unbiased dataset by transforming predictors and then training a classifier on it. The authors propose a geometric repair scheme, which partially moves the two subpopulation distributions $\mu_{i,0}$ and $\mu_{i,1}$ of predictor X_i along the Wasserstein geodesic towards their (unidimensional) Wasserstein barycenter $\mu_{i,B}$, a distribution minimizing the variance of the collection $\{\mu_{i,0}, \mu_{i,1}\}$; see Appendix A. The transformed dataset is then used to train a model that reduces disparate impact.

Gordaliza et al. (2019) proposes a method for transforming the multivariate distribution of predictors called random repair. Given two subpopulation distributions of predictors $\mu_k = P_{X|G=k}$, with $k \in \{0, 1\}$, and a repair parameter $\lambda \in [0, 1]$, the algorithm randomly chooses between the Wasserstein barycenter μ_B of $\{\mu_0, \mu_1\}$ and the original subpopulation distribution μ_k, with λ determining the probability of selecting μ_B.

The authors establish the upper bound on the disparate impact (DI) and balanced error rate (BER) of classifiers with respect to (X, G) using the total variance distance between the subpopulation distributions of predictors,

$$\min_h BER(h, X, G) = \tfrac{1}{2}(1 - d_{TV}(\mu_0, \mu_1)),$$

and show that the TV-distance between repaired subpopulation distributions $\tilde{\mu}_{0,\lambda}$,$\tilde{\mu}_{1,\lambda}$ is bounded by $1 - \lambda$. This, in turn, allows to control the bound on the DI and BER, and hence the closely related statistical parity bias on the repaired dataset is bounded by

$$\max_h bias^C(h|G, \tilde{X}_\lambda) = d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq 1 - \lambda.$$

The random repair algorithm allows for a tight control of TV-distance between repaired subpopulations unlike the geometric repair approach. They also establish bounds on the loss in performance due to modifying the data by the Wasserstein distance between the two subpopulation distributions of predictors. The performance loss is expressed as the difference in classification risk between the repaired and original data on (X, G).

Remark 2 Given the regulatory constraints, the approaches of Feldman et al. (2015) and Gordaliza et al. (2019) would not be permitted in financial institutions that extend credit because a) the protected attribute cannot be used in training or prediction, and b) introducing randomness into the input dataset is prohibited; for details see (Hall et al. 2021). To take into account the regulatory constraints and practical applications, in our companion paper (Miroshnikov et al. 2021b) we propose a post-processing approach that relies on the fairness interpretability framework presented in the current article.

2.6.2 Individual fairness

The work of Dwork et al. (2012) studies individual fairness of randomized classifiers. To understand the main results of the article, we first provide relevant definitions.

Definition 5 Let $(\mathscr{X}, d)$ be a metric space and D a distance on $\mathscr{P}(\{0,1\})$.

(i) A map $M : \mathscr{X} \to \mathscr{P}(\{0,1\})$ is called a randomized classifier.
(ii) $Lip_1(\mathscr{X}, \mathscr{P}(\{0,1\}); d, D) = \{M : \mathscr{X} \to \mathscr{P}(\{0,1\}),\ D(M(x), M(y)) \le d(x,y)\}$.
(iii) Given $\nu \in \mathscr{P}(\mathscr{X})$ the averaged M_ν is defined by $M_\nu(a) = \mathbb{E}_{x\sim\nu}[M(x)(a)]$, $a \subset \{0,1\}$.
(iv) The distance D_{rc} between $\mu, \nu \in \mathscr{P}(\mathscr{X})$ is defined by

$$D_{rc}(\mu, \nu; D, d) := \sup\left\{ M_\mu(\{0\}) - M_\nu(\{0\}),\ M \in Lip_1(\mathscr{X}, \mathscr{P}(\{0,1\}); d, D) \right\} \in [0,1].$$

Individual fairness is defined by imposing a Lipschitz property on the map $x \to M(x) \in \mathscr{P}(\{0,1\})$, $x \in \mathscr{X}$. As in Gordaliza et al. (2019), the work of Dwork et al. (2012) relates the bias in the output to the bias in the input. In particular, the paper establishes the upper bound $D_{TV}(M_{P_0}, M_{P_1}) \le D_{rc}(P_0, P_1)$ for the statistical parity bias of Lipschitz randomized classifiers. Roughly speaking, the above bound means that when two subpopulations P_0, P_1 are "similar" in the sense of the D_{rc} metric, then the Lipschitz condition ensures that the statistical parity bias is small.

The D_{rc} metric has transport-like properties and is related to the Wasserstein metric; see (Dwork et al. 2012, Theorem 3.3) and Theorem 3 in Sect. 3.5.

3 Model bias metric

In our work we shift the focus from measuring the bias in classifiers to the bias in regressor outputs. This is motivated by the fact that many strategies and decisions in the real-world make use of the regressor values or the classification scores of the trained ML models. Furthermore, in the case of classification scores, the bias assessment in FIs is carried out before any classifier threshold is determined.

In this section, we discuss how to measure the regressor bias using optimal transport. We also establish the connection between the regressor bias and the bias in the collection of classifiers induced by thresholding the regressor, and make use of this integral relationship to design generic regressor fairness metrics that incorporate group-based parity criteria, such as equalized odds (Hardt et al. 2015), into the transport formulation.

Definition 6 (***D*-model bias**) Let $X \in \mathbb{R}^n$ be predictors, f be a model, and $G \in \{0, 1\}$ the protected attribute. Let $D(\cdot, \cdot)$ be a metric on the space of probability measures $\mathscr{P}_q(\mathbb{R})$, with $q \geq 0$. Provided $\mathbb{E}[|f(X)|^q]$ is finite, the D-based model bias is defined as the distance between the subpopulation distributions of the model:

$$\text{Bias}_D(f|X, G) := D(P_{f(X)|G=0}, P_{f(X)|G=1}), \tag{1}$$

where $P_{f(X)|G=k}$ is the pushforward probability measure of $f(X)|G = k$. We say that the model (X, f) is *fair* up to the D-based bias $\epsilon \geq 0$ if $\text{Bias}_D(f|X, G) \leq \epsilon$.

Figure 2b illustrates the model bias for two choices of D: the 1-Wasserstein metric W_1 and the Kolmogorov-Smirnov distance KS. Notice the stark difference between the two model biases. This raises the general question on which metric should one use to evaluate the bias. We discuss this issue in the following section.

In what follows we suppress the explicit dependence of the model bias on X.

3.1 Wasserstein distance

To determine an appropriate metric D to be used in (1) is not a trivial task. The choice depends on the context in which the model bias is measured. We argue that it is desirable for the metric to have the following properties:

(P1) It should be continuous with respect to the change in the geometry of the distribution.
(P2) It should be non-invariant with respect to monotone transformations of the distributions.

The property (P1) makes sure that the metric keeps track of changes in the geometry. For instance, suppose an "income" of the group $\{G = 0\}$ is x_0 and that of $\{G = 1\}$ is x_1. A metric that measures income inequality should be able to sense the distance between x_0 and $x_0 + \varepsilon$. That is, having two delta measures δ_{x_0} and $\delta_{x_0+\varepsilon}$ the metric must ensure that as $\varepsilon \to 0$ the distance $D(\delta_{x_0}, \delta_{x_0+\varepsilon})$ approaches zero. The property (P1) also makes sure that slight changes in the subpopulation distributions lead to a slight change in bias measurements, which is important for stability with respect to changes in the dataset X.

The property (P2) makes sure that the metric is non-invariant with respect to monotone transformations. That is, given two random variables X_0 and X_1 and a continuous, strictly increasing transformation $T : \mathbb{R} \to \mathbb{R}$, one would expect the change in distance between $T(X_0)$ and $T(X_1)$ whenever T is not a shift. For example, if $T(x) = \alpha x$, we would expect the distance between $T(X_0) = \alpha X_0$ and $T(X_1) = \alpha X_1$ depend continuously on α.

In what follows, we consider the Wasserstein distance W_q as a potential candidate for fairness interpretability; for use cases in the ML fairness community see Dwork et al. (2012); Feldman et al. (2015); Gordaliza et al. (2019).

To introduce the metric and investigate its properties we switch our focus to probability measures; recall that any random variable Z gives rise to the pushforward probability measure $P_Z(A) = \mathbb{P}(Z \in A)$ on $\mathbb{R}$, and the reverse is true, for any $\mu \in \mathscr{P}(\mathbb{R})$ with the CDF

$F_\mu(a) = \mu((-\infty, a])$ there is a random variable Z such that $P_Z = \mu$. Similar remarks apply for random vectors; see Shiryaev (1980). Given $T : \mathbb{R}^k \to \mathbb{R}^m$ and $\mu \in \mathscr{P}(\mathbb{R}^k)$, we denote by $T_\# \mu$ a measure such that $T_\# \mu(B) = \mu\big(T^{-1}(B)\big)$.

The Wasserstein distance W_q is connected to the concept of optimal mass transport. Given two probability measures $\mu_1, \mu_2 \in \mathscr{P}_q(\mathbb{R})$ with finite q-th moment and the cost function $c(x_1, x_2) = |x_1 - x_2|^q$, the Wasserstein distance W_q is defined by

$$W_q(\mu_1, \mu_2) := \mathscr{T}^{1/q}_{|x_1 - x_2|^q}(\mu_1, \mu_2)$$

where

$$\mathscr{T}_{|x_1 - x_2|^q}(\mu_1, \mu_2) = \inf_{\gamma \in \mathscr{P}(\mathbb{R}^2)} \left\{ \int_{\mathbb{R}^2} |x_1 - x_2|^q \, d\gamma(x_1, x_2), \quad \text{with marginals } \mu_1, \mu_2 \right\}$$

is the minimal cost of transporting the distribution μ_1 into μ_2, and vice versa in view of the symmetry of the cost function. A joint probability measure $\gamma \in \mathscr{P}(\mathbb{R}^2)$ with marginals μ_1 and μ_2 is called a *transport plan*. It specifies how each point x_1 from supp(μ_1) gets distributed in the course of the transportation; specifically, the transport of x_1 is described by the conditional probability measure $\gamma_{x_2|x_1}$.

It can be shown that the Wasserstein metric for probability measures on $\mathbb{R}$ can be expressed in terms of the quantile functions

$$W_q(\mu_1, \mu_2) = \left(\int_0^1 |F^{[-1]}_{\mu_1}(p) - F^{[-1]}_{\mu_2}(p)|^q \, dp \right)^{1/q}, \tag{2}$$

which makes the computation straightforward; see Theorem 7.

To get an understanding of the behavior of W_q, consider two delta measures located at x_0 and $x_0 + \varepsilon$, respectively. By definition of the metric it follows that

$$W_q(\delta_{x_0}, \delta_{x_0+\varepsilon}) = \varepsilon.$$

Thus, W_q is continuous with respect to a shift of a point mass. Furthermore, for any two random variables X_0 and X_1 and $\alpha > 0$

$$W_q(P_{\alpha X_0}, P_{\alpha X_1}) = \alpha W_q(P_{X_0}, P_{X_1})$$

which implies that a multiplicative map $T(x) = \alpha x$ affects the Wasserstein distance.

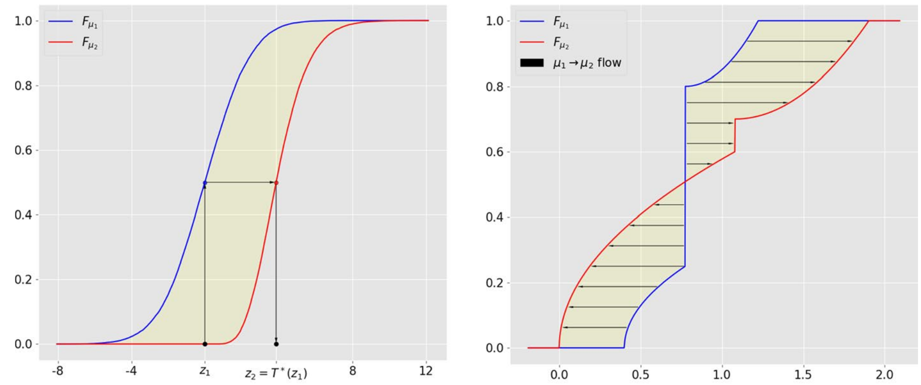
To formally show that properties (P1) and (P2) are satisfied by the Wasserstein metric, we provide the following theorem.

Theorem 1 *The distance W_q satisfies*:

(a) *W_q on $\mathscr{P}_q(\mathbb{R})$ is continuous with respect to the geometry of the distribution.*

(b) *Let $T : \mathbb{R} \to \mathbb{R}$ be a continuous, strictly increasing map. W_q is non-invariant under T, provided, $T(x) \neq x + C$ and $T_\# \mu \in \mathscr{P}_q(\mathbb{R})$, $\mu \in \mathscr{P}_q(\mathbb{R})$.*

Proof See Appendix B. □



(a) Optimal transport map T^*.

(b) Transport flow of π^*.

Fig. 3 Transporting μ_1 to μ_2 under the monotone transport plan π^*

Theorem 1 states that the Wasserstein metric relies on the geometry of the distribution. In particular, the distance is affected in a continuous way by the change in the geometry of the distribution. This, in turn, provides the desired sensitivity of the Wasserstein metric with respect to slight changes in the dataset distribution, including shifts, which is relevant for ML models with ragged CDFs, which makes the Wasserstein metric an appropriate candidate for the model bias measurement. In addition, as we will see, the Wasserstein distance enables us to assess the favorability at the level of the model, which is useful for applications in financial institutions.

3.2 Negative and positive flows under order preserving optimal transport plan

We now provide several properties of the Wasserstein metric, which we employ in the following sections.

Given two probability measures $\mu_1, \mu_2 \in \mathscr{P}_q(\mathbb{R})$, it can be shown that the joint probability measure $\pi^* \in \mathscr{P}(\mathbb{R}^2)$ with the CDF

$$F_{\pi^*}(a,b) = \min(F_{\mu_1}(a), F_{\mu_2}(b)) \tag{3}$$

is an *optimal transport plan* for transporting μ_1 into μ_2 with the cost function $c(x_1,x_2) = |x_1 - x_2|^q$, and thus,

$$W_q^q(\mu_1,\mu_2) = \mathscr{T}_{|x_1-x_2|^q}(\mu_1,\mu_2) = \int_{\mathbb{R}^2} |x_1 - x_2|^q d\pi^*(x_1,x_2). \tag{4}$$

Most importantly, π^* is the only monotone (order preserving) transport plan such that

$$(x_1,x_2),(x_1',x_2') \in \mathrm{supp}(\pi^*), \quad x_1 < x_1' \quad \Rightarrow \quad x_2 \le x_2'.$$

In a special case, when μ_1 is atomless, π^* is determined by the monotone map

$$T^* = F_{\mu_2}^{[-1]} \circ F_{\mu_1}, \tag{5}$$

called an optimal transport map. Specifically, each point x_1 of the distribution μ_1 is transported to the point $x_2 = T^*(x_1)$; see Fig. 3a for an illustration. Thus, $\mu_2 = T^*_\# \mu_1$, and the conditional probability measure $\pi^*_{x_2|x_1} = \delta_{T^*(x_1)}$ for $x_1 \in \operatorname{supp}(\mu_1)$. In this case, (4) reads

$$W_q^q(\mu_1, \mu_2) = \mathscr{T}_{|x_1-x_2|^q}(\mu_1, \mu_2) = \int_{\mathbb{R}} |x_1 - T^*(x_1)|^q d\mu_1(x_1). \tag{6}$$

The results (3)-(6) follow from Theorem 7 for the cost function $c(x_1, x_2) = |x_1 - x_2|^q$.

In a general case, under the transport plan π^*, points $x_1 \in \operatorname{supp}(\mu_1)$ for which $\mu_1(\{x_1\}) = 0$ are transported as a whole, while the "atoms", points x_1 for which $\mu_1(\{x_1\}) > 0$, are allowed to be split or spread along $\mathbb{R}$; see Fig. 3b that illustrates the transport flow under π^* in the general case. The plot also provides a depiction of the order preservation; notice how the arrows do not intersect.

To compute the portion of the transport cost used for moving points of μ_1 to the right or left, it is necessary to restrict the attention to the regions $x_1 < x_2$ and $x_1 > x_2$, respectively.

Lemma 2 *Let* $\mu_1, \mu_2 \in \mathscr{P}_q(\mathbb{R})$, $q \in [1, \infty)$. *Under the monotone plan* π^* *the transport efforts to the left and right for the cost function* $c(x_1, x_2) = |x_1 - x_2|^q$ *are given by*:

$$\begin{aligned}\mathscr{T}^{\leftrightarrows}_{|x_1-x_2|^q}(\mu_1,\mu_2) &= \int_{\{\pm(x_2-x_1)>0\}} |x_1-x_2|^q d\pi^*(x_1,x_2)\\ &= \int_{\left\{\pm(F^{[-1]}_{\mu_2}(p)-F^{[-1]}_{\mu_1}(p))>0\right\}} |F^{[-1]}_{\mu_1}(p) - F^{[-1]}_{\mu_2}(p)|^q dp.\end{aligned} \tag{7}$$

Hence, the Wasserstein distance W_q *can be expressed as*

$$W_q(\mu_1, \mu_2) = \left(\mathscr{T}^{\leftarrow}_{|x_1-x_2|^q}(\mu_1, \mu_2) + \mathscr{T}^{\rightarrow}_{|x_1-x_2|^q}(\mu_1, \mu_2)\right)^{1/q}. \tag{8}$$

Furthermore, if μ_1 *is atomless*, (7) *reads*

$$\mathscr{T}^{\leftrightarrows}_{|x_1-x_2|^q}(\mu_1,\mu_2) = \int_{\left\{\pm(T^*(x_1)-x_1)>0\right\}} |x_1 - T^*(x_1)|^q d\mu_1(x_1), \quad T^* = F^{[-1]}_{\mu_2} \circ F_{\mu_1} \tag{9}$$

Proof By (3) the monotone plan can be expressed as

$$\pi^* = (F^{-1}_{\mu_1}, F^{-1}_{\mu_2})_\# \lambda|_{[0,1]} \in \mathscr{P}(\mathbb{R}^2)$$

where $\lambda|_{[0,1]}$ denotes the Lebesgue measure restricted to [0, 1]. Then, by Proposition 6, for any Borel set $B \subset \mathbb{R}^2$ we have

$$\int_B |x_1 - x_2|^q d\pi^*(x_1, x_2) = \int_{\{p \in (0,1):\, (F^{[-1]}_{\mu_1}(p), F^{[-1]}_{\mu_2})(p)) \in B\}} |F^{[-1]}_{\mu_1}(p) - F^{[-1]}_{\mu_2}(p)|^q dp.$$

Then (7) follows from the above identity with $B = \{(x_1, x_2) : \pm(x_1 - x_2) > 0\}$. Next, by (4) and (7), we obtain (8).

Finally, if μ_1 is atomless, by Theorem 7 the monotone plan $\pi^* = (I, T^*)_\# \mu_1$, where T^* is the optimal transport map given by (5). Then using Proposition 6 we obtain (9). □

3.3 W_1-based model bias and its components

For $q=1$ the Wasserstein distance W_1 is known as the *Earth Mover distance*. Since the distance is symmetric, $\mathrm{Bias}_{W_1}(f|X,G)$ is the cost of transporting the distribution of $f(X)|G=0$ into that of $f(X)|G=1$ or vice versa.

It can be shown that the W_1-based model bias formulation is consistent with both statistical parity fairness criterion as well as quantile parity criterion, which is shown by the following theorem.

Lemma 3 *Let f be a model and G* $\in \{0,1\}$ *the protected attribute. Then*

$$\mathrm{Bias}_{W_1}(f|G) = \int_0^1 bias_p^Q(f|G)\,dp = \int_{\mathbb{R}} bias_t^C(f|G)\,dt.$$

Proof By assumption $\mathbb{E}|f(X)| < \infty$ and hence $\mathbb{E}[|f(X)|G=k|] < \infty$ for $k \in \{0,1\}$. Then, we have (Shorack and Wellner (1986))

$$\begin{aligned} W_1\big(f(X)|G=0, f(X)|G=1\big) &= \int_0^1 |F^{[-1]}_{f(X)|G=0}(p) - F^{[-1]}_{f(X)|G=1}(p)|\,dp \\ &= \int_{\mathbb{R}} |F_{f(X)|G=0}(t) - F_{f(X)|G=1}(t)|\,dt \, < \, \infty. \end{aligned}$$

Hence, the result follows from Definitions 2 and 3, and the above equality. □

Remark 3 The above lemma establishes the representation of the model bias as an integration over the statistical parity bias of classifiers obtained by considering all thresholds. Here, the consistency of the model bias with statistical parity is understood in the sense of the equality in the above lemma. In comparison, Dwork et al. (2012) establishes a connection of statistical parity of Lipschitz randomized classifiers and subpopulations in a dataset upon which the models are built.

While the results in Dwork et al. (2012) do not imply the above lemma, it is appealing to provide a connection between the two. For example, consider the triplet (X, G, Y) with $Y \in \{0,1\}$ and a smooth regressor $f(X) = P(Y=1|X)$. Consider a randomized classifier $z \to \mu_z$ where $z=(x,g,y)$, and $\mu_z(1) = f(x)$. Let $P_g = P_{Z|G=g}$. Then, the upper bound on statistical parity bias of μ_z provided by Dwork et al. (2012) reads

$$D_{TV}(\mu_{P_0}, \mu_{P_1}) = |\mathbb{E}[f(X)|G=0] - \mathbb{E}[f(X)|G=1]| \le W_1(P_0, P_1),$$

which illustrates the difference between Lemma 3.1 of Dwork et al. (2012) and our lemma.

Positive and negative model bias. According to Lemma 2, the cost of transporting a distribution is the sum of the transport effort to the left and the transport effort to the right. This motivates us to define the positive bias as the transport effort for moving the particles of $f(X)|G=0$ in the non-favorable direction and the negative bias as the transport effort in the favorable one; equivalently the latter is the transport effort for moving the particles of $f(X)|G=1$ into the favorable direction and the former is the transport effort into the non-favorable one.

Motivated by Lemma 2 we define positive and negative model biases as follows:

Definition 7 Let f, G, ς_f and F_k be as in Definition 2.

- The positive and negative W_1 based model biases are defined by

$$\text{Bias}^{\pm}_{W_1}(f|G) = \int_{\mathcal{P}_\pm} \pm(F_0^{[-1]}(p) - F_1^{[-1]}(p)) \cdot \varsigma_f \, dp$$

  where

$$\mathcal{P}_\pm = \left\{ p \in (0,1) : \ \pm\widetilde{bias}^{Q}_{p}(f|G) = \pm(F_0^{-1}(p) - F_1^{-1}(p)) \cdot \varsigma_f > 0 \right\}.$$

  In this case, the model bias is disaggregated as follows:

$$\text{Bias}_{W_1}(f|G) = \text{Bias}^{+}_{W_1}(f|G) + \text{Bias}^{-}_{W_1}(f|G).$$

- The net model bias is defined by

$$\text{Bias}^{net}_{W_1}(f|G) = \text{Bias}^{+}_{W_1}(f|G) - \text{Bias}^{-}_{W_1}(f|G).$$

We next establish that the positive and negative W_1 model biases can be expressed in terms of classifier biases. To establish this, we first prove the following auxiliary lemma.

Lemma 4 *Let X_0, X_1 be random variables with $\mathbb{E}|X_i| < \infty$, $i \in \{0,1\}$. Let F_i denote the CDF of X_i and let*

$$\begin{aligned}
\mathcal{T}_0 &= \{t \in \mathbb{R} : F_1(t) < F_0(t)\}, & \mathcal{T}_1 &= \{t \in \mathbb{R} : F_0(t) < F_1(t)\} \\
\mathcal{P}_0 &= \{p \in (0,1) : F_1^{[-1]}(p) < F_0^{[-1]}(p)\}, & \mathcal{P}_1 &= \{p \in (0,1) : F_0^{[-1]}(p) < F_1^{[-1]}(p)\}.
\end{aligned}$$

Then

$$\begin{aligned}
0 &\le \int_{\mathcal{T}_0} F_0(t) - F_1(t)\, dt = \int_{\mathcal{P}_1} F_1^{[-1]}(p) - F_0^{[-1]}(p)\, dp < \infty \\
0 &\le \int_{\mathcal{T}_1} F_1(t) - F_0(t)\, dt = \int_{\mathcal{P}_0} F_0^{[-1]}(p) - F_1^{[-1]}(p)\, dp < \infty.
\end{aligned}$$

Proof See Appendix B. □

Theorem 2 *Let $f, G, \varsigma_f, \mathcal{P}^{\pm}$ and F_k be as in Definition 7. Then*

$$\text{Bias}^{\pm}_{W_1}(f|G) = \int_{\mathcal{P}_\pm} bias^{Q}_{p}(f|G)\, dp = \int_{\mathcal{T}_\pm} bias^{C}_{t}(f|G)\, dt \tag{10}$$

where

$$\mathcal{T}_\pm = \left\{ t \in \mathbb{R} : \ \pm\widetilde{bias}^{C}_{t}(f|G) = \pm(F_1(t) - F_0(t)) \cdot \varsigma_f > 0 \right\}.$$

The net bias satisfies

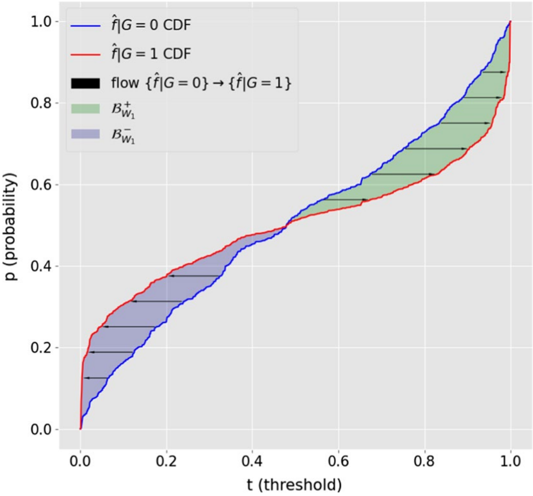


Fig. 4 Positive and negative model biases for the trained XGBoost model (M2), $\varsigma_f = -1$

$$\begin{aligned}\mathrm{Bias}^{net}_{W_1}(f|G) &= \int_0^1 \widetilde{bias}^{Q}_{p}(f|G)dp = \int_{\mathbb{R}} \widetilde{bias}^{C}_{t}(f|G)\,dt \\ &= \big(\mathbb{E}[f(X)|G=0] - \mathbb{E}[f(X)|G=1]\big)\cdot\varsigma_f\end{aligned} \tag{11}$$

Proof Suppose first that favorable direction is ↑. Since $\mathbb{E}|f(X)| < \infty$, we have $\mathbb{E}[|f(X)||G=k] < \infty$ for $k \in \{0,1\}$. Then by Lemma 4

$$\mathrm{Bias}^{\pm}\big(f|G\big) = \pm\int_{\mathcal{P}^{\pm}} F^{[-1]}_{f|G=0}(p) - F^{[-1]}_{f|G=1}(p)\,dp = \pm\int_{\mathcal{T}^{\pm}} F_{f|G=1}(t) - F_{f|G=0}(t)\,dt < \infty.$$

Hence (10) follows from Definitions 2 and 3, and the above equality.

Next, by (10) and Lemma 17 we have

$$\begin{aligned}\mathrm{Bias}^{net}(f|G) &= \mathrm{Bias}^{+}(f|G) - \mathrm{Bias}^{-}(f|G)\\ &= \int_{\mathcal{T}^{+}} \big(F_{f|G=1}(t) - F_{f(X)|G=0}(t)\big)dt - \int_{\mathcal{T}^{-}} \big(F_{f|G=0}(t) - F_{f|G=1}(t)\big)dt\\ &= \int_{-\infty}^{0} \big(F_{f|G=1}(t) - F_{f|G=0}(t)\big)dt + \int_{0}^{\infty} \big((1-F_{f|G=0}(t)) - (1-F_{f|G=1}(t))\big)dt\\ &= \mathbb{E}[f(X)|G=0] - \mathbb{E}[f(X)|G=1].\end{aligned}$$

This proves (11). If the favorable direction is ↓, the proof of (10) and (11) is similar. □

In the context of classification, Theorem 2 states that the positive W_1-based model bias is the integrated classifier bias over the set of thresholds $t \in \mathcal{T}_+$ where the classifiers $Y_t = \mathbb{1}_{\{f(X)>t\}}$ favor the non-protected class $G=0$. Similar remark holds for the negative model.

Furthermore, the property (10) of $\mathrm{Bias}^{\pm}_{W_1}$ allow one to use thresholds and quantiles interchangeably, which is beneficial in classification problems. For this reason, we choose W_1 as our primary metric.

Example To understand the statement of Theorem 2 consider the following classification risk model ($\varsigma_f = -1$) with a predictor whose variance depends on the attribute G:

$$\begin{aligned} &X \sim N(\mu, (1+G)\sqrt{\mu}), \quad \mu = 5 \\ &Y \sim Bernoulli(f(X)), \quad f(X) = \mathbb{P}(Y=1|X) = \sigma(\mu - X). \end{aligned} \tag{M2}$$

which leads to the presence of both positive and negative bias components in the score distribution. Figure 4 depicts the subpopulation score CDFs of the trained GBM classifier and illustrates the fact that the integrated positive quantile and classifier biases yield the positive model bias (green region), and a similar relationship holds for the negative model bias (purple region). The monotone transport flows are also depicted, showing the connection between the signed model bias and the favorability. Since $\varsigma_f = -1$, in the green region the non-protected class is transported towards the non-favorable direction, while in the purple region it is transported towards the favorable one.

On renormalization of model bias If $f(X)$ is a classification score then $\mathrm{Bias}_{W_1}(f|G) \in [0,1]$, which makes it easy to interpret the amount of the bias in the model distribution.

For regressors, however, the model bias can take any value in $[0,\infty)$. One approach is to normalize the model bias as follows. First, pick an appropriate reference scale $L > 0$ corresponding to the response variable. Given the scale L one can define a generalized Wasserstein-based model bias as follows:

$$Bias_{g,W_1}(f|G) = g\Big(\frac{1}{L}\mathrm{Bias}_{W_1}(f|G)\Big) \tag{12}$$

where the link function g is strictly increasing and satisfies

$$g(x) = \begin{cases} x, & x \in [0, 0.5] \\ g\text{increases to}1. & \end{cases}$$

Having this setup yields $Bias_{g,W_1}(f|G) = \frac{1}{L}Bias_{W_1}(f|G)$ whenever the transport effort is within the scale of interest L, that is, when $Bias_{W_1}(f|G) \leq \frac{L}{2}$. In practice, for bounded distributions, one can pick $L = \mathrm{supp}\, P_{f(X)}$, while for unbounded distributions one can pick $L = 2\sigma(f(X))$.

In our work, we develop the bias explanation methods to explain the actual amount of transport effort between subpopulations. The generalization to (12) is trivial.

3.4 Generalized group-based parity model bias

In this section, we will generalize the notions of the Wasserstein-based bias to the case of generic group-based parity for protected attributes with multiple classes. We then apply the generalization to the equalized odds and the equal opportunity parity conditions.

Definition 8 Let f be a model, $X \in \mathbb{R}^n$ predictors, $G \in \{0, 1, \dots, K-1\}$ protected attribute, $G = 0$ non-protected class, and ς_f the sign of the favorable direction of f. Let $\mathcal{A} = \{A_1, \dots, A_M\}$ be a collection of disjoint subsets of the sample space Ω. Define events

$$A_{km} = \{G = k\} \cap A_m, \quad k \in \{0, 1, \dots, K-1\}, \quad m \in \{1, \dots, M\}.$$

(i) We say that $Y_t = \mathbb{1}_{\{f(X)>t\}}$ satisfies $\mathcal{A}$ group-based parity if

$$\mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{km}) = \mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{0m}), \quad k \in \{1,\dots,K-1\}, \quad m \in \{1,\dots,M\}. \tag{13}$$

(ii) $(W_1,\mathcal{A})$-based (weighted) model bias is defined by

$$\mathrm{Bias}^{(w)}_{W_1,\mathcal{A}}(f|G) = \sum_{k=1}^{K-1}\sum_{m=1}^{M} w_{km}\mathrm{Bias}_{W_1}(f|\{A_{0m},A_{km}\}), \quad w_{km} \ge 0,$$

where the weights satisfy $\sum_{k=1}^{K-1}\sum_{m=1}^{M} w_{km} = 1$.

(iii) The positive and negative $(W_1,\mathcal{A})$ weighted model biases are defined by

$$\mathrm{Bias}^{(w)\pm}_{W_1,\mathcal{A}}(f|G) = \sum_{k,m} w_{km}\mathrm{Bias}^{\pm}_{W_1}(f|\{A_{0m},A_{km}\}).$$

Lemma 5 *Let G and $\mathcal{A}$ be as in Definition 8. The $(W_1,\mathcal{A})$ model bias is consistent with the generic parity criterion (13) as given by the following:*

$$\begin{aligned}\mathrm{Bias}_{W_1,\mathcal{A}}(f|G) &= \sum_{k,m} w_{km}\int_0^1 |F^{[-1]}_{f|A_{0m}} - F^{[-1]}_{f|A_{km}}|\,dt\\ &= \sum_{k,m} w_{km}\int_{\mathbb{R}} |\mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{km}) - \mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{0m})|\,dt.\end{aligned}$$

Similarly, the signed model biases can be expressed

$$\begin{aligned}\mathrm{Bias}^{(w)\pm}_{W_1,\mathcal{A}}(f|G) &:= \sum_{k,m} w_{km}\int_{\mathcal{P}_{km\pm}} \pm\big(F^{[-1]}_{f|A_{0m}}(p) - F^{[-1]}_{f|A_{km}}(p)\big)\cdot\varsigma_f\,\,dp\\ &= \sum_{k,m} w_{km}\int_{\mathcal{T}_{km\pm}} |\mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{km}) - \mathbb{P}(Y_t = \mathbb{1}_{\{\varsigma_f=1\}}|A_{0m})|\,dt,\end{aligned}$$

where

$$\begin{aligned}\mathcal{P}_{km\pm} &= \Big\{p \in [0,1] \,:\, \pm\big(F^{[-1]}_{f|A_{0m}}(p) - F^{[-1]}_{f|A_{km}}(p)\big)\cdot\varsigma_f > 0\Big\}\\ \mathcal{T}_{km\pm} &= \Big\{t \in \mathbb{R} \,:\, \pm\big(F_{f|A_{km}}(t) - F_{f|A_{0m}}(t)\big)\cdot\varsigma_f > 0\Big\}.\end{aligned}$$

Proof The claim follows directly from Theorem 4. □

Example Suppose that the favorable direction is ↑. Suppose that $G \in \{0,1\}$ and that the response variable $Y \in \{0,1\}$. Let $\mathcal{A} = \{\{Y=0\},\{Y=1\}\}$. In that case, the group-based parity condition (13) reads

$$\mathbb{P}(Y_t = 1|G=0, Y=m) = \mathbb{P}(Y_t = 1|G=1, Y=m), \quad m = 0,1,$$

which is the equalized odds criterion; Hardt et al. (2015). Then apply the above Lemma.

3.5 Integral probability metrics for fairness assessment

When assessing fairness of model regressors, it is crucial to pick an appropriate metric because the model output is often used to make decisions. A wide class of candidate metrics could be integral probability metrics (IPMs). These provide a notion of "distance" between probability distributions and are designed as generalizations of the Kantorovich-Rubinstein variational formula. They can be defined directly using variational formulas (Müller 1997; Sriperumbudur et al. 2009). Specifically, IPMs can be defined by maximizing the difference of expected values over a function space $\mathcal{A}$,

$$W_{\mathcal{A}}(v_0, v_1) := \sup_{\varphi \in \mathcal{A}} \left\{ \int \varphi(x)\, v_0(dx) - \int \varphi(x)\, v_1(dx) \right\}, \tag{14}$$

where $v_0, v_1 \in \mathscr{P}(\mathscr{X})$ and $(\mathscr{X}, d)$ is a metric space. For example, the Wasserstein metric can be obtained by taking $\mathcal{A} = \{\varphi : [\varphi]_{Lip} \le 1\}$ in (14), where $[\varphi]_{Lip}$ is the Lipschitz constant of the function φ; The Dudley metric is obtained by taking $\mathcal{A} = \{\varphi : [\varphi]_{Lip} + \|\varphi\|_\infty \le 1\}$. Dropping the regularity of test functions leads to a discontinuous response to shifting of delta masses. For example, by setting $\mathcal{A} = \{\varphi : \|\varphi\|_\infty \le 1\}$, one obtains the total variation metric D_{TV}. An interesting aspect of the above variational formula is that it can be generalized to include a broader family of distances between probability distributions, namely divergences such as the Kullback-Leibler divergence; see Birrell et al. (2020) for more information.

Thus, IPMs with regular test functions serve as good candidates for assessing the fairness of the regressor via formula (1). One of the interesting contenders is $W_{\mathcal{A}^*}$ where $\mathcal{A}^* := \{\varphi : \|\varphi\|_\infty \le \frac{1}{2}, [\varphi]_{Lip} \le 1\}$, which is an equivalent metric to the Dudley metric and has the appealing property that its values are in the unit interval. $W_{\mathcal{A}^*}$ provides meaning in fairness assessment, as it could be expressed via a supremum over all "agents" in the form of regular randomized classifiers that detect the differences between two probability subpopulations. Specifically, it can be shown that $W_{\mathcal{A}^*}$ coincides with the D_{rc} metric introduced in Dwork et al. (2012) and discussed in Sect. 2.6.

Lemma 6 *Let $(\mathscr{X}, d)$ be a metric space. Then $D_{rc}(\mu, \nu; D_{TV}, d) = W_{\mathcal{A}^*}(\mu, \nu)$.*

Proof See Appendix B. □

Recall that Dwork et al. (2012) established that the statistical parity bias of a randomized classifier is bounded by the D_{rc} distance between subpopulation input distributions. In contrast, we focus on measuring and explaining the bias in the output of non-randomized regressors, including classification scores, for which the notion of statistical parity is not, in general, applicable. In particular, we assess the distance between regressor output subpopulations via the W_1 metric. In general, any transport metric can be considered for this task, such as $W_{\mathcal{A}^*}$. Furthermore, we propose a framework that quantifies the contribution of predictors to that distance, which serves as a mechanism that pinpoints the main drivers to the regressor bias.

The lemma below illustrates the different behavior of the two metrics under scaling.

Lemma 7 *Let $d(x, y)$ be a norm on $\mathbb{R}^n$. Let $T(x) = cx + x_0$ with $c > 0$. Then*

$$D_{rc}(T_\# \mu, T_\# \nu; D_{TV}, d) = D_{rc}(\mu, \nu; D_{TV}, d_c), \quad \frac{1}{c} W_1(T_\# \mu, T_\# \nu; d) = W_1(\mu, \nu; d)$$

where $\mu, \nu \in \mathscr{P}_1(\mathbb{R}^n; d)$ *and* $d_c(x, y) = cd(x, y)$.

Proof See Appendix B. □

Notice that for large c the values of D_{rc} with the d_c norm saturate and approximate one, which is an upper bound for the metric. However, W_1 is unbounded and the distance between the pushforward measures $T_\# \mu, T_\# \nu$ scales linearly by c, which is an appealing property.

Dwork et al. (2012) establishes the connection between D_{rc} and W_1 under the assumption that the subpopulation distributions are discrete and $d \le 1$. In what follows, we prove a more general version of (Dwork et al. 2012, Theorem 3.3) that connects the two metrics and holds for all probability measures with bounded support.

Theorem 3 *Let* $\mu, \nu \in \mathscr{P}_1(\mathbb{R}^n; d)$ *have bounded supports and* $d(x, y)$ *be a norm. Then*

$$\frac{1}{L} W_1(\mu, \nu\, ; d) = D_{rc}(\mu, \nu; D_{TV}, d_{(1/L)}) \tag{15}$$

for any $L > 0$ *such that* $\operatorname{supp}(\mu), \operatorname{supp}(\nu) \subset B(x_*, \frac{L}{2}; d) = \{x : d(x, x_*) \le \frac{L}{2}\}$.

Proof See Appendix B. □

When using D_{rc} for fairness, the above theorem implies that saturation can be partially avoided via scaling. For example, the rescaling factor can be chosen as the second moment of the two probability measures. However, in our paper we focus on the Wasserstein metric because of its appealing scaling property.

4 Bias explanations

4.1 Relationship between model fairness and predictors

It is shown in Gordaliza et al. (2019) that the statistical parity bias of (non-randomized) classifiers can be bounded by the total variance distance between predictors subpopulations, while the Wasserstein metric, in general, does not allow for such control (in the sense of a bound). In contrast to the bound in Gordaliza et al. (2019), W_1-bias in predictors can control the statistical parity bias of Lipschitz randomized classifiers as shown in Dwork et al. (2012), as well as the W_1-regressor bias as shown by the following lemma.

Lemma 8 *Let* X, G, f *be as in Definition 8. If* f *is Lipschitz continuous then*

$$\mathrm{Bias}_{W_1}(f|X, G) \le [f]_{Lip} \mathrm{Bias}_{W_1}(X|G). \tag{16}$$

Proof The proof follows directly from the Kantorovich-Rubinstein variational formula. □

While the fairness of predictors as a bound is of theoretical importance, it provides little information on the contribution of each predictor to the model unfairness. This is because fairness of predictors is a sufficient requirement for fairness of the model, but not a necessary one. In particular, a model can be slightly unfair while having wildly biased predictors. For example, consider the data generating model

$$X_1 \sim N(\tau G, \sigma), \quad X_2 \sim N(0, \sigma), \quad Y = f(X) = \frac{\varepsilon}{\tau} X_1 + X_2. \tag{17}$$

Note that $\mathrm{Bias}_{W_1}(X|G) \to \infty$ as $\tau \to \infty$, while $\mathrm{Bias}_{W_1}(f|X, G) = \epsilon$ for any $\tau > 0$.

This pedagogical example motivates us to directly assess the contribution of predictors to the model bias. To accomplish this, we design an interpretability framework that employs optimal transport theory in order to pinpoint the main drivers of the model bias. Information from these drivers can then be used for policy decision-making, regulatory-compliant bias mitigation (Miroshnikov et al. 2021b), as well as in other settings.

4.2 Model interpretability

The bias explanations we develop in the next section make use of model explainers, whose objective is to quantify the contribution of each predictor to the value of $f(x)$. Several methods of interpreting ML model outputs have been designed and used over the years. Some notable ones are Partial Dependence Plots (PDP) (Friedman 2001) and SHAP values (Lundberg and Lee 2017).

Partial dependence plots PDP marginalizes out the variables whose impacts to the output are not of interest, quantifying an overall impact of the values of the remaining features.

Let $X \in \mathbb{R}^n$ be predictors, X_S with $S \subseteq \{1, 2, \dots, n\}$ a subvector of X, and $-S$ the complement set. Given a model f, the partial dependence plot of f on X_S is defined by

$$PDP_S(x; f) = \mathbb{E}[f(x_S, X_{-S})] \approx \frac{1}{N} \sum_{j=1}^{N} f(x_S, X_{-S}^{(j)}), \tag{18}$$

where we abuse the notation and ignore the variable ordering in f.

Shapley additive explanations In its original form the Shapley values appear in the context of cooperative games; see Shapley (1953); Young (1985). A cooperative game with n players is a super-additive set function v that acts on $N = \{1, 2, \dots, n\}$ and satisfies $v(\emptyset) = 0$. Shapley was interested in determining the contribution by each player to the game value $v(N)$. It turns out that under certain symmetry assumptions the contributions are unique and they are called Shapley values; furthermore, the super-additivity assumption can in principle be dropped (uniqueness and existence still hold).

It is shown in Shapley (1953) that there exists a unique collection of values $\{\varphi_i\}_{i=1}^n$ satisfying the axioms of symmetry, efficiency, and law of aggregation, ((A1)-(A3) in Shapley (1953)), it is given by

$$\varphi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} [v(S \cup \{i\}) - v(S)], \quad s = |S|, \; n = |N|. \tag{19}$$

The values provide a disaggregation of the value $v(N)$ of the game into n parts that represent a contribution to the worth by each player: $\sum_{i=1}^{n} \varphi_i[v] = v(N)$.

The explanation techniques explored in Štrumbelj and Kononenko (2010) and Lundberg and Lee (2017) utilize cooperative game theory to compute the contribution of each predictor to the model value. In particular, given a model f, Lundberg and Lee (2017) consider the games

$$v^{CE}(S;X,f) = \mathbb{E}[f|X_S], \quad v^{ME}(S;X,f) = \mathbb{E}[f(X_S, X_{-S})]|_{x_S=X_S} \tag{20}$$

with

$$v^{CE}(\varnothing;X,f) = v^{ME}(\varnothing;X,f) = \mathbb{E}[f(X)].$$

The games defined in (20) are not cooperative since they do not satisfy the condition $v(\varnothing) = 0$. However, by setting $\varphi_0 = \mathbb{E}[f(X)]$, the values satisfy the additivity property:

$$\sum_{i=0}^{n} \varphi_i[v(\cdot\,;X,f)] = f(X), \quad v \in \{v^{CE}, v^{ME}\}.$$

Throughout the text when the context is clear we suppress the explicit dependence of $v(S;X,f)$ on X and f. Furthermore, we will refer to values $\varphi_i[v^{ME}]$ and $\varphi_i[v^{CE}]$ as SHAP values and abusing the notation we write

$$\varphi_i(X;f,v) = \varphi_i[v(S;X,f)], \quad v \in \{v^{CE}, v^{ME}\}.$$

Conditional and marginal games In our work, we refer to the games v^{CE} and v^{ME} as conditional and marginal, respectively. If predictors X are independent, the two games coincide. In the presence of dependencies, however, the games are very different. Roughly speaking, the conditional game explores the data by taking into account dependencies, while the marginal game explores the model f in the space of its inputs, ignoring the dependencies. Strictly speaking, the conditional game is determined by the probability measure P_X, while the marginal game is determined by the product probability measures $P_{X_S} \otimes P_{X_{-S}}$, $S \subset N$ as stated below.

Lemma 9 (stability) *The SHAP explanations have the following properties*:

(i) $\|\varphi(X;f,v^{CE})\|_{L^2(\mathbb{P})} \leq \|f\|_{L^2(P_X)}$.
(ii) $\|\varphi(X;f,v^{ME})\|_{L^2(\mathbb{P})} \leq C\|f\|_{L^2(\widetilde{P}_X)}$, *with* $\widetilde{P}_X = \frac{1}{2^n}\sum_{S\subset N} P_{X_S} \otimes P_{X_{-S}}$.

Proof By the properties of the conditional expectation and (19) we have

$$\|\varphi_i(X;f,v^{CE})\|_{L^2(\Omega)} \leq \sum_{S\subset N\setminus\{i\}} \frac{s!(n-s-1)!}{n!} \|\mathbb{E}[f(X)|X_S]\|_{L^2(\Omega)} \leq \|f\|_{L^2(P_X)}.$$

Since φ is linear, the map in (i) is a bounded, linear operator with the unit norm. This proves (i).

By (19) and (20) we have

$$\|\varphi_i(X;f,v^{ME})\|_{L^2(\Omega)} \leq max_{s\in\{0,\dots,n-1\}} \frac{s!(n-s-1)!}{n!} \sum_{S\subset N\setminus\{i\}} \|f\|_{L^2(P_{X_S}\otimes P_{X_{-S}})} \leq C\|f\|_{L^2(\tilde{P}_X)}.$$

where $C = C(n)$ is a constant that depends on n. This proves (ii). □

To clarify the notation, we let $L^2(\widetilde{P}_X)$ denote the space of functions defined on $\mathbb{R}^n$ such that

$$\int f^2(x)\widetilde{P}_X(dx) := \frac{1}{2^n}\sum_{S\subset N}\int f^2(x_S,x_{-S})[P_{X_S}\otimes P_{X_{-S}}](x_S,x_{-S}) < \infty,$$

where as before we ignore the variable ordering in f, and for $S=\varnothing$ we assign $P_{X_\varnothing}\otimes P_X = P_X$.

We should point out that under dependencies the marginal explanation map (*ii*) in Lemma 9 is in general not continuous in $L^2(P_X)$. Hence the algorithm that produces marginal explanations may fail to satisfy the stability bounds in the sense discussed in Kearns and Ron (1999); Bousquet and Elisseeff (2002). For a more general version of the above proposition see Miroshnikov et al. (2021a).

In general, SHAPs are computationally intensive to evaluate due to the different combinations of predictors that need to be considered; in addition, computing $\varphi[v^{CE}]$ is challenging when the predictor's dimension is large in light of the curse of dimensionality; see Hastie et al. (2016). Lundberg et al. (2019) created a fast method called TreeSHAP but it can only be applied to ML algorithms that incorporate tree-based techniques. The algorithm evaluates $\varphi[v]$ for the game v that can be chosen as either one that is based upon tree paths and resembles v^{CE}, or the marginal game v^{ME}. To understand the difference between the two games, see Janzing et al. (2019); Sundararajan and Najmi (2019); Chen et al. (2020); Miroshnikov et al. (2021a).

4.3 Bias explanations of predictors

In this section, given a model, we define the bias explanation (or contribution) of each predictor. An extension to groups of predictors maybe found in Sect. 4.6.

In what follows we will be using the following notation. Given predictors $X=(X_1,X_2,\dots,X_n)$ and a model f, a generic single feature explainer of f that quantifies the attribution of each predictor X_i to the model value $f(X)$ is denoted by

$$E(X;f) = (E_1(X;f), E_2(X;f),\dots,E_n(X;f)).$$

For example, a simple way of setting up an explainer E_i is by specifying each component via a conditional or marginal expectation $E_i(X;f) = v(\{i\};X,f)$, $v\in\{v^{CE},v^{ME}\}$.

A more advanced way of computing single feature explanations is via the Shapley value $E(X;f) = \varphi[v(\cdot;X,f)]$, $v\in\{v^{CE},v^{ME}\}$. For more details on appropriate game values and their properties see Miroshnikov et al. (2021a).

Definition 9 Let $X\in\mathbb{R}^n$ be predictors, f a model, $G\in\{0,1\}$ the protected attribute, $G=0$ the non-protected class, and ς_f the sign of the favorable direction of f. Let $E(X;f)$ be an explainer of f that satisfies $\mathbb{E}\big[|E(X;f)|\big] < \infty$.

- The bias explanation of the predictor X_i is defined by

$$\beta_i(f|X,G;E_i) = W_1(E_i(X;f)|G=0, E_i(X;f)|G=1) = \int_0^1 |F^{[-1]}_{E_i|G=0} - F^{[-1]}_{E_i|G=1}|\,dp.$$

- The positive bias and negative bias explanations of the predictor X_i are defined by

$$\beta_i^{\pm}(f|X,G;E_i) = \int_{\mathcal{P}_{i\pm}} (F^{[-1]}_{E_i|G=0} - F^{[-1]}_{E_i|G=1}) \cdot \varsigma_f \, dp$$

where

$$\mathcal{P}_{i\pm} = \{p \in [0,1] : \pm(F^{[-1]}_{E_i|G=0} - F^{[-1]}_{E_i|G=1}) \cdot \varsigma_f > 0\}.$$

In this case the X_i bias explanation is disaggregated as follows:

$$\beta_i(f|X,G;E_i) = \beta_i^+(f|X,G;E_i) + \beta_i^-(f|X,G;E_i).$$

- The X_i net bias explanation is defined by

$$\beta_i^{net}(f|X,G;E_i) = \beta_i^+(f|X,G;E_i) - \beta_i^-(f|X,G;E_i).$$

- The classifier (or statistical parity) bias of the explainer E_i for a threshold $t \in \mathbb{R}$ is defined by

$$\widetilde{bias}_t^C(E_i|G) = \big(F_{E_i|G=1}(t) - F_{E_i|G=0}(t)\big) \cdot \varsigma_f.$$

By design the contribution β_i^+ measures the positive contribution to the total model bias, not the positive one. In particular, it measures the contribution to the increase in the positive flow and the decrease to the negative one. The meaning of β_i^- is similar. To better understand their meaning, consider the following data generating model:

$$f(X) = X_1 + X_2, \quad X_1 = N(\mu + \tau G, \sigma), \quad X_2 = N(\mu - \tau G, \sigma) \tag{21}$$

where X_1, X_2 are independent. Note that $\text{Bias}_{W_1}(f|X,G) = 0$, while the bias explanations are $\beta_1^+ = \tau$, $\beta_1^- = 0$, $\beta_2^+ = 0$, $\beta_2^- = \tau$ for either model explainer discussed in this section. Note also that both positive and negative model biases are zero. The positive contribution $\beta_1^+ = \tau$ measures how much in total is added to the positive model bias and subtracted from the negative one. A similar discussion holds for β_i^-. Thus, the amount that X_1 contributes to the positive bias is offset by the amount that X_2 resists to its increase. This leads to zero positive model bias. A similar discussion applies to the negative model bias.

Lemma 10 *Let $X, f, G, E_i(X;f)$, and ς_f be as in the definition 9. Then*

$$\beta^{net}(f|X,G;E_i) = \big(\mathbb{E}[E_i(X;f)|G=0] - \mathbb{E}[E_i(X;f)|G=1]\big) \cdot \varsigma_f. \tag{22}$$

Proof Similar to the proof of Theorem 2 with the assumption $\varsigma_{E_i} = \varsigma_f$. □

Observe that the bias explanations for a classification score always lie in the unit interval.

Lemma 11 *Let f be a classification score and $G \in \{0,1\}$ the protected attribute. Let the explainer E_i be either $v(\{i\};X,f)$ or $\varphi_i[v(\cdot;X,f)]$, where $v \in \{v^{CE}, v^{ME}\}$. Then $\beta_i, \beta_i^-, \beta_i^+ \in [0,1]$.*

Proof The lemma follows from the fact that $f \in [0,1]$ and the definition of explainer values. □

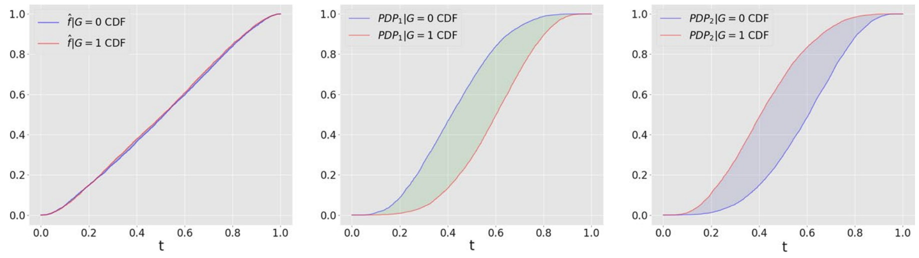


Fig. 5 Model and PDP biases for the model (M3), $\varsigma_{\hat{f}} = -1$

The explainer E_i that appears in Definition 9 is a generic one. In the examples that follow we chose to work with explainers based on marginal SHAPs because of the ease of computation. Note that when predictors are independent then the two types of explanations coincide; for the case when dependencies are present see the discussion at the end of the section.

Intuition. For a given model f and the explainer E_i the explanation β_i quantifies the W_1 distance between the distributions of the explainer $E_i|G=0$ and $E_i|G=1$, that is, this value is an assessment of the bias introduced by the predictor X_i. The value β_i is the area between the corresponding subpopulation explainer CDFs $F_{E_i|G=k}$, $k \in \{0,1\}$, similar to the area depicted in Fig. 4. The value β_i^+ represents the bias across quantiles of the explainer E_i for which the predictor X_i favors the non-protected class $G=0$ and β_i^- represents the bias across quantiles for which X_i favors the protected class $G=1$. The β_i^{net} assesses the net contribution across different quantiles and represents an explanation that allows one to assess whether *on average* the predictor X_i favors class $G=0$ or class $G=1$; see Lemma 10.

In what follows we consider several simple examples to get more intuition behind the bias explanation values as well as discuss their additivity or the lack thereof. To avoid complex notation when the context is clear we suppress the dependence of the bias explanations on X and the explainer E.

Definition 10 Let f, X, G, and E_i be as in Definition 9.

- We say that E_i strictly favors class $G=0\,(G=1)$ if $\beta_i^-(f|G;E_i)=0$ $(\beta_i^+(f|G;E_i)=0)$.
- We say that X_i has mixed bias explanations if $\beta_i^{\pm}(f|G;E_i)>0$.

Offsetting. Since each predictor may favor one class or the other, the predictors may offset each other in terms of the bias contributions to the model bias. To understand the offsetting effect consider a binary classification risk model ($\varsigma_f=-1$) with two predictors:

$$\begin{aligned} &X_1 \sim N(\mu+G,1), \quad X_2 \sim N(\mu-G,1)\\ &Y \sim Bernoulli(f(X)), \quad f(X)=\mathbb{P}(Y=1|X)=logistic(2\mu-X_1-X_2) \end{aligned} \tag{M3}$$

where $\mu=5$, and $\{X_i|G=k\}_{i,k}$ are independent and $\mathbb{P}(G=0)=\mathbb{P}(G=1)$. We next train logistic regression score $\hat{f}(X)$, with $\varsigma_{\hat{f}}=-1$, and choose the explainer to be $E_i=PDP_i$. By construction the explanation E_1 of the predictor X_1 strictly favors class $G=0$, while that of X_2 strictly favors class $G=1$. Moreover,

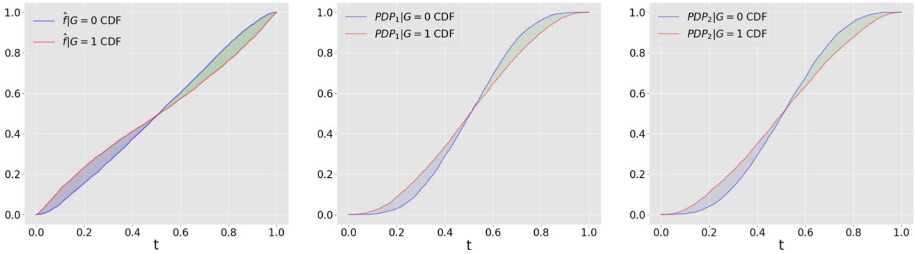


Fig. 6 Model and PDP biases for the model (M4), $\varsigma_{\hat{f}} = -1$

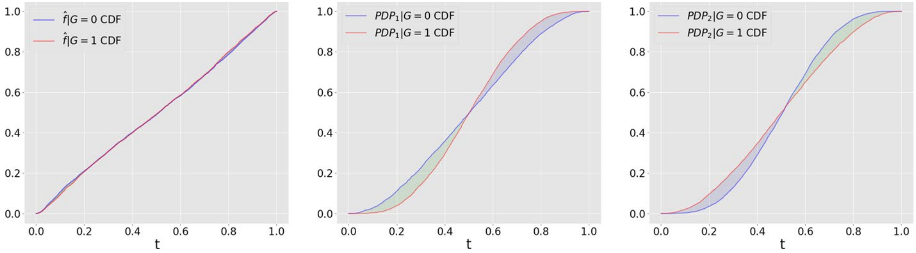


Fig. 7 Model and PDP biases for the model (M5), $\varsigma_{\hat{f}} = -1$

$$\beta_1(\hat{f}|G;E_1) = \beta_1^+(\hat{f}|G;E_1) = \beta_2^-(f|G;E_2) = \beta_1(\hat{f}|G;E_2) \approx 0.17.$$

Combining the two predictors at the model level leads to bias offsetting. By construction the resulting model bias is $\mathrm{Bias}_{W_1}(f|G) = 0$. Figure 5 displays the CDFs for the trained score subpopulations $\hat{f}|G = k$ and the corresponding explainers $E_i|G = k$, which illustrates the offsetting phenomena numerically.

Another important point we need to make is that the equality $\beta_i^{net} = 0$ does not in general imply that the predictor X_i has no effect on the model bias. This is a consequence of (22). Moreover, predictors with mixed bias might amplify the model bias as well as offset it. To understand how mixed bias predictors interact at the level of the model bias consider the following risk classification model ($\varsigma_f = -1$).

$$\begin{aligned} &X_1 \sim N(\mu, 1+G), \quad X_2 \sim N(\mu, 1+G) \\ &Y \sim Bernoulli(f(X)), \quad f(X) = \mathbb{P}(Y=1|X) = logistic(2\mu - X_1 - X_2). \end{aligned} \tag{M4}$$

where $\mu = 5$, and $\{X_i|G=k\}_{i,k}$ are independent and $\mathbb{P}(G=0) = \mathbb{P}(G=1)$. As before we train a logistic regression score $\hat{f}$, with $\varsigma_{\hat{f}} = -1$, and choose $E_i = PDP_i$. By construction, the true classification score f satisfies $\beta_i^{net}(f|G) = 0$ for each predictor X_i. Furthermore, the CDFs of explainers satisfy

$$(F_{E_i(X,f)|G=0}(t) - F_{E_i(X,f)|G=1}(t)) \cdot \mathrm{sgn}(t - 0.5) > 0$$

for any threshold $t \neq 0.5$. Combining the two predictors at the level of the model leads to amplifying the positive and negative model biases and hence the model bias itself. Figure 6 displays the CDFs for the trained score subpopulations $\hat{f}|G = k$ and the corresponding

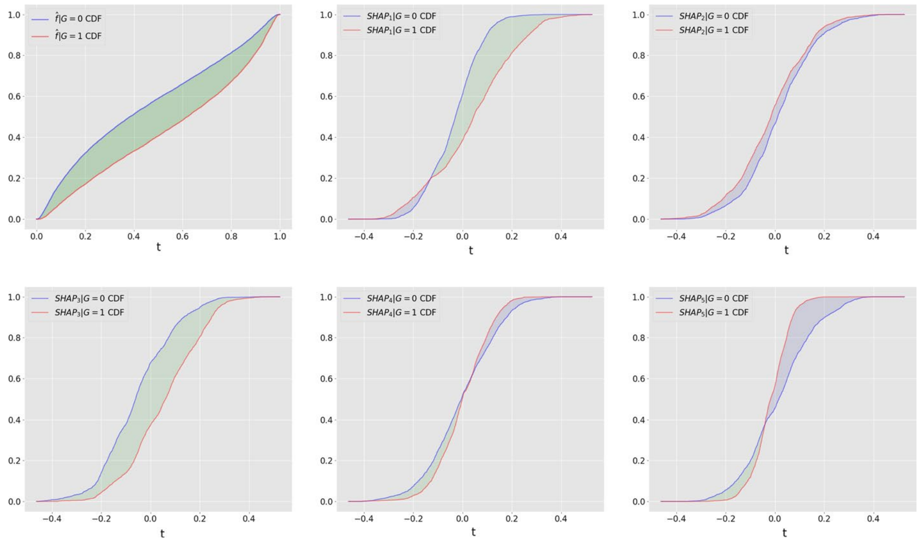


Fig. 8 Model bias and SHAP explainer biases for trained XGBoost (M6), $\varsigma_{\hat{f}} = -1$

explainers $E_i(\hat{f})|G = k$. The numerics illustrate that as long as the regions for positive and negative bias of mixed predictors agree, when combined they will increase the model bias.

If the regions of positive and negative bias for two predictors do not agree, then offsetting will happen. To see this, let us modify the above example as follows:

$$\begin{aligned} &X_1 \sim N(\mu, 2-G), X_2 \sim N(\mu, 1+G) \\ &Y \sim Bernoulli(f(X)), \quad f(X) = \mathbb{P}(Y=1|X) = logistic(2\mu - X_1 - X_2). \qquad \text{(M5)} \end{aligned}$$

By construction, $\beta_i^{net}(f|G) = 0$ for each predictor. However, the region of thresholds where the explainer $E_1(f)$ favors class $G = 0$ coincides with the region where $E_2(f)$ favors class $G = 1$, and the same holds for the two complimentary regions. This leads to bias offsetting so that $\text{Bias}_{W_1}(f|G) = 0$. The numerical results for this example are displayed in Fig. 7.

Bias explanation plots. Given a machine learning model f, predictors $X \in \mathbb{R}^n$, protected attribute G, and the explainers E_i, the corresponding bias explanations

$$\left\{(\beta_i, \beta_i^+, \beta_i^-, \beta_i^{net})(f|G;E_i)\right\}_{i=1}^n$$

are sorted according to any desired entry in the 4-tuple and then displayed in that order. This plot is called *Bias Explanation Plot* (BEP).

To showcase how BEP works, consider a classification risk model ($\varsigma_f = -1$):

$$\begin{aligned} &\mu = 5, \quad a = \tfrac{1}{20}(10, -4, 16, 1, -3) \\ &X_1 \sim N(\mu - a_1(1-G), 0.5+G), \quad X_2 \sim N(\mu - a_2(1-G), 1) \\ &X_3 \sim N(\mu - a_3(1-G), 1), \quad X_4 \sim N(\mu - a_4(1-G), 1-0.5G) \\ &X_5 \sim N(\mu - a_5(1-G), 1-0.75G) \\ &Y \sim Bernoulli(f(X)), \quad f(X) = \mathbb{P}(Y=1|X) = logistic(\textstyle\sum_i X_i - 24.5). \qquad \text{(M6)} \end{aligned}$$

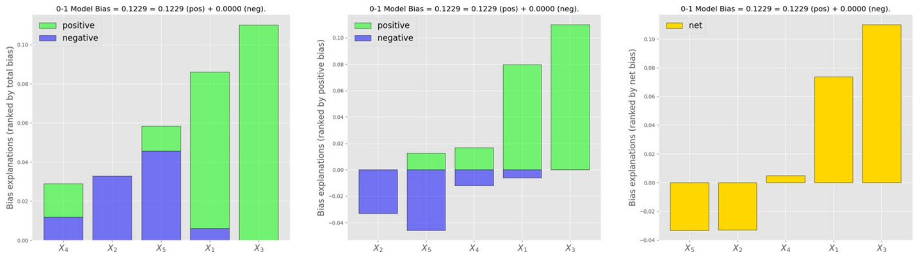


Fig. 9 Bias explanations ranked by β_i and β_i^+ and ranked β_i^{net} for the model (M6), $\varsigma_{\hat{f}} = -1$

where $\{X_i|G=k\}_{i,k}$ are independent and $\mathbb{P}(G=0) = \mathbb{P}(G=1)$. We next generate 20, 000 samples from the distribution (X, Y) and train a regularized XGBoost model which produces the score $\hat{f}$. Figure 8 displays the CDFs of the subpopulation scores $\hat{f}|G=k$ (top left), and those of the explainers $E_i = \varphi_i(\hat{f}, v^{ME})$. We see that there is positive model bias in the plot showing the CDFs, thus class $G=0$ is favored. For the predictors, the bias explanation plots show that X_1, X_4 and X_5 have mixed biases that arise due to differences in subpopulation variances of predictors, while the bias in X_2 strictly favors class $G=1$ and the bias in X_3 favors $G=0$.

The numerically computed model bias and its disaggregation are given by

$$(\text{Bias}_{W_1}, \text{Bias}^+_{W_1}, \text{Bias}^-_{W_1}, \text{Bias}^{net}_{W_1})(\hat{f}|G) = (0.1533, 0.1533, 0, 0.1533)$$

The bias explanations are then computed as the Earth Mover distance, and its disaggregation, between the distributions of subpopulation explainers $E_i(\hat{f})|G=k$. The bias explanations are given by

$$\begin{aligned}
(\beta_1, \beta_1^+, \beta_1^-, \beta_1^{net}) &= (0.0860, 0.0799, 0.0061, 0.0738)\\
(\beta_2, \beta_2^+, \beta_2^-, \beta_2^{net}) &= (0.0328, 0, 0.0328, -0.0328)\\
(\beta_3, \beta_3^+, \beta_3^-, \beta_3^{net}) &= (0.1100, 0.1100, 0, 0.1100)\\
(\beta_4, \beta_4^+, \beta_4^-, \beta_4^{net}) &= (0.0289, 0.0169, 0.0119, 0.0050)\\
(\beta_5, \beta_5^+, \beta_5^-, \beta_5^{net}) &= (0.0584, 0.0127, 0.0457, -0.0330)
\end{aligned}$$

Figure 9 displays the above bias explanations for each predictor in increasing order by total bias (left), positive bias (middle), and ranked net bias (right), respectively. Clearer information can be obtained from these plots compared to Fig. 8. For example, one can now see how mixed X_1, X_4, X_5 are and how the positive and negative parts compare.

Relationship with model bias The positive and negative bias explanations provide an informative way to determine the main drivers for positive and negative bias among predictors, which can be done by ranking the bias attributions. However, though informative, the positive and negative bias explanations are *not additive*. That is, in general

$$\text{Bias}^{\pm}_{W_1}(\hat{f}|G) \neq \textstyle\sum_{i=1}^n \beta_i^{\pm}(\hat{f}|G; E_i).$$

The main reason for lack of additivity is the presence of *bias interactions* which happen at the level of quantiles, or thresholds. The bias explanations by design compute the contribution to the cost of transport but do not track how mass is transported; see Figs. 6, 7. To better understand the bias interactions, motivated by Štrumbelj and Kononenko (2010), we introduce a game theoretic approach in Sect. 4.5 that yields additive bias explanations.

For additive models with independent predictors, however, we have the following result.

Lemma 12 *Let $X \in \mathbb{R}^n$ be predictors. Let the model f be additive, that is, $f(X) = \sum_{i=1}^n f_i(X_i)$. Let an explainer E_i be either $v^{ME}(\{i\};X,f)$ or $\varphi_i[v^{ME}(\cdot;X,f)]$. Let $\{\beta_i, \beta_i^+, \beta_i^-, \beta_i^{net}\}_i$ be the bias explanations of (X, f). Then*

$$\mathrm{Bias}_{W_1}^{net}(f|G) = \mathrm{Bias}_{W_1}^{+}(f|G) - \mathrm{Bias}_{W_1}^{-}(f|G) = \sum_{i=1}^n \left(\beta_i^+ - \beta_i^-\right) = \sum_{i=1}^n \beta_i^{net}.$$

If X are independent then the lemma holds for E_i in the form $v^{CE}(\{i\};X,f)$ or $\varphi_i[v^{CE}(\cdot;X,f)]$.

Proof Suppose that $E_i(X;f) = v^{ME}(\{i\};X,f)$. Then, in view of the additivity of f, we have

$$v^{ME}(\{i\};X,f) = f_i(X_i) - \mathbb{E}[f_i(X_i)] + \mathbb{E}[f(X)]$$

and hence by Lemma 10 we have

$$\beta_i^{net}(f|G;v^{ME}) = \left(\mathbb{E}[f_i(X_i)|G=0] - \mathbb{E}[f_i(X_i)|G=1]\right) \cdot \varsigma_f.$$

Summing up the net bias explanations gives

$$\begin{aligned}\sum_i \beta_i^{net}(f|G;v^{ME}) &= \sum_i \left(\mathbb{E}[f_i(X_i)|G=0] - \mathbb{E}[f_i(X_i)|G=1]\right) \cdot \varsigma_f \\ &= \left(\mathbb{E}[f(X)|G=0] - \mathbb{E}[f(X)|G=1]\right) \cdot \varsigma_f = \mathrm{Bias}_{W_1}^{net}(f|G).\end{aligned} \tag{23}$$

Suppose that $E_i(X;f) = \varphi_i(X;f, v^{ME})$. Since $\{X_i\}_{i=1}^n$ are independent and f is additive,

$$\varphi_i(X;f, v^{ME}) = \varphi_i(X;f, v^{CE}) = f_i(X_i) - \mathbb{E}[f_i(X_i)] = v^{ME}(\{i\};X,f) + \mathbb{E}[f(X)].$$

Since a shift in the distribution does not affect the bias, the bias explanation based on $\varphi_i[v^{ME}]$ coincide with that of v^{ME}. This together with (23) and the independence assumption proves the lemma. □

Example Let f be as in Lemma 12. Suppose that f is either positively biased or negatively biased, that is, $\mathrm{Bias}_{W_1}(f|G) = (1-\delta) \cdot \mathrm{Bias}_{W_1}^{+}(f|G) + \delta \cdot \mathrm{Bias}_{W_1}^{-}(f|G)$ with $\delta \in \{0,1\}$. Then

$$\mathrm{Bias}_{W_1}(f|G) = (-1)^{\delta} \sum_{i=1}^n (\beta_i^+ - \beta_i^-).$$

4.4 Stability of marginal and conditional bias explanations

Under dependencies the marginal and conditional bias explanations differ in their description. The conditional bias explanations rely on the joint distribution (X, Y) and encapsulate the interaction between the bias in predictors and the response variable, while the marginal explanations encapsulate the interaction between bias in predictors and the structure of the model, that is, the map $x \to f(x)$; for details see Miroshnikov et al. (2021a). In particular, we have the following result.

Theorem 4 (stability) *Let $X \in \mathbb{R}^n$ be predictors. Let $E_i = \varphi_i[v]$, $v \in \{v^{CE}, v^{ME}\}$. The bias explanations based on the marginal and conditional Shapley values satisfy the following*:

(i) *For all $f, g \in L^2(P_X)$, we have*

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

(ii) *For all $f, g \in L^2(\widetilde{P}_X)$, we have*

$$|\beta_i^{\pm}(f|X, G, \varphi_i[v^{ME}]) - \beta_i^{\pm}(g|X, G, \varphi_i[v^{ME}])| \leq C\|f - g\|_{L^2(\widetilde{P}_X)}.$$

Proof Take $f, g \in L^2(P_X)$. Take $i \in \{1, 2, \dots, n\}$ and set

$$A = \varphi_i[v^{CE}(\cdot; X, f)], \quad B = \varphi_i[v^{CE}(\cdot; X, g)].$$

Let $\mu_k = P_{A|G=k}$, $\nu_k = P_{B|G=k}$, and $\gamma = P_{(A,B)|G=k}$ for $k \in \{0, 1\}$. By construction $\gamma_k \in \Pi(\mu_k, \nu_k)$ and hence

$$\begin{aligned}\sum_{k\in\{0,1\}} W_1(\mu_k, \nu_k) &\leq \sum_{k\in\{0,1\}} \int |x_1 - x_2| P_{(A,B)|G=k}(dx_1, dx_2)\\ &\leq \sum_{k\in\{0,1\}} \mathbb{E}[|A - B|G = k]\\ &\leq C\|A - B\|_{L^2(\mathbb{P})} \leq C\|f - g\|_{L^2(P_X)}\end{aligned}$$

where $C = \max_{k\in\{0,1\}} \left\{\frac{1}{\mathbb{P}(G=k)}\right\}$ and the last inequality follows from Lemma 9(*i*).

Then, using the triangle inequality and the inequality above, we obtain

$$\begin{aligned}|\beta_i(f|X, G, \varphi_i[v^{CE}]) - \beta_i(g|X, G, \varphi_i[v^{CE}])| &= |W_1(\mu_1, \mu_2) - W_1(\nu_1, \nu_2)|\\ &\leq W_1(\mu_1, \nu_1) + W_1(\nu_2, \mu_2)\\ &\leq C\|f - g\|_{L^2(P_X)}.\end{aligned}$$

We next establish the bounds for the net-bias explanations. Assuming $\varsigma_f = \varsigma_g$ and using Lemma 10 we obtain

$$\begin{aligned}&|\beta_i^{net}(f|X, G, \varphi_i[v^{CE}]) - \beta_i^{net}(g|X, G, \varphi_i[v^{CE}])|\\ &= |\mathbb{E}[A|G = 0] - \mathbb{E}[A|G = 1] - \mathbb{E}[B|G = 0] + \mathbb{E}[B|G = 1]|\\ &\leq \sum_{k\in\{0,1\}} \mathbb{E}[|A - B||G = k]\\ &\leq C\|A - B\|_{L^2(P)} \leq C\|f - g\|_{L^2(P_X)}.\end{aligned}$$

Combining the above inequalities and using the fact that $\beta^{\pm} = \frac{1}{2}(\beta \pm \beta^{net})$ gives (*i*). To prove (*ii*), we follow the same steps as above and use Lemma 9(*ii*). □

Remark 4 Proposition 4 implies that the map $f \to \beta_i^{\pm}(f|X, G, \underset{\sim}{\varphi}_i[v^{CE}])$ is continuous in $L^2(P_X)$ and the map $f \to \beta_i^{\pm}(f|X, G, \varphi_i[v^{ME}])$ is continuous in $L^2(\underset{\sim}{P}_X)$.

4.5 Shapley-bias explanations

As discussed in Sect. 4.2, the non-additive bias explanations measure the positive and negative contributions to the model bias, but not to each flow. To this end, we measure signed contributions to each positive and negative model bias by employing a game-theoretic approach, which has been explored in numerous works in the area of machine learning interpretability; see Lipovetsky and Conklin (2001); Štrumbelj and Kononenko (2010); Lundberg and Lee (2017). In the spirit of Štrumbelj and Kononenko (2010), we define a cooperative game in which the players are predictors and the payoff is their bias contributions and then compute corresponding additive Shapley values.

Group explainers. Let $X \in \mathbb{R}^n$ be predictors and f a model. A generic *group explainer* of f is denoted by

$$E(S;X,f), \quad S \subset \{1, 2, \dots, n\}.$$

We assume that $E(S;\ X,\ f)$ quantifies the attribution of each predictor X_S with $S \subset \{1, 2, \dots, n\}$ to the model value $f(X)$ and satisfies

$$E(\emptyset, X,f) = \mathbb{E}[f(X)], \quad E(\{1, 2, \dots, n\};X,f) = f(X).$$

Relatively straightforward group explainers can be constructed using conditional and marginal game or game value. In particular, for a nonempty $S \subset \{1, 2, \dots, n\}$ one can set a trivial group explainer as

$$v(S;X,f) \quad \text{or} \quad \varphi_S[v] = \varphi_S(X;f, v) = \sum_{i \in S} \varphi_i(X;f, v) \quad \text{where} \quad v \in \{v^{CE}, v^{ME}\}. \tag{24}$$

Definition 11 Let X, G, f, ς_f be as in Definition 9. Let $E(\cdot\,;X,f)$ be a group explainer.

- Cooperative bias-game v^{bias} associated with X, G, f and E is defined by

  $$v^{bias}(S;G, E(\cdot;X,f)) = W_1(E(S;X,f)|G = 0, E(S;X,f)|G = 1), \quad S \subset \{1, 2, \dots, n\}.$$

  $v^{bias}(S)$ is the minimal cost of transporting $E(S)|G = 0$ to $E(S)|G = 1$ and vice versa.
- Under optimal transport the positive bias-game and negative bias-game, respectively, are defined by:

  $v^{bias+}(S)$ is the effort of transporting $E(S)|G = 0$ in the non-favorable direction.
  $v^{bias-}(S)$ is the effort of transporting $E(S)|G = 0$ in the favorable direction.

  The above values are specified in Lemma 2 for $q = 1$.
- Net bias-game is defined by

  $$v^{bias,net} = v^{bias+} - v^{bias+}.$$

- The Shapley-bias explanations of (X, f) based on the group explainer E are defined by

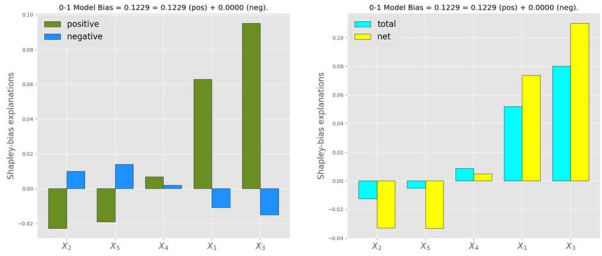


Fig. 10 Additive Shapley-bias explanations based on the game $v^{bias,ME}$ for the model (M6)

$$\varphi^{bias}(f|G) = \varphi[v^{bias}], \quad \varphi^{bias\pm}(f|G) = \varphi[v^{bias\pm}], \quad \varphi^{bias,net}(f|G) = \varphi[v^{bias,net}] \tag{25}$$

where φ denotes the Shapley value (19) and where we suppressed the dependence on X and E.

Unlike the regular bias explanations which by construction are always non-negative, the Shapley-bias explanations are signed, that is, they can be both positive and negative.

Lemma 13 *Given* (X, f) *and the explainer* E, *the Shapley bias-explanations defined in* (25) *satisfy*

$$\sum_{i=1}^{n} \varphi_i^{bias} = \mathrm{Bias}_{W_1}(f|G), \quad \sum_{i=1}^{n} \varphi_i^{bias\pm} = \mathrm{Bias}_{W_1}^{\pm}(f|G), \quad \sum_{i=1}^{n} \varphi_i^{bias,net} = \mathrm{Bias}_{W_1}^{net}(f|G)$$

and, thus,

$$\begin{aligned} \varphi[v^{bias}] &= \varphi[v^{bias+}] + \varphi[v^{bias-}] \\ \varphi[v^{bias,net}] &= \varphi[v^{bias+}] - \varphi[v^{bias-}]. \end{aligned}$$

Proof The result follows from Shapley (1953) and the properties of the W_1-based model bias. □

For Shapley-bias explanations based on the conditional and marginal games we have the following.

Theorem 5 *Given* (X, f), *let the conditional and marginal bias games be defined by*

$$\begin{aligned} v^{bias,CE}(S;X,f) &= v^{bias}(S\,;\varphi_S[v^{CE}(\cdot;X,f)]) \\ v^{bias,ME}(S;X,f) &= v^{bias}(S\,;\varphi_S[v^{ME}(\cdot;X,f)]) \end{aligned}$$

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

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

Proof The proof follows the same steps as in Theorem 4. □

Example Applying the above methodology to $\hat{f}$ and G of the model (M6) we compute the Shapley-bias explanations of predictors X_i, $i \in \{1, 2, \dots, 5\}$ using the group explainer $E(S) = \varphi_S[v^{ME}]$ defined in (24) for the construction of the bias-game.

The results are displayed in Fig. 10. On the left, the explanations are plotted in increasing order of the positive bias, and in the middle plot by the total bias, while the right plot contains the information on all four types of biases. By comparing these to the non-additive bias explanation plots in Fig. 9 we see how the signed values provide further information on how the predictors contribute to the model bias.

For example, from (M6) we have that X_3, as a contributor to the model $\hat{f}$, favors the class $G = 0$ since $\beta_3^+ > 0$ and $\beta_3^- = 0$. Recall that β_3^+ captures the total contribution to the increase of the positive model bias plus the decrease (or resistance) to the negative model bias. The Shapley-bias explanations, however, allow one to estimate separately the (signed) contributions to both positive and negative model bias.

In particular, the left plot of Fig. 10 informs us that X_3 in $\hat{f}$ contributes to the increase of the positive model bias (green), measuring the contribution to pushing the subpopulation of the non-protected class towards the favorable direction, while its contribution to the negative model bias (blue) is negative, which indicates the resistance towards the subpopulation's pull in the non-favorable direction.

4.6 Group Shapley-bias explanations

It might be important for a practitioner to understand the main factors within the data itself that contribute to the bias in the response variable and not how the model structure contributes to it. To do this, one needs to generate bias explanations based on the conditional game v^{CE}. The conditional game, when predictors are independent, coincides with the marginal game and the conditional expectations $\mathbb{E}[f(X)|X_S]$ can be computed through averaging with error control. However, under dependencies, the conditional expectations and corresponding Shapley-bias explanations are difficult to compute in light of the curse of dimensionality.

Another important aspect to consider is that highly dependent predictors carry similar information. For instance, in the case where a group of predictors is represented via a smaller collection of latent variables, the latent variable explanations are spread out among the predictors in that group; see Chen et al. (2020). Under dependencies, for practical and business purposes, one may want to explain the information carried by the entire group rather than the predictors themselves.

The two issues mentioned above can be addressed simultaneously by adapting the ideas from Aas et al. (2020); Miroshnikov et al. (2021a). In particular, grouping predictors based on dependencies and utilizing specially-designed group explainers to compute the contribution of the group help unite the marginal and conditional approaches. Therefore, applying similar techniques, one can approximate the conditional Shapley-bias explanations of weakly independent groups using the marginal approach, which only requires averaging over a small dataset. Furthermore, grouping allows one to reduce complexity.

In what follows we adapt the techniques from Miroshnikov et al. (2021a) to construct group Shapley-bias explanations. To this end, we first introduce required notation. Let $X \in \mathbb{R}^n$ and $\{S_j\}_{j=1}^m$ be disjoint sets that partition the set of the predictors' indexes,

$$N = \{1, 2, \dots, n\} = \bigcup_{j=1}^m S_j, \quad \mathcal{P} = \{S_1, S_2, \dots, S_m\}, \tag{26}$$

so that $X_{S_1}, X_{S_2}, \dots, X_{S_m}$ form weakly independent groups such that within each group the predictors share significant amount of mutual information. Given a cooperative game v on N, we define the quotient game by

$$v^{\mathcal{P}}(A) = v\big(\textstyle\bigcup_{j\in U} S_j\big), \quad A \subset M = \{1, 2, \dots, m\}.$$

By design, $v^{\mathcal{P}}(A)$ is played by the groups, viewing the elements of the partition as players.

Definition 12 Given X, f, G as in Definition (9), and the partition $\mathcal{P}$ as in 26.

- The conditional and marginal group bias-games are defined by

$$v^{bias}_{\mathcal{P}}(A; X, G, f, v) = W_1\big(v^{\mathcal{P}}(A)|G = 0, v^{\mathcal{P}}(A)|G = 1\big), \quad v \in \{v^{CE}, v^{ME}\}. \tag{27}$$

- The corresponding Shapley-bias explanations of $\{X_{S_j}\}_{j=1}^m$ are then defined by

$$\varphi^{bias,\mathcal{P}}_{S_j}(f|X, G; v) = \varphi_j[v^{bias}_{\mathcal{P}}(\cdot\,; v)], \quad v \in \{v^{CE}, v^{ME}\}.$$

Lemma 14 *Given $X, f, G, \mathcal{P}$ as in Definition 12. If $\{X_{S_j}\}_{j=1}^m$ are independent, then*

$$\varphi^{bias,\mathcal{P}}_{S_j}(f|X, G; v^{CE}) = \varphi^{bias,\mathcal{P}}_{S_j}(f|X, G; v^{ME}), \quad S_j \in \mathcal{P}. \tag{28}$$

Consequently,

$$|\varphi^{bias,\mathcal{P}}_{S_j}(f|X, G; v) - \varphi^{bias,\mathcal{P}}_{S_j}(g|X, G; v)| \le C\|f - g\|_{L^2(P_X)}, \quad v \in \{v^{CE}, v^{ME}\}.$$

Proof By independence, we have $v^{ME,\mathcal{P}} = v^{CE,\mathcal{P}}$. Hence by (27) we obtain

$$v^{bias}_{\mathcal{P}}(A; v^{CE}) = v^{bias}_{\mathcal{P}}(A; v^{ME}), \quad A \subset M$$

and this yields (28). The stability argument can be carried out similarly to Lemma 4. □

Similar construction is used to compute positive and negative bias explanations $\varphi^{bias+,\mathcal{P}}_{S_j}$ and $\varphi^{bias-,\mathcal{P}}_{S_j}$, respectively.

Remark 5 The importance of equality (28) is that the expression on the right-hand side can be computed via averaging using a dataset with $O(\tau^{-2})$ samples for a given error tolerance τ. This makes the computation of the conditional explanation feasible. Furthermore, the complexity of computations becomes $O(2^m)$ where m is the number of independent groups. For example, given a classification score and $X \in \mathbb{R}^{100}$, having 100 predictors split into 10 independent groups, it is sufficient to use a dataset with 10000 samples in order to compute conditional group Shapley-bias explanations of a classification score with error tolerance $\tau = 0.01$ and complexity $O(2^{10} \cdot 10000^2)$, which is feasible and easily parallelizable. If the number of independent groups is still large the above technique can modified to incorporate recursive groupings.

5 On the application of the framework

5.1 Bias mitigation under regulatory constraints

In this section, we will discuss how the fairness interpretability framework can be used for real-world applications in financial institutions that work under regulatory constraints.

An operational flow for model development in many FIs may consists of the following stages: (1) Model training, (2) Fair Lending Compliance governance review, and (3) Production, which includes model prediction and decision-making steps. Steps 1 and 3 are carried out by quantitative departments, while step 2 by the dedicated Compliance Office (CO), a department separate from business. The CO provides oversight to the company's compliance with federal and state regulations.

FIs are explicitly prohibited from collecting some protected information on customers such as race and ethnicity (apart from mortgage lending), as stated by the ECOA. Furthermore, protected attributes cannot be used in training or inference. However, proxy information on the protected attribute such as the one derived from Bayesian Improved Surname and Geocoding (BISG) is allowed to be used by the compliance office solely for fairness analysis (Elliot et al. 2009). Proxy information, however, must remain within the compliance office and the business does not (and should not) have access to the proxy data.

For fairness assessment, the CO carries out the bias assessment step. The CO can determine the main drivers contributing to model bias using our method and subsequently utilize bias mitigation methods. The bias mitigation step can include model postprocessing. However, in order to adhere to regulations, a post-processed model must not utilize the proxy attribute $\tilde{G}$ or any information on the joint distribution $(X, \tilde{G})$, such as probabilities $\mathbb{P}(\tilde{G}|X)$. The reasons for that are a) in the production step one can only have access to X, and b) a post-processed model is shared with business units that should be prevented from inferring the protected attribute from X.

Some rudimentary techniques for bias mitigation include recommendations on which predictors to drop from training or model post-processing via nullifying a given predictor by fixing its value. A more efficient technique has been proposed in our companion paper Miroshnikov et al. (2021b). There we construct an efficient frontier over a family of compliant post-processed models utilizing the interpretability framework developed in the current article. Other examples of compliant methods include those that vary hyper-parameters to get an efficient frontier, such as those in Schmidt et al. (2021).

5.2 Pedagogical example on bias mitigation

In this section we provide a pedagogical example that showcases how to properly utilize the information on the positive and negative bias explanations when it comes to bias mitigation. A rudimentary mitigation technique one can employ is to construct a regulatory-compliant post-processed model by neutralizing an appropriate collection of predictors X_S. This is accomplished by fixing their values in the model to some reference values x_S^* and setting $\tilde{f}(x;S,x^*) = f(x_S^*, x_{-S})$.

Often the objective of the bias mitigation in FIs is the reduction of the positive model bias which quantifies how much the model favors the majority class. In practice, regressor models are usually positively-biased, meaning $\text{Bias}_{W_1}^{+}(f|G) > 0$ and $\text{Bias}_{W_1}^{-}(f|G) = 0$.

Taking into account the above discussion, let us assume for the sake of explanation that $f(X) = \sum_{i=1}^{n} f_i(X_i)$ is an additive and positively-biased model. Let β_i^+, β_i^-, where $i \in N = \{1, \ldots, n\}$, be the positive and negative marginal bias explanations, respectively. Finally, let us decompose the predictor index set as follows: $N = N_+ \cup N_- \cup N_0$ where

$$N_+ = \{i : \beta_i^+ > \beta_i^-\}, \quad N_- = \{i : \beta_i^- > \beta_i^+\}, \quad N_0 = \{i : \beta_i^+ = \beta_i^-\}.$$

In this case, by Lemma 12 the model bias is given by

$$\text{Bias}_{W_1}(f|X,G) = \text{Bias}^+_{W_1}(f|X,G) = \sum_{i \in N_+} (\beta_i^+ - \beta_i^-) - \sum_{i \in N_-} (\beta_i^- - \beta_i^+) > 0$$

which illustrates the bias offsetting mechanism.

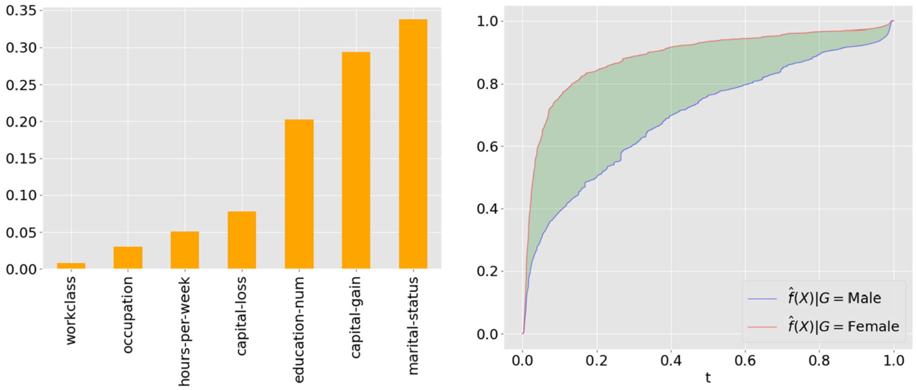
Note that neutralizing the predictor $i_0 \in N_-$ would cause the model bias, which is equal to the positive model bias, to increase, while neutralizing $i_1 \in N_+$would cause the model bias to decrease.

Thus, one approach to reduce the model bias is to rank order the predictors in N_+ by their net-bias explanations and, subsequently, neutralize them one by one in that order. This will incrementally reduce the positive model bias until the point where neutralizing the next predictor causes the model bias to become equal to the negative model bias (with the positive model bias being zero), which operates as a stopping criterion of the approach. This simple and rather naÃ¯ve strategy illustrates that a) the decomposition of explanations is useful for bias mitigation and that b) neutralization of biased predictors ranked by total bias contribution is not always the optimal strategy.

5.3 Example on census income dataset

In this section, we showcase the application of the framework to the 1994 Census Income dataset from the UCI Machine Learning Repository (Dheeru et al. 2017).

This dataset contains fourteen predictors and a dependent variable Y that indicates if an individual earns more or less than \$50K annually. After investigating the predictors, we removed the protected attributes ‘sex’, ‘race’, ‘age’, and ‘native-country’. We also excluded ‘fnlwght’ and ‘relationship’, the latter due to its high dependence with ‘sex’ since in the dataset the categorical values ‘Husband’ and ‘Wife’ correspond to ‘Male’ and ‘Female’, respectively. The remaining seven predictors used for model training are ‘workclass’, ‘education-num’, ‘occupation’, ‘marital-status’, ‘capital-gain’, ‘capital-loss’, and ‘hours-per-week’.



(a) Feature importance.

(b) Score subpopulation CDFs; bias ≈ 0.19.

Fig. 11 Model training and protected attribute analysis

Algorithm 1: Model bias and bias explanations

```
Data: Model $f$, dataset $D = (X, G)$ with $m$ samples, $X \in \mathbb{R}^n$ and $G \in \{0,1\}$, and model explainer $E_i$.
Result: Output the model biases $\text{Bias}^{\pm}_{W_1}(f|X,G)$ and predictor bias explanations $\{\beta_i^{\pm}\}_{i=1}^{n}$.
Partition the predictors in $D$ according to $G = k$, $k \in \{0,1\}$. This yields subsets $D_{X,0}$, $D_{X,1}$.
For each $k \in \{0,1\}$ evaluate $f(x)$ on $D_{X,k}$ to obtain the set of subpopulation model values $S_k$.
For each $k \in \{0,1\}$ compute the empirical CDF $\hat{F}_k$ of $f(X)|G = k$ based upon $S_k$.
$\text{Bias}^{\pm}_{W_1}(f|X,G) := \int_{\mathscr{P}_{\pm}} |\hat{F}_0^{[-1]} - \hat{F}_1^{[-1]}|\,dp$ with $\mathscr{P}_{\pm}$ as in Definition 7.
for $i$ in $\{1,\dots,n\}$ do
    For each $k \in \{0,1\}$ evaluate $E_i(x)$ on $D_{X,k}$ to obtain the set of subpopulation values $S_{i,k}$.
    For each $k \in \{0,1\}$ compute the empirical CDF $\hat{F}_{i,k}$ of $E_i(X)|G = k$ based upon $S_{i,k}$.
    $\beta_i^{\pm} := \int_{\mathscr{P}_{i\pm}} |\hat{F}_{i,0}^{[-1]} - \hat{F}_{i,1}^{[-1]}|\,dp$ with $\mathscr{P}_{i\pm}$ as in Definition 9.
end
```

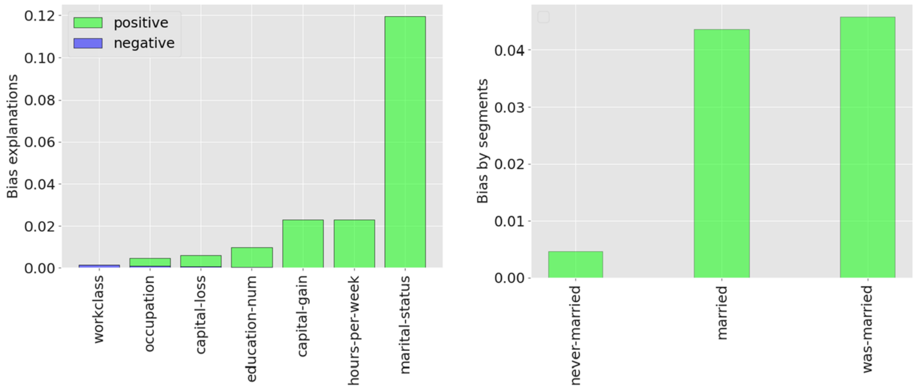
For the model training, we use the training dataset D_{train} with 32561 samples to build a classification score

$$\hat{f}(x) = \widehat{\mathbb{P}}(Y = \} > 50\text{K}'|X = x),$$

using Gradient Boosting. For training we use the following parameters: `n_estimators`=200, `min_samples_split`=5, `subsample`=0.8, `learning_rate`=0.1. The feature importance of each predictor can be seen in Fig. 11a, with the most significant predictors being ‘marital-status’, ‘capital-gain’, and ‘education-num’.

Performance metrics for the GBM model on the trained dataset, and test dataset with 16251 samples, were evaluated. Specifically, the mean logloss on the train and test set is approximately 0.288 and 0.292 respectively, and the AUC is 0.922 and 0.918 respectively.

The focus of the application is to evaluate and explain the model bias with respect to the protected attribute $G =$‘sex’, with values ‘Female’ and ‘Male’, where ‘Female’ is the protected class. To this end, following the steps in Algorithm 1, we form the dataset S containing the classification scores



(a) Bias explanations. **(b)** Model bias segmented by marital status.

Fig. 12 Model bias explanations

$$S = \left\{\hat{f}(x^{(i)}) \,:\, (x^{(i)}, y^{(i)}) \in D_{train}\right\},$$

and partition it based on each class of G. This yields the sets S_M and S_F containing the classification scores for 'Female' and 'Male' respectively, which we use to construct the empirical CDFs of the subpopulation scores, $\hat{F}_{Female}$ and $\hat{F}_{Male}$, using the `ECDF` class from the `statsmodels` library.

Figure 11b depicts the empirical CDFs, where we see that the model has almost exclusively positive bias, and the positive direction is assumed to be $\varsigma_{\hat{f}} = 1$. To confirm this observation, we subsequently compute the positive and negative model biases by integrating the difference of the two CDFs over the sets where $\hat{F}_{Female} > \hat{F}_{Male}$ and $\hat{F}_{Female} < \hat{F}_{Male}$, respectively, as indicated in Definition 7. This yields the following values:

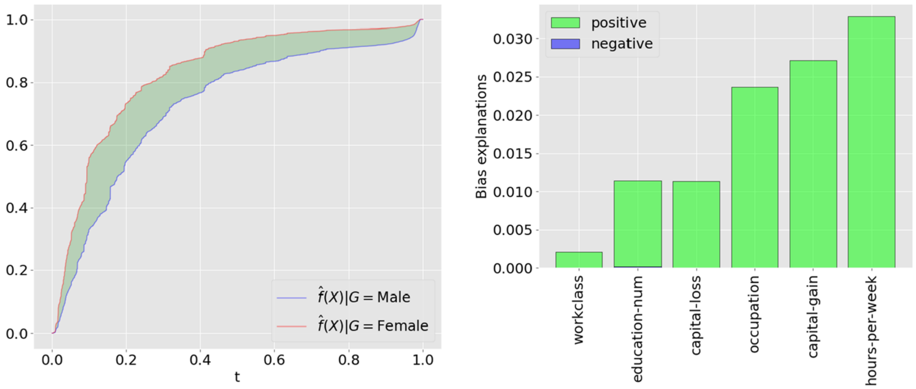
$$\text{Bias}^{+}_{W_1}(\hat{f}|X, G) \approx 0.19, \quad \text{Bias}^{-}_{W_1}(\hat{f}|X, G) \approx 0.00.$$

To understand the contributions of the predictors to the model bias, we next construct the bias explanations based on the marginal model explainer. To accomplish this, we subsample the predictors from the training set, and obtain a background dataset D_X with $m = 4000$ samples. Next, we compute the model explanations for each predictor X_i yielding the sets

$$S_{E_i} = \left\{ \frac{1}{m} \sum_{x \in D_X} \hat{f}(x_i^*, x_{-i}),\ x^* \in D_X \right\}.$$

Similar to obtaining the model bias, we then partition S_{E_i} based on each class of G and obtain the empirical CDFs of $E_i(X)|G = g,\ g \in \{\text{'Female'}, \text{'Male'}\}$, which are then used to compute the bias explanations $\beta_i^{\pm}$ according to Definition 9. These are depicted in Fig. 12a and are ranked in ascending order of the positive bias. All the values for the negative bias explanations are close to zero, which further indicates the positively biased nature of the predictors. Observe that the most positively contributing predictor to the model bias is 'marital-status' by far with value ≈ 0.12.

Since 'marital-status' is the most impactful, it merits further investigation into its effect on the model bias. To this end, we group the different values of 'marital-status' into three categories: M_1 ='never-married', M_2 ='married', and M_3 ='was-married'. Then, we segment



(a) Score subpopulation CDFs; bias ≈ 0.10.

(b) Bias explanations.

Fig. 13 Bias explanations for the re-trained model without 'marital-status' predictor

the dataset S of classification scores into three subsets S_{M_i}, $i \in \{1, 2, 3\}$, that correspond to the aforementioned categories. To gain further understanding on how each of these categories contributes to the model bias, we compute the model bias on each segment. The negative model bias on each segment turns out to be zero, while the positive model biases are plotted in Fig. 12b. The plot indicates that the category 'never-married' exhibits an insignificant level of bias, while there is some substantial positive bias in 'married' and 'was-married'.

Given the above analysis, one can attempt to reduce the model bias either by applying the postprocessing technique discussed in Sect. 5.2, or, alternatively, retrain the model by dropping some of the biased predictors. We showcase the latter approach by dropping 'marital-status' and retraining the model with the same parameters. We check the performance of the new model on the train and test sets. The mean logloss is 0.358 and 0.363 respectively, and AUC is 0.862 and 0.855 respectively. We then compute the model bias and bias explanations; see Fig. 13. The positive model bias has been reduced to approximately 0.10, while the negative stays zero. The trade-off is a drop in performance, as seen by the performance metric values above. The bias explanations in the retrained model have slightly increased since 'marital-status' was dropped and the importance of the remaining predictors increased.

We would like to point out that the technique used above might not lead to bias reduction under the presence of strong dependencies, since other predictors could be used as proxies for the dropped predictor. However, the postprocessing technique outlined in Sect. 5.2 modifies the model score directly and the dependencies do not play a significant role. Keep in mind that this technique is rather crude and one may opt to employ the postprocessing methods described in Miroshnikov et al. (2021b) which apply to numerical predictors, but can be adjusted for categorical ones.

6 Conclusion

In this paper, we presented a novel bias interpretability framework for measuring and explaining bias in classification and regression models at the level of a distribution that utilizes the Wasserstein metric and the theory of optimal mass transport. We introduced and theoretically characterized bias predictor attributions to the model bias and constructed additive bias explanations utilizing cooperative game theory. To our knowledge, bias interpretability methods at the level of a regressor distribution have not been addressed in the literature before.

At a higher level, the model bias is a non-trivial superposition of predictor bias attributions. The bias explanations we introduced determine the contribution of a given predictor to the model bias. However, any two or more predictors will interact in the context of the bias explanations. For example, if one predictor favors the non-protected class and the other favors the protected class, it might be possible that when both predictors are utilized by the model the total effect on model bias is zero. This phenomenon opens up numerous avenues for future research to investigate the interactions of predictors across subpopulation distributions in the context of bias explanations. This is where ML interpretability techniques can come into play and aid with the study of predictor interactions in the model bias.

To make bias explanations additive we utilized cooperative game theory which lead to additive Shapley-bias explanations. These explanations rely on the Shapley formula, which makes them computationally expensive. The intractability of such calculations can be mitigated by grouping predictors based on dependencies and then computing the Shapley bias attributions for each group (via a quotient game) which reduces the dimensionality. However, if the number of groups is large, the issue of computational intensity remains. Thus, a possible research direction is to investigate methods that allow for approximation of the additive bias explanations and their fast computations.

In this paper, we formulated a methodology that computes the model bias and quantifies the contribution of predictors to that bias. However, an important application of the bias explanation methodology lies in bias mitigation, which will be useful in regulatory settings such as the financial industry, and may utilize information about the main drivers of the model bias. This will be investigated in our upcoming paper. The framework is generic and in principle can be applied to a wide range of predictive ML systems. For instance, it might be insightful to understand the predictor attributions to probabilistic differences of populations studied in physics, biology, medicine, economics, etc.

Appendix

A. Kantorovich transport problem

To formulate the transport problem we need to introduce the following notation. Let $\mathcal{B}(\mathbb{R}^k)$ denote the σ-algebra of Borel sets. The space of all Borel probability measures on $\mathbb{R}^k$ is denoted by $\mathscr{P}(\mathbb{R}^k)$. The space of probability measure with finite q-th moment is denoted by

$$\mathscr{P}_q(\mathbb{R}^k) = \{\mu \in \mathscr{P}(\mathbb{R}^k) : \int_{\mathbb{R}^k} |x|^q d\mu(x) < \infty\}.$$

Definition 13 (*push-forward*)

(a) Let $\mathbb{P}$ be a probability measure on a measurable space $(\Omega, \mathscr{F})$. Let $X \in \mathbb{R}^p$ be a random vector defined on Ω. The push-forward probability distribution of $\mathbb{P}$ by X is defined by

$$P_X(A) := \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \in A\}\big).$$

(b) Let $\mu \in \mathscr{P}(\mathbb{R}^k)$ and $T : \mathbb{R}^k \to \mathbb{R}^m$ be Borel measurable, the pushforward of μ by T, which we denote by $T_\# \mu$ is the measure that satisfies

$$(T_\# \mu)(B) = \mu\big(T^{-1}(B)\big), \quad B \subset \mathcal{B}(\mathbb{R}^k).$$

(c) Given measure $\mu = \mu(dx_1, dx_2, ..., dx_k) \in \mathscr{P}(\mathbb{R}^k)$ we denote its marginals onto the direction x_j by $(\pi_{x_j})_\# \mu$ and the cumulative distribution function by

$$F_\mu(a_1, a_2, \dots, a_k) = \mu((-\infty, a_1] \times (-\infty, a_2] \dots, (-\infty, a_k])$$

Theorem 6 (change of variable) *Let $T : \mathbb{R}^k \to \mathbb{R}^m$ be Borel measurable map and $\mu \in \mathscr{P}(\mathbb{R})$. Let $g \in L^1(\mathbb{R}^m, T_\# \mu)$. Then*

$$\int_{\mathbb{R}^m} g(y) T_\# \mu(dy) = \int_{\mathbb{R}^k} g(T(x))\, \mu(dx).$$

Proof See (Shiryaev 1980, p. 196). □

Definition 14 (*Kantorovich problem on* $\mathbb{R}$) Let $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R})$ and $c(x_1, x_2) \geq 0$ be a cost function. Consider the problem

$$\inf_{\gamma \in \Pi(\mu_1, \mu_2)} \left\{ \int_{\mathbb{R}^2} c(x_1, x_2) \gamma(dx_1, dx_2) \right\} =: \mathscr{T}_c(\mu_1, \mu_2)$$

where $\Pi(\mu_1, \mu_2) = \{\gamma \in \mathscr{P}(\mathbb{R}^2) : (\pi_{x_j})_\# \gamma = \mu_j\}$ denotes the set of transport plans between μ_1 and μ_2, and $\mathscr{T}_c(\mu_1, \mu_2)$ denotes the minimal cost of transporting μ_1 into μ_2.

Definition 15 Let $q \geq 1$ and let d be a metric on $\mathbb{R}$. Let the set $\mathscr{P}_q(\mathbb{R}^n; d) = \{\mu \in \mathscr{P}(\mathbb{R}^n) : \int d(x, x_0)^q d\mu(x) < \infty\}$ where x_0 is any fixed point. The Wasserstein distance W_q on $\mathscr{P}_q(\mathbb{R}^n; d)$ is defined by

$$W_q(\mu_1, \mu_2; d) := \mathscr{T}^{1/q}_{d(x_1, x_2)^q}(\mu_1, \mu_2), \quad \mu_1, \mu_2 \in \mathscr{P}_q(\mathbb{R}^n; d)$$

where

$$\mathscr{T}_{d(x_1, x_2)^q}(\mu_1, \mu_2) = \inf_{\gamma \in \mathscr{P}(\mathbb{R}^2)} \left\{ \int_{\mathbb{R}^2} d(x_1, x_2)^q d\gamma, \quad \gamma \in \Pi(\mu_1, \mu_2) \right\}.$$

We drop the dependence on d in the notation of the Wasserstein metric when $d(x, y) = |x - y|$.

The following theorem contains well-known facts established in the texts such as Shorack and Wellner (1986); Villani (2003); Santambrogio (2015).

Theorem 7 *Let $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R})$. Let $c(x_1, x_2) = h(x - y) \geq 0$ with h convex and let*

$$\pi^* := (F^{-1}_{\mu_1}, F^{-1}_{\mu_2})_\# \lambda|_{[0,1]} \in \mathscr{P}(\mathbb{R}^2)$$

where $\lambda|_{[0,1]}$ *denotes the Lebesgue measure restricted to* $[0,\ 1]$. *Suppose that* $\mathscr{T}_c(\mu_1,\mu_2) < \infty$. *Then*

(1) $\pi^* \in \Pi(\mu_1,\mu_2)$ *and* $F_{\pi^*} = \min(F(a), F(b))$.
(2) π^* *is an optimal transport plan that is*

$$\mathscr{T}_c(\mu_1,\mu_2) = \int_{\mathbb{R}^2} h(x_1 - x_2)\, d\pi^*(x_1,x_2).$$

(3) π^* *is the only monotone transport plan, that is, it is the only plan that satisfies the property*

$$(x_1,x_2),(x_1',x_2') \in \mathrm{supp}(\pi^*) \subset \mathbb{R}^2 \quad x_1 < x_1' \quad \Rightarrow \quad x_2 \le x_2'.$$

(4) *If* h *is strictly convex then* π^* *is the only optimal transport plan.*
(5) *If* μ_1 *is atomless, then* π^* *is determined by the monotone map* $T^* = F_{\mu_2}^{[-1]} \circ F_{\mu_1}$, *called an optimal transport map. Specifically,* $\mu_2 = T^*_{\#}\mu_1$ *and hence* $\pi^* = (I, T^*)_{\#}\mu_1$, *where* I *is the identity map. Consequently,*

$$\int_{\mathbb{R}^2} h(x_1 - x_2)\, d\pi^*(x_1,x_2) = \int_{\mathbb{R}} h(x_1 - T^*(x_1)) d\mu_1(x_1) = \mathbb{E}[X_1 - T^*(X_1)], \quad \mu_1 = P_{X_1}.$$

(6) *For* $q \in [1,\infty)$, *we have*

$$\begin{aligned} W_q{}^q(\mu_1,\mu_2) = \mathscr{T}_{|x_1-x_2|^q}(\mu_1,\mu_2) &= \int_{\mathbb{R}^2} |x_1 - x_2|^q d\pi^*(x_1,x_2) \\ &= \int_0^1 |F_{\mu_1}^{[-1]}(p) - F_{\mu_2}^{[-1]}(p)|^q dp < \infty. \end{aligned}$$

Definition 16 Given a set of probability measures $\{\mu_j\}_{j=1}^J \subset \mathscr{P}_2(\mathbb{R}^n)$, with $J \ge 1$, with finite second moments, and weights $\{\omega_j\}_{j=1}^J$, the Wasserstein barycenter is the minimizer of the map $\nu \to \sum_{j\in J} \omega_j W_2^2(\nu,\mu_j)$.

B. Proofs and auxiliary lemmas

Definition 17 (*geometric continuity*) Let $D(\cdot,\cdot)$ be a metric on $\mathscr{P}_k(\mathbb{R}^n)$, with $k \ge 0$. We say that D is continuous with respect to the geometry of the distribution if for any $\mu \in \mathscr{P}_k(\mathbb{R}^n)$ $\lim_{\epsilon\to 0+} D(\mu, T_{\epsilon\#}\mu) = 0$, for any family $\{T_\epsilon\}_{\epsilon>0}$ of continuously differentiable maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ that satisfy

(i) $det\, \nabla T_\epsilon > 0$.
(ii) The family $\{T_\epsilon - I\}_\epsilon$ has a common compact support.
(iii) $T_\epsilon \to I$ uniformly on $\mathbb{R}^n$ as $\epsilon \to 0$, where I is the identity map.

Definition 18 (*invariance*) Let $D(\cdot,\cdot)$ a metric on $\mathscr{P}_k(\mathbb{R}^n)$. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a map such that $T_{\#}\mu \in \mathscr{P}_k(\mathbb{R}^n)$ for every $\mu \in \mathscr{P}_k(\mathbb{R}^n)$. We say that D is invariant under the transformation T if $D(\mu_1,\mu_2) = D\big(T_{\#}\mu_1, T_{\#}\mu_2\big)$.

Proof of Theorem 1 Let $q \in [1,\infty)$. Let T_ϵ be a family of maps from $\mathbb{R}$ to $\mathbb{R}$ as in Definition 17. Take $\mu \in \mathscr{P}_q(\mathbb{R})$. Since $T_\varepsilon - I$ has compact support, there is a bounded $B \subset \mathbb{R}$ such that $T_\varepsilon(x) = x$ for all $x \in B^c$. Thus,

$$\int_{\mathbb{R}} |x|^q dT_{\varepsilon\#}\mu(x) = \int_{\mathbb{R}} |T_\varepsilon(x)|^q d\mu(x) = \int_B |T_\varepsilon(x)|^q d\mu(x) + \int_{B^c} |x|^q d\mu(x) < \infty$$

and hence $T_{\varepsilon\#}\mu \in \mathscr{P}_q(\mathbb{R})$.

Next, consider a probability measure $\pi = (I, T_\varepsilon)_{\#}\mu$. By construction, its marginals are μ and $T_{\varepsilon\#}\mu$ and hence π is a transport plan. Then, Lemma 6 and the definition of the distance D_{W_q} imply

$$D^q_{W_q}(\mu_\varepsilon, T_{\varepsilon\#}\mu) \le \int_{\mathbb{R}^2} |x_1 - x_2| d\pi(x_1,x_2) = \int_{\mathbb{R}} |x_1 - T_\varepsilon(x_1)| d\mu(x_1).$$

Sending $\varepsilon \to 0$ in the above inequality, and using the assumption that $I - T_\varepsilon \to 0$ uniformly in $\mathbb{R}$, we conclude that $D^q_{W_q}(\mu, T_{\varepsilon\#}\mu) \to 0$. This proves the statement (a).

Let $T : \mathbb{R} \to \mathbb{R}$ be continuous and strictly increasing. Let $q \in [1,\infty)$. Suppose that D_{W_q} on $\mathscr{P}_q(\mathbb{R})$ is invariant under T. Let $\mu_1 = \delta_a$ and $\mu_2 = \delta_b$ for $a < b$. Then by invariance we obtain

$$(T(b) - T(a))^q = D^q_{W_q}(T_{\#}\mu_1, T_{\#}\mu_2) = D^q_{W_q}(\mu_1,\mu_2) = (b-a)^q.$$

Since a, b are arbitrarily chosen, we conclude that $T(x) = x + C$. This proves (b). □

Proof of Lemma 6 First, take any $M \in Lip_1(\mathscr{X}, \mathscr{P}(\{0,1\}))$ and set $\varphi(x) = [M(x)](\{0\})$. Then

$$d(x,y) \ge D_{TV}(M(x), M(y)) = \frac{1}{2} \sum_{a\in\{0,1\}} \big|[M(x)](a) - [M(y)](a)\big| = |\varphi(x) - \varphi(y)|$$

and hence $\tilde{\varphi} = \varphi - \frac{1}{2} \in \mathcal{A}^*$.

Next, take $\tilde{\varphi} \in \mathcal{A}^*$. Let $\varphi = \tilde{\varphi} + \frac{1}{2}$. Take $x \in \mathscr{X}$ and pick $M(x)$ to be a probability measure such that $[M(x)](\{0\}) = \varphi(x)$. Then $M \in Lip_1(\mathscr{X}, \mathscr{P}(\{0,1\}); D_{TV}, d)$

The lemma follows from the above and the fact that $M_\mu(\{0\}) - M_\nu(\{0\}) = \int \tilde{\varphi} d[\mu - \nu]$. □

Lemma 15 *Let $d(x,y) = \|x-y\|$ be a norm on $\mathbb{R}^n$. Let $T(x) = cx + x_0$ with $c > 0$. Then*

$$D_{rc}(T_{\#}\mu, T_{\#}\nu; D_{TV}, d) = D_{rc}(\mu,\nu; D_{TV}, d_c), \quad \mu,\nu \in \mathscr{P}_1(\mathbb{R}^n; d),$$

where $d_c(x,y) = cd(x,y)$.

Proof

□

Lemma 16 *Let $d(x,y) = \|x-y\|$ be a norm on $\mathbb{R}^n$ and $\mu,\nu \in \mathscr{P}_1(\mathbb{R}^n; d)$. Let $c > 0$. Then*

(i) $W_1(\mu, \nu; d_c) = c\, W_1(\mu, \nu; d)$, $d_c(x, y) = cd(x, y)$.

$$\begin{aligned} D_{rc}(T_\# \mu, T_\# \nu; D_{TV}, d) &= \sup_{\varphi \in Lip_1(\mathbb{R}^n, [0,1]; d)} \int \varphi(x)[\tilde{\mu} - \tilde{\nu}](dx) \\ &= \sup_{\varphi \in Lip_1(\mathbb{R}^n, [0,1]; d)} \int \varphi(cx + x_0)[\mu - \nu](dx) \\ &= \sup_{u \in Lip_1(\mathbb{R}^n, [0,1]; d_c)} \int u(x)[\mu - \nu](dx) \\ &= D_{rc}(\mu, \nu; D_{TV}, d_c). \end{aligned}$$

(ii) *For any* $T(x) = cx + x_0$

$$W_1(T_\# \mu, T_\# \nu; d) = cW_1(\mu, \nu; d).$$

Proof The lemma follows directly from the definition of W_1 and the fact that d is a norm. □

Proof of Lemma 7 The proof follows from Lemmas 15 and 16. □

Proof of Theorem 3 Take any $L > 0$ and x_* such that the supports of μ and ν are contained in $B(x_*, \frac{L}{2}; d)$. By the Kantorovich-Rubinstein duality theorem (Kantorovich 1958; Dudley 1976), we have

$$W_1(\mu, \nu; d) = \sup \left\{ \int u(x)[\mu - \nu](dx),\ u \in Lip_1(\mathbb{R}^n; d) \right\}.$$

Since $Lip_1(\mathbb{R}^n, [0, L]; d) \subset Lip_1(\mathbb{R}^n; d)$ we have

$$\sup \left\{ \int \tilde{u}(x)[\mu - \nu](dx),\ \tilde{u} \in Lip_1(\mathbb{R}^n, [0, L]; d) \right\} \leq W_1(\mu_1, \mu_2; d).$$

Next, take any $u \in Lip_1(\mathbb{R}^n; d)$. Observe that

$$u_0 := \inf_{x \in B(x_*, L/2)} u(x) = \big(u(x_*) + \inf_{x \in B(x_*, L/2)} (u(x) - u(x_*))\big) \in [u(x_*) - \tfrac{L}{2}, u(x_*) + \tfrac{L}{2}].$$

Define

$$\tilde{u}(x) = \min(\max(u(x) - u_0, 0), L).$$

Note that $\tilde{u} \in Lip_1(\mathbb{R}^n, [0, L]; d)$. Furthermore,

$$0 \leq u(x) - u_0 \leq \sup_{z \in B(x_*, \frac{L}{2})} d(x, z) \leq L, \quad x \in B(x_*, \tfrac{L}{2})$$

and hence $\tilde{u} = u(x) - u_0$ for $x \in B(x_*, \frac{L}{2})$. Then, since μ and ν have support in $B(x_*, \frac{L}{2})$, we have

$$\int u(x)[\mu - \nu](dx) = \int \tilde{u}(x)[\mu - \nu](dx)$$

and hence

$$\sup\left\{\int \tilde{u}(x)[\mu-\nu](dx),\ \tilde{u}\in Lip_1(\mathbb{R}^n,[0,L];d)\right\}\geq W_1(\mu,\nu;d).$$

Thus, we conclude

$$W_1(\mu,\nu;d)=\sup\left\{\int \tilde{u}(x)[\mu-\nu](dx),\ \tilde{u}\in Lip_1(\mathbb{R}^n,[0,L];d)\right\} \tag{29}$$

for any norm d and any ball $B(x_*,\frac{L}{2};d)$ containing the supports of μ and ν.

Let $d_{(1/L)}(x,y)=\frac{1}{L}\|x-y\|=\frac{1}{L}d(x,y)$. Then $B(x_*,\frac{1}{2};d_{(1/L)})=B(x_*,\frac{L}{2};d)$ and hence using (29) and Lemma 16, we obtain

$$\begin{aligned}&\frac{1}{L}W_1(\mu,\nu;d)=W_1(\mu,\nu;d_{(1/L)})\\&\quad=\sup\left\{\int \tilde{u}(x)[\mu-\nu](dx),\ \tilde{u}\in Lip_1(\mathbb{R}^n,[0,1];d_{(1/L)})\right\}=D_{rc}(\mu,\nu;D_{TV},d_{(1/L)})\end{aligned}$$

which proves the equality. □

Proof of Lemma 4 Define the set

$$A_0=\{(p,t)\in(0,1)\times\mathbb{R}\,:\,F_1(t)<p\leq F_0(t)\}.$$

Note $(p,t)\in A_0$ implies $t\in\mathcal{T}_0$. Hence, applying Lemma 18, we obtain

$$\lambda^2(A_0)=\int_{\mathcal{T}_0}F_0(t)-F_1(t)\,dp<\infty$$

where the finiteness of the right-hand side follows from the fact that $\mathbb{E}|X_i|<\infty$ and Lemma 30.

Observe next that the definition of the generalized inverse implies that

$$F_i^{[-1]}(p)\leq t\Leftrightarrow p\leq F_i(t),\qquad F_i^{[-1]}(p)>t\Leftrightarrow p>F_i(t)$$

and hence

$$A_0=\{(p,t)\in(0,1)\times\mathbb{R}\,:\,F_0^{[-1]}(p)\leq t<F_1^{[-1]}(p)\}.$$

Note by above $(p,t)\in A_0$ implies that $p\in\mathcal{P}_1$. Hence, Lemma 18 imply

$$\lambda^2(A_0)=\int_{\mathcal{P}_1}F_1^{[-1]}(p)-F_0^{[-1]}(p)\,dp$$

and this proves $(4)_1$. The proof of $(4)_2$ is similar. □

Lemma 17 *Let X be a random variable with $\mathbb{E}|X|<\infty$. Let $X^+=\max(0,X)$, $X^-=\max(0,-X)$. Then*

$$\mathbb{E}[X]=\mathbb{E}[X^+]-\mathbb{E}[X^-],\quad \mathbb{E}[X^+]=\int_0^\infty(1-F(t))\,dt,\quad \mathbb{E}[X^-]=\int_{-\infty}^0F(t)dt \tag{30}$$

where F is the CDF of X.

Proof Note that $|X(\omega)| \geq X^{+}(\omega), X^{-}(\omega) \geq 0$ and hence $\mathbb{E}[X^{+}]$ and $\mathbb{E}[X^{-}]$ are finite. Recalling that $X = X^{+} - X_{-}$, we obtain $(30)_1$.

Next, by definition of the expectation, we have

$$\begin{aligned}\infty > \mathbb{E}[X^{+}] &= \int_{\Omega} X^{+}(\omega)\,\mathbb{P}(d\omega) = \int_{\Omega}\left(\int_{\mathbb{R}} \mathbb{1}_{\{0\leq x\leq X^{+}(\omega)\}}dx\right)\mathbb{P}(d\omega)\\ &= \int_{\mathbb{R}} \mathbb{1}_{\{0\leq x\}}\left(\int_{\Omega} \mathbb{1}_{\{x\leq X^{+}(\omega)\}}\mathbb{P}(d\omega)\right)dx = \int_{0}^{\infty}(1 - F(x))\,dx\end{aligned}$$

where we applied the Tonelli's theorem to exchange the order of integration. This proves $(30)_2$. The proof for $(30)_3$ is similar. □

Lemma 18 *Let λ denote the Lebesgue measure on $\mathbb{R}$. Let f, g be λ-measurable functions such that $g \leq f$.*

(i) *If $f - g \in L^1(\mathbb{R})$, then*

$$\begin{aligned}\lambda\otimes\lambda\Big(\big\{(x,y)\,:\,g(x)<y<f(x)\big\}\Big) &= \int_{\mathbb{R}}(f-g)\,d\lambda\\ &= \lambda\otimes\lambda\Big(\big\{(x,y)\,:\,g(x)\leq y\leq f(x)\big\}\Big) < \infty.\end{aligned} \tag{31}$$

(ii) *If $\lambda\otimes\lambda\Big(\big\{(x,y)\,:\,g(x)<y<f(x)\big\}\Big) < \infty$, then $f - g \in L^1(\mathbb{R})$ and (31) holds.*

Proof Suppose that $f - g \in L^1(\mathbb{R})$. Since f and g are measurable, the set $\big\{(x,y)\,:\,g(x)<y<f(x)\big\}$ is measurable with respect to the product measure $\lambda^2 = \lambda\otimes\lambda$. Then by the Tonelli's theorem we obtain

$$\begin{aligned}\infty > \int_{\mathbb{R}}(f(x)-g(x))\,d\lambda(x) &= \int_{\mathbb{R}}\left(\int_{\mathbb{R}}\mathbb{1}_{\{y:g(x)<y<f(x)\}}\,d\lambda(y)\right)d\lambda(x)\\ &= \int_{\mathbb{R}^2}\mathbb{1}_{\{(x,y):g(x)<y<f(x)\}}\,d(\lambda\otimes\lambda) = \lambda^2\big(\{(x,y)\,:\,g(x)<y<f(x)\}\big),\end{aligned}$$

which proves the first equality in (31). The second equality (31) is proved similarly. This gives (i).

Suppose that $\lambda^2\big(\{(x,y)\,:\,g(x)<y<f(x)\}\big) < \infty$. Following the calculations above in the reverse order we conclude that $f - g \in L^1(\mathbb{R})$ and hence (31) holds. This proves (ii).

□

Acknowledgements The authors would like to thank Steve Dickerson (CAO, Decision Management at Discover Financial Services (DFS)), Raghu Kulkarni (VP, Data Science at DFS) and Melanie Wiwczaroski (former Sr. Director, Enterprise Fair Banking at DFS) for formulation of the problem as well as helpful business and compliance insights. We also thank Patrick Haggerty (former Director & Senior Counsel at DFS), Charla Hausler (Vice President Enterprise Compliance Programs at DFS) and Kate Prochaska (Sr. Counsel & Director, Regulatory Policy at DFS) for their helpful comments relevant to regulatory issues that arise in the financial industry. We also would like to thank professors Markos Katsoulakis and Robin Young from the University of Massachusetts Amherst, professor Matthias Steinrücken from the University of Chicago and professor Hangjie Ji from North Carolina State University for their valuable comments and suggestions that aided us in writing this article.

Authors' contributions Not applicable.

Availability of data and material All data in the examples of the article are synthetic, except for the dataset used in Sect. 5.3 which was taken from the UCI Machine Learning Repository which can be found at https://archive.ics.uci.edu/ml/datasets/census+income. In the synthetic examples, we provide explicit formulas for the joint data distribution.

Code availability Algorithm 1 in Sect. 5.3 presents the steps on the application of our interpretability framework. Since the code of Algorithm 1 is the intellectual property of Discover Financial Services, we provide the list of publicly available Python packages for its implementation. We used the package `sklearn.ensemble` for model training, `numpy.random` for generating random parameters and the joint distribution of synthetic data, `sklearn.metrics` for computing performance metrics, `empirical_distribution.ECDF` from the package `statsmodels.distributions` for computing the Wasserstein distance, and for the computation of Shapley values we used `itertools.combinations` and `numpy`.

Declarations

Funding Not applicable.

Conflict of interest Not applicable.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Aas, K., Jullum, M., & Løland, A., (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*.

Barocas, S., Hardt, M., Narayanan, A., *Fairness and Machine Learning: Limitations and Opportunities*. Available at: https://fairmlbook.org/.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., & Rey-Bellet, L. (2022). (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research*, *23*, 1–70.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research, 2*, 499–526.

Chen, H., Danizek, J., Lundberg, S., Lee, S.-I. (2020). True to the Model or True to the Data. *arXiv preprint* arXiv:2006.1623v1.

Chen, J., Kallus, N., Mao, X., Svacha, G., Udell, M. (2019). Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved, *FAT* '19: Proceedings of the conference on fairness, accountability, and transparency*, p. 339-348, https://doi.org/10.1145/3287560.3287594.

Dheeru, D., & Taniskidou, E.K. *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences, (2017), https://archive.ics.uci.edu/ml/datas ets/.

Dudley, R.M., (1976). *Probabilities and Metrics* Lecture Notes Series (Vol. 45). Matematisk Institut: Aarhus University, Aarhus.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R.S. (2012). Fairness through awareness. In *Proc. ACM ITCS*, 214-226.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities health services and outcomes research. *Methodology, 9*(2), 69–83.

Equal Credit Opportunity Act (1974). https://www.fdic.gov/regulations/laws/rules/6000-1200.html.

Equal employment opportunity act, (1972). https://www.dol.gov/sites/dolgov/files/ofccp/regs/compliance/posters/pdf/eeopost.pdf.

Fair housing Act, (1968). https://www.fdic.gov/regulations/laws/rules/2000-6000.html.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *In Proc. 21st ACM SIGKDD*, 259-268.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Gordaliza, P., Barrio, E.D., Fabrice, G., Loubes J.-M. (2019). Obtaining Fairness using Optimal Transport Theory, Proceedings of the 36th International conference on machine learning, PMLR 97:2357-2365.

Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., & Schmidt, N. (2021). United States fair lending perspective on machine learning. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2021.695301

Hardt, M., Price, E., & Srebro, N. (2015). Equality of opportunity in supervised learning. *In Advances in neural information processing systems, 3315-3323.*

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning* (2nd ed.). Springer.

Janzing, D., Minorics, L., Blöbaum, P. (2019). Feature relevance quantification in explainable AI: A causal problem. *arXiv preprint* arXiv:1910.13413v2.

Jiang, H., Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *Proceedings of the 23-rd International conference on artificial intelligence and statistics (AISTATS).*

Kamishima, T., Akaho, S., Asoh, H., & Sakuma J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer, *Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases (ECMLPKDD), Part II, pp.35-50.*

Kamiran F., & Calders, T. (2009). Classifying without discriminating, *2009 2nd International conference on computer, control and communication*, Karachi, pp. 1-6, https://doi.org/10.1109/IC4.2009.4909197,(2009).

Kantorovich, L. V., & Rubinstein, G. (1958). On a space of completely additive functions. *Vestnik Leningradskogo Universiteta, 13*(7), 52–59.

Kearns, M., & Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation, 11*(6), 1427–1453.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Analysis of regression in game theory approach, 17*, 319–330.

Lundberg S.M., Erion G.G., & Lee S.-I. (2019). Consistent individualized feature attribution for tree ensembles, *arXiv preprint* arxiv:1802.03888.

Lundberg, S.M., & Lee S.-I. (2017). A unified approach to interpreting model predictions, *31st Conference on neural information processing systems*.

Miroshnikov, A., Kotsiopoulos, K., Franks, R., Ravi Kannan, A. (2021). Model-agnostic bias mitigation methods with regressor distribution control for Wasserstein-based fairness metrics. *arXiv preprint* arxiv:2111.11259.

Miroshnikov, A., Kotsiopoulos, K., Ravi Kannan, A., (2021). Mutual information-based group explainers with coalition structure for machine learning model explanations. *arXiv preprint* arxiv:2102.10878.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability, 29*(2), 429–443.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Basel: Birkäuser Springer.

Schmidt, N., Curtis, J., Siskin, B., & Stocks, C. (2021). Methods for Mitigation of Algorithmic Bias Discrimination, Proxy Discrimination, and Disparate Impact. *U.S. Provisional Patent 63/153,692*.

Shapley, L. S. (1953). A value for n-person games. *Annals of Mathematics Studies, 28*, 307–317.

Shiryaev, A. (1980). *Probability*. Springer.

Shorack, G. R., & Wellner, J. A. (1986). *Empirical processes with applications to statistics*. New York: Wiley.

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B., & Lanckriet, G.R.G. (2009). On integral probability metrics, ϕ-divergences and binary classification, *arXiv preprint* arxiv:0901.2698.

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research, 11*, 1–18.

Sundararajan, M., Najmi, A. (2019). The many shapley values for model explanation. *arXiv preprint* arXiv:1908.08474.

Villani, C. (2003). *Topics in Optimal Transportation.* American Mathematical Society.

Woodworth, B., Gunasekar, S., Ohannessian, M.I., & Srebro, N. (2017). Learning nondiscriminatory predictors. *In Proc. of Conference on Learning Theory*, p. 1920-1953.

Young, H. P. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory, 14*(2), 65–72.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C. (2013). Learning Fair representations. *In Proc. of Intl. Conf. on Machine Learning, p. 325-333.*

Zhang, B.H., Lemoine, B., Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *In Proc. of the 2018 AAAI/ACM Conference on AI, ethics and society* (pp. 335-340)