On the Direction of Discrimination:
An Information-Theoretic Analysis of Disparate Impact in Machine Learning

Hao Wang, Berk Ustun, and Flavio P. Calmon
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Emails: hao_wang@g.harvard.edu, {berk, flavio}@seas.harvard.edu

Abstract—**In the context of machine learning, disparate impact refers to a form of systematic discrimination whereby the output distribution of a model depends on the value of a sensitive attribute (e.g., race or gender). In this paper, we propose an information-theoretic framework to analyze the disparate impact of a binary classification model. We view the model as a fixed channel, and quantify disparate impact as the divergence in output distributions over two groups. Our aim is to find a *correction function* that can perturb the input distributions of each group to align their output distributions. We present an optimization problem that can be solved to obtain a correction function that will make the output distributions statistically indistinguishable. We derive closed-form expressions to efficiently compute the correction function, and demonstrate the benefits of our framework on a recidivism prediction problem based on the ProPublica COMPAS dataset.**

I. INTRODUCTION

Machine learning (ML) models aim to exploit biases in the training data to predict an outcome of interest. In many real-world applications, however, effective prediction should not be achieved by discriminating on a *sensitive attribute*, such as race or gender [1, 2].

Discrimination can occur directly when a sensitive attribute is used as an input to the model, known as *disparate treatment*. More pervasive today is a phenomenon known as *disparate impact* [3], where a sensitive attribute is omitted from the model, but still affects its predictions through correlations with "proxy" variables (e.g., income, education level). The potential to discriminate by proxy is not unique to ML. In the United States, for example, racial minorities were indirectly denied financial services by exploiting correlations between race, address, and income – a practice known as *redlining* [4].

Disparate impact can arise as an artifact of empirical loss minimization when a sensitive attribute is valuable for prediction and can be approximately inferred using other *proxy* variables in the training data. A large body of recent work has documented this phenomenon in real-world applications ranging from online advertising [5] to recidivism prediction [6]. When used in human or algorithmic decision-making, models with disparate impact may violate anti-discrimination laws [3] and inadvertently amplify societal biases [7].

These issues have motivated a growing stream of technical work on disparate impact in ML, focusing on topics such as:

(i) how to identify and quantify disparate impact [8, 9, 10, 11]; (ii) how to train models that mitigate disparate impact [12, 13, 14]; and (iii) how to identify causal factors of discrimination [15]. The present work is inscribed within the first research direction.

In this paper, we consider the disparate impact problem from an information-theoretic perspective. Our goal is to derive a *correction function* that can be used to identify features that act as proxies of a sensitive attribute for a fixed prediction model. We first present a correction function that has an information-theoretic interpretation in terms of error exponents of binary hypothesis testing (Section II). We then derive closed-form expressions for the correction function that can easily be computed using the prediction model, which uses the features X to predict the outcome Y, and a group membership distribution, which uses the features X to "predict" the sensitive attribute S (Section III). Our approach is inspired by recent work in information-theoretic privacy [16], which takes a similar route to analyzing the behavior of error exponents under small perturbations. We illustrate our framework on a recidivism prediction problem derived from the ProPublica COMPAS dataset [6] (Section IV).

II. FRAMEWORK

We consider a channel $W_{Y|X}$, which takes as input a vector of d random variables $X = (X_1, \ldots, X_d) \in \mathcal{X}$ and produces as output a random variable $Y \in \mathcal{Y}$. We assume that the support sets $\mathcal{X}$ and $\mathcal{Y}$ are finite. In practice, $W_{Y|X}$ represents a predictive model (e.g., a linear classifier to predict recidivism), X represents a vector of features (e.g., *Age*, *Salary*), Y represents the predicted output of $W_{Y|X}$ given X (e.g., $Y = 1$ iff the model predicts that a prisoner with features X will commit a crime after being released from prison).

We seek to characterize differences in the output distribution of the channel $W_{Y|X}$ with respect to a *sensitive attribute S*. We focus on the case where the sensitive attribute is binary $S \in \{0, 1\}$, and use $P_{X|S=0}, P_{X|S=1}$ and $P_{Y|S=0}, P_{Y|S=1}$ to denote the conditional distributions of inputs and outputs, respectively. A channel $W_{Y|X}$ is said to have *disparate impact* with respect to S when $P_{Y|S=0} \neq P_{Y|S=1}$. We assume that $W_{Y|X}$ does not use the sensitive attribute S, as doing so would violate legal constraints in applications such as hiring and
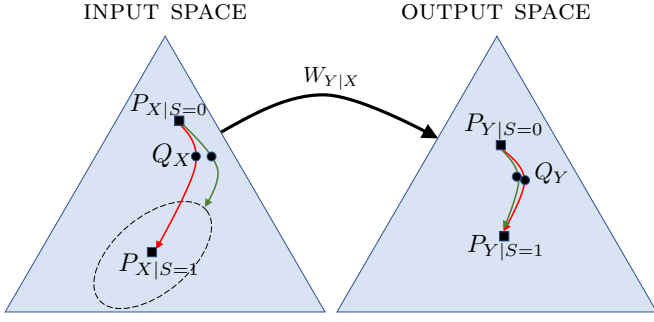
Fig. 1: Two correction paths on the probability simplex. The green line depicts a path that reduces disparate impact by aligning only output distributions. The red line depicts a path that reduces disparate impact by aligning the input distributions and output distributions.

credit scoring (see, e.g., [3]). In this setting, the Markov condition $S \to X \to Y$ ensures that $P_{Y|S=0} = W_{Y|X} \circ P_{X|S=0}$ and $P_{Y|S=1} = W_{Y|X} \circ P_{X|S=1}$. Thus, disparate impact occurs only when $P_{X|S=0} \neq P_{X|S=1}$.

Given a channel $W_{Y|X}$, disparate impact can be reduced by perturbing $P_{X|S=0}$ to a new distribution Q_X so that the resulting output distribution $Q_Y = W_{Y|X} \circ Q_X$ is "closer" to $P_{Y|S=1}$ (cf. Fig. 1). Intuitively, larger disparities between output distributions require larger perturbations, and the direction between Q_X and $P_{X|S=0}$ reflects which components of X contribute to this disparity. In what follows, we will define this setup formally.

Definition 1. Given a convex divergence metric $J(\cdot\|\cdot)$ (e.g., total variation or KL-divergence), a fixed joint distribution $P_{S,X,Y}$ such that $S \to X \to Y$, and $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ such that $\lambda_i \geq 0$, the objective function is defined as:

$$L_{\boldsymbol{\lambda}}(Q_X) \triangleq \lambda_1 J(Q_X\|P_{X|S=0}) + \lambda_2 J(Q_X\|P_{X|S=1}) \\ + \lambda_3 J(Q_Y\|P_{Y|S=0}) + \lambda_4 J(Q_Y\|P_{Y|S=1}). \tag{1}$$

The *correction path* between $P_{X|S=0}$ and $P_{X|S=1}$ is defined as an optimal solution to (1) for a fixed value of $\boldsymbol{\lambda}$:

$$L_{\boldsymbol{\lambda}}^{\text{opt}} \triangleq \min_{Q_X} L_{\boldsymbol{\lambda}}(Q_X). \tag{2}$$

From standard convexity results [e.g., 17], as the values of λ_i are changed, the distribution Q_X that minimizes (2) describes the lower boundary of the set

$$\{(J(Q_X\|P_{X|S=0}), J(Q_X\|P_{X|S=1}), \\ J(Q_Y\|P_{Y|S=0}), J(Q_Y\|P_{Y|S=1})) \mid \text{supp}(Q_X) \subseteq \mathcal{X}\}.$$

Solving (2) will produce different paths on the probability simplex depending on the components of $\boldsymbol{\lambda}$ (see Fig. 1). When $\lambda_2 = \lambda_4 = 0$, the optimal solution is $Q_X = P_{X|S=0}$. When $\lambda_2 = \lambda_3 = 0$, Q_X will traverse the shortest path (as measured by $J(\cdot\|\cdot)$) between $P_{X|S=0}$ and the set $\{P_X \mid J(W_{Y|X} \circ P_X\|P_{Y|S=1}) = 0\}$ (i.e., the green line in Fig. 1). Note that this path transforms $P_{X|S=0}$ into a distribution devoid of disparate impact, but potentially different from $P_{X|S=1}$. Perhaps of greater interest is the path obtained when $\lambda_3 = 0$ (i.e., the red line in Fig. 1). In this case, varying $\lambda_1, \lambda_2, \lambda_4 \geq 0$

corresponds to traversing between $P_{X|S=0}$ and $P_{X|S=1}$, while controlling the similarity of the induced distribution on Y.

Our goal is to produce a *correction function* that indicates which components of X contribute to disparate impact. More precisely, the correction function is a local (multiplicative) perturbation of $P_{X|S=0}$ that decreases the objective function (1) the most (see Definition 2). This definition leads to correction functions that can be cast in terms of predictive models for S and Y given X, as we show in Section III.

Definition 2. For a given function $f \in \mathcal{L}(P_{X|S=0})$, we define the perturbed distribution $\widetilde{P}_{X|S=0}$ as

$$\widetilde{P}_{X|S=0}(\boldsymbol{x}) \triangleq P_{X|S=0}(\boldsymbol{x})(1 + \epsilon f(\boldsymbol{x})), \tag{3}$$

where $\epsilon > 0$ is chosen so that $\widetilde{P}_{X|S=0}$ is a valid probability distribution, and

$$\mathcal{L}(P_{X|S=0}) \triangleq \\ \{f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}\left[f(X)|S=0\right] = 0, \mathbb{E}\left[f(X)^2|S=0\right] = 1\}.$$

Moreover, $\widetilde{P}_{Y|S=0} \triangleq W_{Y|X} \circ \widetilde{P}_{X|S=0}$.

Next, we denote $\Delta_{\boldsymbol{\lambda}}(f)$ as the decrease in the objective function (1) by locally perturbing the distribution $P_{X|S=0}$.

Definition 3. For a given $\boldsymbol{\lambda}$, we define $\Delta_{\boldsymbol{\lambda}}(f)$ as

$$\Delta_{\boldsymbol{\lambda}}(f) \triangleq \lim_{\epsilon \to 0^+} \frac{L_{\boldsymbol{\lambda}}(\widetilde{P}_{X|S=0}) - L_{\boldsymbol{\lambda}}(P_{X|S=0})}{\epsilon}. \tag{4}$$

Definition 4. For a given $\boldsymbol{\lambda}$, the correction function $\widetilde{f}$ is the minimizer of $\Delta_{\boldsymbol{\lambda}}(f)$:

$$\widetilde{f} = \operatorname*{argmin}_{f \in \mathcal{L}(P_{X|S=0})} \Delta_{\boldsymbol{\lambda}}(f). \tag{5}$$

We remark that the influence of local perturbations on probability distributions has been studied both in statistics (see, e.g., [18]) and information theory (see e.g., [19]).

Connections to Binary Hypothesis Testing

In the remainder of this paper, we consider settings where the outcome variable is binary $Y \in \{0, 1\}$, and the divergence measure is the KL-divergence $J(\cdot\|\cdot) = D_{\text{KL}}(\cdot\|\cdot)$. In this case, the objective function (1) can be expressed as:

$$L_{\boldsymbol{\lambda}}(Q_X) = \lambda_1 D_{\text{KL}}(Q_X\|P_{X|S=0}) + \lambda_2 D_{\text{KL}}(Q_X\|P_{X|S=1}) \\ + \lambda_3 D_{\text{KL}}(Q_Y\|P_{Y|S=0}) + \lambda_4 D_{\text{KL}}(Q_Y\|P_{Y|S=1}). \tag{6}$$

Our choice of $D_{\text{KL}}(\cdot\|\cdot)$ is motivated by its relationship with the error exponent in hypothesis testing (see Ch 11 in [17], [20, 21]). Specifically, when $\lambda_3 = \lambda_4 = 0$ in (2), the correction path describes the best trade-off (in terms of the first order term in the exponent) between the Type I and Type II error for a hypothesis test that seeks to distinguish S given an observation of X. The optimal value of (2) can then be expressed in terms of the Rényi's α-divergence (cf. [21, Section II-A]), i.e., $D_{\alpha}(P_{X|S=0}\|P_{X|S=1})$. In other words, under the choice of KL-divergence the correction path can be understood as the trade-off between error exponents of two (independent) binary

hypothesis tests: one to distinguish S from X; and another to distinguish S from Y.

III. MAIN RESULTS

In this section, we derive closed-form expressions for the correction function for the objective function defined in (6). We leverage the connection between small perturbations of KL-divergences and maximal correlation [as noted, for example, by 22, 23].

Our main results consist of Theorems 1 and 2, where we prove that the correction function that minimizes (5) is a linear combination of two components: f_l, which aligns the perturbed input distribution with $P_{X|S=1}$; and f_m, which aligns the corresponding output distribution with $P_{Y|S=1}$. In Theorem 2, we show that f_l and f_m (and thus $\widetilde{f}$) can be expressed in terms of the group membership distribution $P_{S|X}$ and the channel $W_{Y|X}$. This result has an important practical benefit: it allows us to compute the correction function $\widetilde{f}$ directly using only $P_{S|X}$ and $W_{Y|X}$, without computing the complete joint distribution $P_{S,X,Y}$. In what follows, we present a formal statement of these results.

We start our derivation of the correction function by providing a simplified expression for $\Delta_{\boldsymbol{\lambda}}(f)$ in Lemma 1.

Lemma 1. *For a given $\boldsymbol{\lambda}$, $\Delta_{\boldsymbol{\lambda}}(f)$ can be expressed as*

$$
\Delta_{\boldsymbol{\lambda}}(f) = \lambda_2 \mathbb{E}\left[f(X) \log \frac{P_{X|S=0}(X)}{P_{X|S=1}(X)} \middle| S = 0 \right]
$$
$$
+ \lambda_4 \mathbb{E}\left[g(Y) \log \frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)} \middle| S = 0 \right], \tag{7}
$$

where $g(y) = \mathbb{E}\left[f(X)|Y = y, S = 0\right]$.

Next, we introduce definitions used to derive the correction function.

Definition 5. *The log-likelihood ratio functions f_l and g_l are given by*

$$
f_l(\boldsymbol{x}) \triangleq \log \frac{P_{X|S=0}(\boldsymbol{x})}{P_{X|S=1}(\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_{X|S=0}(X)}{P_{X|S=1}(X)} \middle| S = 0 \right], \tag{8}
$$

$$
g_l(y) \triangleq \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)} - \mathbb{E}\left[\log \frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)} \middle| S = 0 \right]. \tag{9}
$$

Definition 6. *The maximal correlation* (see, e.g., [24]) *between X and Y given $S = 0$ is defined as*

$$
\rho_m(P_{X|S=0}; W_{Y|X}) \triangleq \max_{\substack{f \in \mathcal{L}(P_{X|S=0}) \\ g \in \mathcal{L}(P_{Y|S=0})}} \mathbb{E}\left[f(X)g(Y)|S = 0\right].
$$

The maximal correlation can be equivalently given by

$$
\rho_m(P_{X|S=0}; W_{Y|X}) = \sqrt{\mathbb{E}\left[\mathbb{E}\left[g_m(Y)|X, S = 0\right]^2 \middle| S = 0 \right]}.
$$

We refer to the functions that attain the maximum as the *principal functions* and denote them as (f_m, g_m).

Theorems 1 and 2 characterize the correction function $\widetilde{f}$. Theorem 1 shows that the correction function is the linear combination of the log-likelihood ratio function f_l and the principal function f_m. Theorem 2 shows that f_l and f_m can be expressed in terms of the group membership distribution $P_{S|X}$ and the channel $W_{Y|X}$, respectively.

Theorem 1. *Given $\boldsymbol{\lambda}$, the correction function $\widetilde{f}$ has the form*

$$
\widetilde{f} = n_l f_l + n_m f_m, \tag{10}
$$

where n_l and n_m are constants computed as:

$$
n_l = \frac{-\lambda_2}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}},
$$
$$
n_m = \frac{-\lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}},
$$
$$
a_1 = \mathbb{E}\left[f_l(X)f_m(X)|S = 0\right],
$$
$$
a_2 = \sqrt{\mathbb{E}\left[\left(f_l(X) - a_1 f_m(X)\right)^2 \middle| S = 0 \right]},
$$
$$
b_1 = \mathbb{E}\left[g_l(Y)g_m(Y)|S = 0\right].
$$

Theorem 2. *The log-likelihood ratio function f_l and the principal function f_m can be expressed as*

$$
f_l(\boldsymbol{x}) = \log \frac{P_{S|X}(0|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_{S|X}(0|X)}{P_{S|X}(1|X)} \middle| S = 0 \right], \tag{11}
$$

$$
f_m(\boldsymbol{x}) = \frac{(g_m(1) - g_m(0))W_{Y|X}(1|\boldsymbol{x}) + g_m(0)}{\rho_m(P_{X|S=0}; W_{Y|X})}. \tag{12}
$$

Here, $g_m(0) = \sqrt{p/(1-p)}$ and $g_m(1) = -\sqrt{(1-p)/p}$ where $p \triangleq P_{Y|S=0}(1)$.

Combining Theorems 1 and 2, we obtain a closed-form expression for the correction function $\widetilde{f}$. Note that, due to our definition of $\widetilde{f}$ in terms of local perturbations, the correction function does not depend on λ_1 and λ_3 as in (6). In Corollary 1, we provide an expression for $\Delta_{\boldsymbol{\lambda}}(\widetilde{f})$ in this case:

Corollary 1. *For a given $\boldsymbol{\lambda}$, $\Delta_{\boldsymbol{\lambda}}(\widetilde{f})$ is given by*

$$
\Delta_{\boldsymbol{\lambda}}(\widetilde{f}) = -\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}
$$

where a_1, a_2, b_1 are defined in Theorem 1.

We now instantiate our results for the case where $\lambda_2 = 0$, where our objective is to align only the output distributions (i.e., the green line in Fig. 1). As expected, the following corollary shows that the correction function $\widetilde{f}$ under this scenario is (up to a sign difference) the principal function f_m.

Corollary 2. *When $\lambda_2 = 0$, the correction function $\widetilde{f}$ is:*

$$
\widetilde{f} = -\text{sign}\left(\mathbb{E}\left[g_l(Y)g_m(Y)|S = 0\right]\right)f_m \tag{13}
$$

and

$$
\Delta_{\boldsymbol{\lambda}}(\widetilde{f})
$$
$$
= -\lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})\sqrt{\text{Var}\left[\log \frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)} \middle| S = 0 \right]}.
$$

We conclude this section with Example 1, where we compute $\widetilde{f}$ when $P_{S|X}$ follows a logistic distribution.

Example 1. When $P_{S|X}(1|\boldsymbol{x}) = (1 + \exp(\theta_0 + \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle))^{-1}$, the log-likelihood ratio function f_l is the linear function

$$f_l(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle - \sum_{i=1}^{d} \mathbb{E}\left[\theta_i X_i \Big| S = 0\right]. \quad (14)$$

Combining (10), (12) and (14), $\widetilde{f}$ can be expressed as a linear combination of $W_{Y|X}$, a linear function of $\boldsymbol{x}$, and a constant term. Further, when $P_{S|Y}(1|y) = (1 + \exp(\gamma_0 + \gamma_1 y))^{-1}$, $\Delta_{\boldsymbol{\lambda}}(f)$ can be expressed as

$$\Delta_{\boldsymbol{\lambda}}(f) = \lambda_2 \sum_{i=1}^{d} \theta_i \mathbb{E}\left[f(X) X_i | S = 0\right] + \lambda_4 \gamma_1 \mathbb{E}\left[g(Y) Y | S = 0\right].$$

for any perturbation f, including $\widetilde{f}$.

IV. NUMERICAL EXPERIMENTS

We now discuss a numerical experiment where we compute correction functions for a recidivism prediction model. We consider the ProPublica COMPAS dataset [6], which contains information on the criminal history and demographic makeup of prisoners in Brower County, Florida from 2013–2014. Our goal is to illustrate the technical feasibility of our approach on a real-world dataset, and to show that the correction function can be computed using standard predictive models for S and Y given X (i.e., without the need to compute the distribution $P_{S,X,Y}$). We provide code to reproduce our analysis at [25].

Setup

We restrict our analysis to individuals who are African American ($S = 0$) or Caucasian ($S = 1$). We process the raw dataset by dropping records with missing information and converting categorical variables to numerical values. Our final dataset contains 5278 records (3175 African American + 2103 Caucasian), where the record for individual i consists of a feature vector $\boldsymbol{x}_i = (Age, ChargeDegree, Sex, PriorCounts, LengthOfStay)$, and an outcome variable, set as $y_i = 1$ iff they are arrested for a crime within 2 years of release from prison.

We use the entire dataset to train two logistic regression models: (i) $W_{Y|X}$, which uses the features to predict the outcome; (ii) $P_{S|X}$, which uses the features to predict group membership. Although $W_{Y|X}$ does not use S as an input, it has significant disparate impact over S, assigning higher scores on average to African Americans compared to Caucasians ($\mathbb{E}[Y|S = 0] = 0.543$ vs. $\mathbb{E}[Y|S = 1] = 0.438$). Using $W_{Y|X}$ and $P_{S|X}$, we apply Theorems 1 and 2 to compute the correction function $\widetilde{f}$ for $\lambda_2 = \lambda_4 = 1$.

Results

In Fig. 2, we show how the correction function can identify features that contribute to the disparate impact of a predictive model. Here, we plot the conditional distribution of $\widetilde{f}$ for distinct values of *PriorCounts* and *LengthOfStay*. As shown, the

FEATURES	$\operatorname{argmax} \widetilde{f}(\boldsymbol{x})$	$\operatorname{argmin} \widetilde{f}(\boldsymbol{x})$	$\operatorname{argmin} |\widetilde{f}(\boldsymbol{x})|$
Age	> 45	< 25	25 to 45
ChargeDegree	Misdemeanor	Felony	Felony
Sex	Female	Male	Male
PriorCounts	0	> 3	1 to 3
LengthOfStay	< 3 Months	< Week	< Week

Table 1: Prototypical examples for the correction function $\widetilde{f}$. We show the features of African American entries that attain the maximum, minimum, and smallest values of $\widetilde{f}$.

distribution of $\widetilde{f}$ is similar across all values of *LengthOfStay*, which suggests that *LengthOfStay* does not affect the disparate impact of the model. In contrast, the distribution of $\widetilde{f}$ differs based on the value of *PriorCounts* (see e.g., the differences between *PriorCounts* > 3). This suggests that the model may be using *PriorCounts* to discriminate between African Americans and Caucasians.

In Table 1, we show prototypical examples for the correction function, which correspond to feature vectors for which the correction function $\widetilde{f}$ attains its maximum value/minimum value/value closest to 0. Recalling that the $\widetilde{f}$ represents a local perturbation that minimizes the objective in (4), we see that the disparity in the output distributions of $W_{Y|X}$ is maximal for African American males that are under 25 years old, with > 3 priors, and charged with a felony (middle column). This is to be expected, as the training dataset for $W_{Y|X}$ shows a large correlation between *PriorCounts* and the outcome, and higher *PriorCounts* for African Americans on average.
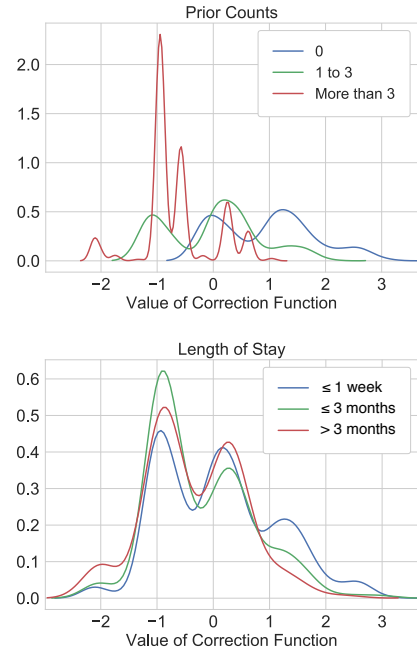

Fig. 2: Empirical distributions of the correction function $\widetilde{f}$ conditioned on distinct values of *PriorCounts* (top) and *LengthOfStay* (bottom). We use a kernel density estimator to smooth the histograms.

V. DISCUSSION

Disparate impact in machine learning is a critical issue with important societal implications. In this paper, we proposed an information-theoretic framework to study disparate impact. We derived a correction function that reflects how components of the input variables X affect the disparity in the output distributions. We then demonstrated how our framework could be used on a recidivism prediction application derived from a real-world dataset. Interesting directions for future work include extending our analysis to a broader class of predictive models, and using correction functions to design machine learning algorithms that mitigate disparate impact. We are confident that information-theoretic tools can inspire exciting new solutions to the problem.

REFERENCES

[1] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 582–638, 2014.

[2] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: from discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2125–2126.

[3] S. Barocas and A. Selbst, "Big data's disparate impact," 2016.

[4] D. B. Hunt, "Redlining," *Encyclopedia of Chicago*, 2005.

[5] L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, no. 3, p. 10, 2013.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016.

[7] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing," 2017.

[8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 259–268.

[9] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," in *Proceedings of ICDM 2016*. IEEE, 2016, pp. 1–10.

[10] J. Adebayo and L. Kagal, "Iterative orthogonal feature projection for diagnosing bias in black-box models," *arXiv preprint arXiv:1611.04967*, 2016.

[11] C. Simoiu, S. Corbett-Davies, and S. Goel, "The problem of infra-marginality in outcome tests for discrimination," Working paper available at https://arxiv. org/pdf/1607.05376. pdf, Tech. Rep., 2017.

[12] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 2009, pp. 1–6.

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 325–333.

[14] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3995–4004.

[15] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *arXiv preprint arXiv:1706.02744*, 2017.

[16] J. Liao, L. Sankar, V. Y. Tan, and F. P. Calmon, "Hypothesis testing in the high privacy limit," in *Proc. of 54th IEEE Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2016, pp. 649–656.

[17] T. M. Cover and J. A. Thomas, "Elements of information theory 2nd edition," 2006.

[18] P. J. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 1248–1251.

[19] S. Borade and L. Zheng, "Euclidean information theory," in *IEEE International Zurich Seminar on Communications*. IEEE, 2008, pp. 14–17.

[20] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. on Info. Theory*, vol. 20, no. 4, pp. 405–417, 1974.

[21] E. Tuncel, "On error exponents in hypothesis testing," *IEEE Trans. on Info. Theory*, vol. 51, no. 8, pp. 2945–2950, 2005.

[22] A. A. Gohari and V. Anantharam, "Evaluation of Marton's inner bound for the general broadcast channel," *IEEE Trans. on Info. Theory*, vol. 58, no. 2, pp. 608–619, 2012.

[23] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint arXiv:1304.6133*, 2013.

[24] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[25] "Disparate Impact Repository," *GitHub*, 2018. [Online]. Available: http://www.github.com/ustunb/disparate-impact

A. Proof of Lemma 1

Proof. First, note that we can compute the distribution $\widetilde{P}_{Y|S=0}$ by passing $\widetilde{P}_{X|S=0}$ through the given channel $W_{Y|X}$ and get the following expression.

$$
\begin{aligned}
\widetilde{P}_{Y|S=0}(y) &= \sum_{\boldsymbol{x}} W_{Y|X}(y|\boldsymbol{x})P_{X|S=0}(\boldsymbol{x})(1+\epsilon f(\boldsymbol{x})) \\
&= P_{Y|S=0}(y) + \epsilon \sum_{\boldsymbol{x}} W_{Y|X}(y|\boldsymbol{x})P_{X|S=0}(\boldsymbol{x})f(\boldsymbol{x}) \\
&= P_{Y|S=0}(y) + \epsilon \sum_{\boldsymbol{x}} P_{X|Y,S=0}(\boldsymbol{x}|y)P_{Y|S=0}(y)f(\boldsymbol{x}) \\
&= P_{Y|S=0}(y)\left(1+\epsilon \mathbb{E}\left[f(X)|Y=y,S=0\right]\right) \\
&= P_{Y|S=0}(y)(1+\epsilon g(y)).
\end{aligned}
$$

By the definition of $\widetilde{P}_{X|S=0}$, we can compute the KL-divergence between $\widetilde{P}_{X|S=0}$ and $P_{X|S=1}$ in the following way.

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\widetilde{P}_{X|S=0}\|P_{X|S=1}) \\
&= \sum_{\boldsymbol{x}} P_{X|S=0}(\boldsymbol{x})(1+\epsilon f(\boldsymbol{x}))\log\frac{P_{X|S=0}(\boldsymbol{x})(1+\epsilon f(\boldsymbol{x}))}{P_{X|S=1}(\boldsymbol{x})} \\
&= \mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1}) \\
&\quad + \epsilon \sum_{\boldsymbol{x}} P_{X|S=0}(\boldsymbol{x})f(\boldsymbol{x})\log\frac{P_{X|S=0}(\boldsymbol{x})}{P_{X|S=1}(\boldsymbol{x})} \\
&\quad + \sum_{\boldsymbol{x}} P_{X|S=0}(\boldsymbol{x})(1+\epsilon f(\boldsymbol{x}))\log(1+\epsilon f(\boldsymbol{x})) \\
&= \mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1}) \\
&\quad + \epsilon \mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right] \\
&\quad + \epsilon \sum_{\boldsymbol{x}} P_{X|S=0}(\boldsymbol{x})(1+\epsilon f(\boldsymbol{x}))f(\boldsymbol{x}) + O(\epsilon^2) \\
&= \mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1}) \\
&\quad + \epsilon \mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right] \\
&\quad + \epsilon \mathbb{E}\left[f(X)|S=0\right] + \epsilon^2 \mathbb{E}\left[f(X)^2|S=0\right] + O(\epsilon^2) \\
&= \mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1}) \\
&\quad + \epsilon \mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right] + O(\epsilon^2).
\end{aligned} \tag{15}
$$

Following similar computation, we have

$$
\mathrm{D}_{\mathrm{KL}}(\widetilde{P}_{X|S=0}\|P_{X|S=0}) = O(\epsilon^2). \tag{16}
$$

Since $\mathbb{E}\left[g(Y)|S=0\right]=0$, we have

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\widetilde{P}_{Y|S=0}\|P_{Y|S=1}) - \mathrm{D}_{\mathrm{KL}}(P_{Y|S=0}\|P_{Y|S=1}) \\
&= \epsilon \mathbb{E}\left[g(Y)\log\frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)}\middle| S=0\right] + O(\epsilon^2).
\end{aligned} \tag{17}
$$

Also,

$$
\mathrm{D}_{\mathrm{KL}}(\widetilde{P}_{Y|S=0}\|P_{Y|S=0}) = O(\epsilon^2). \tag{18}
$$

Combining (15), (16), (17), (18) together and letting $\epsilon \to 0$, we obtain

$$
\begin{aligned}
\Delta_{\boldsymbol{\lambda}}(f) =&\lambda_2 \mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right] \\
&+ \lambda_4 \mathbb{E}\left[g(Y)\log\frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)}\middle| S=0\right].
\end{aligned}
$$

$\square$

B. Proof of Theorem 1

Proof. Note that when Y is a binary random variable, for any function $g(y)$ with $\mathbb{E}\left[g(Y)\right]=0$, we have that $g(Y) = \mathbb{E}\left[g(Y)g_m(Y)\right]g_m(Y)$. Furthermore, for any function $f(\boldsymbol{x})$ with $\mathbb{E}\left[f(X)\right]=0$, if $\mathbb{E}\left[f(X)f_m(X)\right]=0$, then $\mathbb{E}\left[f(X)|Y\right]=0$.

We define

$$
f_L(\boldsymbol{x}) \triangleq \frac{f_l(\boldsymbol{x}) - a_1 f_m(\boldsymbol{x})}{a_2}, \tag{19}
$$

when $a_2 \neq 0$. When $a_2 = 0$, we can choose an arbitrary function f_L such that $f_L \in \mathcal{L}(P_{X|S=0})$ and $\mathbb{E}\left[f_L(X)f_m(X)|S=0\right]=0$. Note that $f_L \in \mathcal{L}(P_{X|S=0})$ from the definition. Furthermore, we have $g_l(y)=b_1 g_m(y)$ with $b_1 = \mathbb{E}\left[g_l(Y)g_m(Y)|S=0\right]$. Similarly,

$$
f(\boldsymbol{x}) = m_1 f_m(\boldsymbol{x}) + m_2 f_L(\boldsymbol{x}) + m_3 f_r(\boldsymbol{x}) \tag{20}
$$

where,

$$
\begin{aligned}
m_1 &= \mathbb{E}\left[f(X)f_m(X)|S=0\right], \\
m_2 &= \mathbb{E}\left[f(X)f_L(X)|S=0\right], \\
m_3 &= \sqrt{\mathbb{E}\left[\left(f(X)-m_1 f_m(X)-m_2 f_L(X)\right)^2\middle| S=0\right]}.
\end{aligned}
$$

$$
f_r(\boldsymbol{x}) \triangleq \frac{f(\boldsymbol{x}) - m_1 f_m(\boldsymbol{x}) - m_2 f_L(\boldsymbol{x})}{m_3},
$$

when $m_3 \neq 0$. When $m_3 = 0$, we can choose an arbitrary function f_r such that $f_r \in \mathcal{L}(P_{X|S=0})$ and $\mathbb{E}\left[f_r(X)f_m(X)|S=0\right] = \mathbb{E}\left[f_r(X)f_L(X)|S=0\right]=0$. Note that $f_r \in \mathcal{L}(P_{X|S=0})$ following the definition. Since, by the definition of f_L and f_r, $\mathbb{E}\left[f_m(X)f_L(X)|S=0\right]=0$ and $\mathbb{E}\left[f_m(X)f_r(X)|S=0\right]=0$, then

$$
\begin{aligned}
g(y) &= \mathbb{E}\left[f(X)|Y=y,S=0\right] \\
&= \rho_m(P_{X|S=0};W_{Y|X})m_1 g_m(y).
\end{aligned}
$$

Therefore, following previous discussions and using Lemma 1, we have

$$\Delta_{\boldsymbol{\lambda}}(f)$$

$$= \lambda_2 \mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right]$$

$$+ \lambda_4 \mathbb{E}\left[g(Y)\log\frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)}\middle| S=0\right]$$

$$= \lambda_2 \mathbb{E}\left[f(X)f_l(X)\middle| S=0\right] + \lambda_4 \mathbb{E}\left[g(Y)g_l(Y)\middle| S=0\right]$$

$$= \lambda_2 a_1 m_1 + \lambda_2 a_2 m_2 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1 m_1.$$

Since $f \in \mathcal{L}(P_{X|S=0})$, we have that $m_1^2 + m_2^2 + m_3^2 = 1$. Accordingly, we can minimize $\Delta_{\boldsymbol{\lambda}}(f)$ by solving the optimization problem:

$$\min_{m_1, m_2} \quad \left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right) m_1 + \lambda_2 a_2 m_2$$

$$\text{s.t.} \quad m_1^2 + m_2^2 + m_3^2 = 1.$$

By the Cauchy-Schwarz inequality, the minimal value is $-\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}$ which is achieved by setting

$$m_1 = \frac{-\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}},$$

$$m_2 = \frac{-\lambda_2 a_2}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}}$$

$$m_3 = 0.$$

Therefore, the function f, which achieves this minimal value, is $n_m f_m + n_l f_l$ where

$$n_m = \frac{-\lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}}$$

and

$$n_l = \frac{-\lambda_2}{\sqrt{\left(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X})b_1\right)^2 + (\lambda_2 a_2)^2}}.$$

$\square$

C. Proof of Theorem 2

Proof. Suppose that $g_m(0) = a \geq 0$ and $g_m(1) = b$. Then $\mathbb{E}[g_m(Y)|S=0] = a(1-p) + bp = 0$ which implies that $b = \frac{-(1-p)}{p}a$. Since $\mathbb{E}[g_m(Y)^2|S=0] = a^2(1-p) + b^2 p = a^2(1-p) + a^2 \frac{(1-p)^2}{p} = 1$, then $a = \sqrt{\frac{p}{1-p}}$.

Next,

$$f_m(\boldsymbol{x})$$

$$= \frac{\mathbb{E}[g_m(Y)|X=\boldsymbol{x}, S=0]}{\rho_m(P_{X|S=0}; W_{Y|X})}$$

$$= \frac{g_m(1)W_{Y|X}(1|\boldsymbol{x}) + g_m(0)W_{Y|X}(0|\boldsymbol{x})}{\rho_m(P_{X|S=0}; W_{Y|X})}$$

$$= \frac{g_m(1) - g_m(0)}{\rho_m(P_{X|S=0}; W_{Y|X})}W_{Y|X}(1|\boldsymbol{x}) + \frac{g_m(0)}{\rho_m(P_{X|S=0}; W_{Y|X})}.$$

Note that

$$\log\frac{P_{X|S=0}(\boldsymbol{x})}{P_{X|S=1}(\boldsymbol{x})} = \log\frac{P_{S|X}(0|\boldsymbol{x})P_S(1)}{P_{S|X}(1|\boldsymbol{x})P_S(0)}$$

$$= \log\frac{P_{S|X}(0|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} + \log\frac{P_S(1)}{P_S(0)}.$$

$$f_l(\boldsymbol{x}) = \log\frac{P_{X|S=0}(\boldsymbol{x})}{P_{X|S=1}(\boldsymbol{x})} - \mathbb{E}\left[\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right]$$

$$= \log\frac{P_{S|X}(0|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log\frac{P_{S|X}(0|X)}{P_{S|X}(1|X)}\middle| S=0\right]$$

$$= \log\frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log\frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle| S=0\right].$$

$\square$

D. Details for Example 1

When $P_{S|X}(1|\boldsymbol{x}) = (1 + \exp(\theta_0 + \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle)^{-1}$, then

$$f_l(\boldsymbol{x}) = \log\frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log\frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle| S=0\right]$$

$$= \theta_0 + \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle - \mathbb{E}\left[\theta_0 + \langle\boldsymbol{\theta}, X\rangle\middle| S=0\right]$$

$$= \langle\boldsymbol{\theta}, \boldsymbol{x}\rangle - \sum_{i=1}^{d}\mathbb{E}\left[\theta_i X_i\middle| S=0\right].$$

Next,

$$\mathbb{E}\left[f(X)\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle| S=0\right]$$

$$= \mathbb{E}\left[f(X)\log\frac{P_{S|X}(0|X)}{P_{S|X}(1|X)}\middle| S=0\right]$$

$$+ \mathbb{E}\left[f(X)\log\frac{P_S(1)}{P_S(0)}\middle| S=0\right]$$

$$= \mathbb{E}\left[f(X)(\theta_0 + \langle\boldsymbol{\theta}, X\rangle)|S=0\right]$$

$$= \theta_0\mathbb{E}[f(X)|S=0] + \sum_{i=1}^{d}\theta_i\mathbb{E}[f(X)X_i|S=0]$$

$$= \sum_{i=1}^{d}\theta_i\mathbb{E}[f(X)X_i|S=0].$$

Similarly, if we also assume that $P_{S|Y}(1|y) = (1 + \exp(\gamma_0 + \gamma_1 y))^{-1}$, then

$$\mathbb{E}\left[g(Y)\log\frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)}\middle| S=0\right] = \gamma_1\mathbb{E}[g(Y)Y|S=0].$$

Thus, we can express,

$$\Delta_{\boldsymbol{\lambda}}(f)$$

$$= \lambda_2\sum_{i=1}^{d}\theta_i\mathbb{E}[f(X)X_i|S=0] + \lambda_4\gamma_1\mathbb{E}[g(Y)Y|S=0].$$