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 present 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. We then aim to find a correction function that can be used to perturb the input distributions of each group in order 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 for the correction function that can be used to compute it efficiently. We illustrate the use of the correction function for a recidivism prediction application derived from the ProPublica COMPAS dataset.

I. INTRODUCTION

Machine learning (ML) models aim to exploit biases in the training data to predict an outcome variable 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 manifests itself in contexts outside of ML. In the United States, for example, racial minorities were more likely to be charged with a crime, sentenced to longer terms, and released from prison under supervision, even when controlling for various socioeconomic variables such as address, age, and income. This phenomenon is known as redlining [4].

Disparate impact is likely to arise as an artifact of empirical loss minimization when a sensitive attribute is valuable for the purposes of prediction, and has 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 not only violate anti-discrimination laws [3], but inadvertently amplify societal biases [7].

Such 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., $Y = 1$ if 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. We say that a channel $W_{Y|X}$ has 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 credit scoring (see, e.g., [3]). In this setting, the Markov condition $S \rightarrow X \rightarrow Y$ 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 when $P_X|S=0 \neq P_X|S=1$.

Given a fixed channel $W_Y|X$, disparate impact can be reduced by perturbing $P_X|S=0$ into a new distribution $Q_X$ such that the corresponding 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 $Q$ and $P_X|S=0$ identifies which components of $X$ contribute to this disparity. We formally define this setup next.

**Definition 1.** Let $L\left(\cdot \mid \cdot\right)$ be a convex divergence measure (e.g. total variation, KL-divergence), and $P_{X,S,Y}$ be fixed, where $S \rightarrow X \rightarrow Y$. For $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ with $\lambda_1 \geq 0$, we define the objective function as follows.

\[
L_\lambda(Q_X) \triangleq \lambda_1 J(Q_X \mid P_X|S=0) + \lambda_2 J(Q_X \mid P_X|S=1) + \lambda_3 J(Q_Y \mid P_Y|S=0) + \lambda_4 J(Q_Y \mid P_Y|S=1)
\]

where $Q_Y \triangleq W_Y|X \circ Q_X$.

The correction path between $P_X|S=0$ and $P_X|S=1$ is given by the solution of the optimization problem

\[
P^\text{opt}_\lambda \triangleq \min_{Q_X} L_\lambda(Q_X).
\]

From standard convexity results, e.g. [17], as the values of $\lambda_i$ are changed, the distribution $Q_X$ which achieves the minimal value given in (2) describes the lower boundary of the set

\[
\{ (J(Q_X \mid P_X|S=0), J(Q_X \mid P_X|S=1), J(Q_Y \mid P_Y|S=0), J(Q_Y \mid P_Y|S=1)) \mid \text{supp}(Q_X) \subseteq \mathcal{X} \}.
\]

The solutions of (2) form different paths on the probability simplex, as illustrated in Fig. 1. For $\lambda_2 = \lambda_4 = 0$, 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 $J(\cdot \mid \cdot)$) between $P_X|S=0$ and the set \{ $P_X \mid J(W_Y|X \circ P_X \mid P_Y|S=1) = 0$ \}, corresponding to 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 delineated by setting $\lambda_2 = 0$ and varying $\lambda_1$, $\lambda_2$, and $\lambda_4$. This corresponds to the red path in Fig. 1, 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 $X$ are contributing to the observed disparate impact. More precisely, the correction function is a local (multiplicative) perturbation of $P_X|S=0$ (see Definition 2) that decreases the objective function (1) the most. This definition leads to correction functions that can be cast in terms of predictive models for $S$ and $Y$ given $X$, as illustrated in Section III.

**Definition 2.** For a given function $f \in \mathcal{L}(P_X|S=0)$, we define the perturbed distribution $\tilde{P}_X|S=0$ as

\[
\tilde{P}_X|S=0(x) \triangleq P_X|S=0(x)(1 + \epsilon f(x)),
\]

where $\epsilon > 0$ is sufficiently small so that $\tilde{P}_X|S=0$ is a valid probability distribution, and

\[
\mathcal{L}(P_X|S=0) \triangleq \{ f : \mathcal{X} \rightarrow \mathbb{R} \mid E[f(X)|S=0] = 0, E[f(X)^2|S=0] = 1 \}.
\]

Moreover, $\tilde{P}_Y|S=0 \triangleq W_Y|X \circ \tilde{P}_X|S=0$.

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

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

\[
\Delta_\lambda(f) \triangleq \lim_{\epsilon \to 0^+} \frac{L_\lambda(\tilde{P}_X|S=0) - L_\lambda(P_X|S=0)}{\epsilon}.
\]

The function $f$ which achieves the minimal value of $\Delta_\lambda(f)$ is defined as the correction function next.

**Definition 4.** For a given $\lambda$, the correction function $\tilde{f}$ is defined as

\[
\tilde{f} = \arg\min_{f \in \mathcal{L}(P_X|S=0)} \Delta_\lambda(f).
\]

**Connections to Binary Hypothesis Testing**

In the remainder of this paper, we consider $Y$ binary and KL-divergence as the divergence measure, i.e., $J(\cdot \mid \cdot) = D_{\text{KL}}(\cdot \mid \cdot)$. In this case, the objective function (1) can be expressed as:

\[
L_\lambda(Q_X) = \lambda_1 D_{\text{KL}}(Q_X \mid P_X|S=0) + \lambda_2 D_{\text{KL}}(Q_X \mid P_X|S=1) + \lambda_3 D_{\text{KL}}(Q_Y \mid P_Y|S=0) + \lambda_4 D_{\text{KL}}(Q_Y \mid P_Y|S=1).
\]

Our choice of $D_{\text{KL}}(\cdot \mid \cdot)$ is motivated by its relationship with the error exponent in hypothesis testing ([17, Chap. 11], [18, 19]). 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. [19, Section II-A]), i.e., $D_\alpha(P_X|S=0 \parallel P_X|S=1)$. 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 under the objective given in (6). We leverage the connection between small perturbations of KL-divergences and maximal correlation (noted, for example, in [20, 21]).

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: $f_1$, 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 $f_1$ and $f_m$ (and thus $\tilde{f}$) 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 $\tilde{f}$ 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 $\Delta_\lambda(f)$ in the following lemma.

**Lemma 1.** For a given $\lambda$, $\Delta_\lambda(f)$ can be simplified as

$$\Delta_\lambda(f) = \lambda_2 \mathbb{E} \left[ f(X) \log \frac{P_X|S=0}{P_X|S=1} | S = 0 \right] + \lambda_3 \mathbb{E} \left[ g(Y) \log \frac{P_Y|S=0}{P_Y|S=1} | S = 0 \right],$$  \hspace{1cm} (7)

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

**Proof.** See Appendix A-B.

We now present definitions of log-likelihood ratio functions and principal functions that will be used to derive the correction function (see, e.g., [22] for further details on maximal correlation).

**Definition 5.** The log-likelihood ratio functions $f_1$ and $g_1$ are given by

$$f_1(x) \triangleq \log \frac{P_X|S=0(x)}{P_X|S=1(x)} - \mathbb{E} \left[ \log \frac{P_X|S=0}{P_X|S=1} | S = 0 \right],$$  \hspace{1cm} (8)

$$g_1(y) \triangleq \log \frac{P_Y|S=0(y)}{P_Y|S=1(y)} - \mathbb{E} \left[ \log \frac{P_Y|S=0}{P_Y|S=1} | S = 0 \right].$$  \hspace{1cm} (9)

**Definition 6.** The maximal correlation between $X$ and $Y$ given $S = 0$ is defined as

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

We refer to the functions $(f_m, g_m)$ that attain the maximum as the principal functions. The maximal correlation can be equivalently given by

$$\rho_m(P_X|S=0; W_{Y|X}) = \sqrt{\mathbb{E} \left[ g_m(Y) | X, S = 0 \right]^2 | S = 0}.$$  \hspace{1cm} (10)

Theorems 1 and 2 characterize the correction function $\tilde{f}$.

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

$$\tilde{f} = n_t f_1 + n_m f_m,$$  \hspace{1cm} (11)

where $n_t$ and $n_m$ are constants computed as:

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

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

where

$$a_1 = \mathbb{E} [f_1(X)f_m(X)|S=0],$$

$$a_2 = \mathbb{E} [f_1(X)g_m(Y)|S=0],$$

$$b_1 = \mathbb{E} [g_1(Y)g_m(Y)|S=0].$$

**Proof.** See Appendix A-B.

**Theorem 2.** The log-likelihood ratio function $f_1$ and the principal function $f_m$ can be expressed as

$$f_1(x) = \log \frac{P_{S|X}(0|x)}{P_{S|X}(1|x)} - \mathbb{E} \left[ \log \frac{P_{S|X}(0)}{P_{S|X}(1)} | X \right],$$

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

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

**Proof.** See Appendix A-C.

Combining Theorems 1 and 2, we obtain a closed-form expression for the correction function $\tilde{f}$. Note that, due to our definition of $\tilde{f}$ in terms of local perturbations, the correction function does not depend on $\lambda_1$ and $\lambda_3$ in (6). Corollary 1 states the corresponding value of $\Delta_\lambda(\tilde{f})$.

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

$$\Delta_\lambda(\tilde{f}) = \frac{\lambda_2 a_1 + \lambda_4 \rho_m(P_X|S=0; W_{Y|X}) b_1}{\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_X|S=0; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}},$$

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

We instantiate our results for $\lambda_2 = 0$ next. This corresponds to the green line in Fig. 1, where our main objective is to align only the output distributions. Unsurprisingly, the following
within 2 years of release from prison.

Corollary 2. When $\lambda_2 = 0$, the correction function $\tilde{f}$ has the form

$$\tilde{f} = -\text{sign} \left( \mathbb{E} [g(Y) g_m(Y) | S = 0] \right) f_m$$

and

$$\Delta_\lambda(f) = -\lambda_4 \rho_m(P_X|S=0; W_Y|X) \sqrt{\text{Var} \left[ \log \frac{P_Y|S=0(Y)}{P_Y|S=1(Y)} \right] \mid S = 0}.$$

We conclude this section with Example 1, where we compute $\tilde{f}$ when $P_{S|X}$ can be expressed by a logistic regression.

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

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

Combining (10), (12) and (14), $\tilde{f}$ can be expressed in terms of a simple linear combination of $W_{Y|X}$, a linear function of $x$, and a constant term. In addition, when $P_{S|Y}(1|y) = (1 + \exp(\gamma_0 + \gamma_1 y))^{-1}$, $\Delta_\lambda(f)$ can be simplified for any perturbation $f$ (including $\tilde{f}$) as

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

See Appendix A-D for details.

IV. NUMERICAL EXPERIMENTS

We discuss next 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 make no comment on associated societal interpretations. Code to reproduce our analysis can be found in [23].

Setup

We restrict our analysis to individuals who are African American ($S = 0$) or Caucasian ($S = 1$). We mildly 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 $x_i = (\text{Age}, \text{ChargeDegree}, \text{Sex}, \text{PriorCounts}, \text{LengthOfStay})$, and an outcome variable, set as $y_i = 1$ if he/she is arrested for a crime within 2 years of release from prison.

We use the full 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, 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 $\tilde{f}$ 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 $\tilde{f}$ for distinct values of PriorCounts and LengthOfStay. As shown, the distribution of $\tilde{f}$ 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 $\tilde{f}$ 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 $\tilde{f}$ attains its maximum value/minimum value/value closest to 0. Recalling that the $f$ 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).
is to be expected, as the training dataset for $W_{Y|X}$ shows a large correlation between \( \text{PriorCounts} \) and the outcome, and higher \( \text{PriorCounts} \) for African Americans on average.

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 to link components of \( X \) to the disparity observed in the output distributions, and illustrated its value 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.

ACKNOWLEDGMENTS

F.P. Calmon would like to thank the Harvard Dean’s Competitive Fund for Promising Scholarship for supporting this research.

REFERENCES

APPENDIX A
Proofs

A. Proof of Lemma 1

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

\[
\tilde{P}_{Y|S=0}(y) = \sum_x W_{Y|X}(y|x)P_{X|S=0}(x)(1 + \epsilon f(x)) = P_{Y|S=0}(y) + \epsilon \sum_x W_{Y|X}(y|x)P_{X|S=0}(x)f(x)
\]

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

\[
D_{KL}(\tilde{P}_{X|S=0}||P_{X|S=1}) = \sum_x P_{X|S=0}(x)(1 + \epsilon f(x)) \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)}
\]

\[
+ \epsilon \sum_x P_{X|S=0}(x) f(x) \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)} + \epsilon \sum_x P_{X|S=0}(x)(1 + \epsilon f(x)) \log(1 + \epsilon f(x))
\]

\[
= \sum_x \epsilon f(x) \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)}
\]

\[
+ \epsilon \sum_x P_{X|S=0}(x)(1 + \epsilon f(x)) f(x) + O(\epsilon^2)
\]

\[
= \sum_x \epsilon f(x) \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)}
\]

\[
+ \epsilon \sum_x P_{X|S=0}(x)(1 + \epsilon f(x)) f(x) + O(\epsilon^2)
\]

Following similar computation, we have

\[
D_{KL}(\tilde{P}_{X|S=0}||P_{X|S=1}) = O(\epsilon^2).
\]

Since \( E[g(Y)|S=0] = 0 \), we have

\[
D_{KL}(\tilde{P}_{Y|S=0}||P_{Y|S=0}) - D_{KL}(P_{Y|S=0}||P_{Y|S=1}) = \epsilon E[g(Y) \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)}|S=0] + O(\epsilon^2).
\]

Also,

\[
D_{KL}(\tilde{P}_{Y|S=0}||P_{Y|S=0}) = O(\epsilon^2).
\]

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

\[
\Delta_{\lambda}(f) = \lambda \sum_x f(x) \log \frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}|S=0
\]

\[
+ \lambda \sum_x g(Y) \log \frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)}|S=0.
\]

B. Proof of Theorem 1

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

We define

\[
f_L(x) = f(x) - a_1 f_m(x),
\]

where

\[
m_1 = E[f(X)g_m(X)|S=0],
\]

\[
m_2 = E[f(X)g_m(X)|S=0],
\]

\[
m_3 = \sqrt{E[(f(X) - m_1 f_m(X) - m_2 f_L(X))^2]|S=0].
\]

\[
f_r(x) = f(x) - m_1 f_m(x) - m_2 f_L(x),
\]

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

\[
g(y) = E[f(X)|Y = y, S=0] = \rho_m(P_{X|S=0};W_{Y|X})m_1 g_m(y).
Therefore, following previous discussions and using Lemma 1, we have
\[
\Delta_{\lambda}(f) = \lambda_2 \mathbb{E} \left[ f(X) \log \frac{P_{X|S=0}(X)}{P_{X|S=1}(X)} \right]_{S = 0} + \lambda_4 \mathbb{E} \left[ g(Y) \log \frac{P_{Y|S=0}(Y)}{P_{Y|S=1}(Y)} \right]_{S = 0} + \lambda_2 a_1 m_1 + \lambda_2 a_2 m_2 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1 m_1.
\]
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 \( \Delta_{\lambda}(f) \) by solving the optimization problem:
\[
\min_{m_1, m_2} \quad (\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1) m_1 + \lambda_2 a_2 m_2
\]
\[\text{s.t.} \quad m_1^2 + m_2^2 + m_3^2 = 1.
\]
By the Cauchy-Schwarz inequality, the minimal value is
\[
-\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}
\]
which is achieved by setting
\[
m_1 = \frac{-\lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1}{\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}},
\]
\[
m_2 = \frac{-\lambda_2 a_2}{\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}},
\]
\[
m_3 = 0.
\]
Therefore, the function \( f \), which achieves this minimal value, is
\[
n_m f_m + n_i f_i
\]
where
\[
n_m = \frac{-\lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1}{\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}}.
\]
and
\[
n_i = \frac{-\lambda_2 a_2}{\sqrt{(\lambda_2 a_1 + \lambda_4 \rho_m(P_{X|S=0}; W_{Y|X}) b_1)^2 + (\lambda_2 a_2)^2}}.
\]
\[\square\]

C. Proof of Theorem 2

Proof. Suppose that \( g_m(0) = a \geq 0 \) and \( g_m(1) = b \). Then
\[
\mathbb{E} \left[ g_m(Y) | S = 0 \right] = a(1 - p) + b p = 0 \quad \text{which implies that} \quad b = \frac{(1-p) a}{p}.
\]
Since \( g_m(Y)^2 | S = 0 \right] = a^2 (1 - p) + b^2 p = a^2 (1 - p) + a^2 \frac{(1-p)^2}{p} = 1, \text{ then} a = \sqrt{\frac{p}{1-p}}.
\]
Next,
\[
f_m(x) = \mathbb{E} \left[ g_m(Y) | X = x, S = 0 \right] = \rho_m(P_{X|S=0}; W_{Y|X}) g_m(1) W_{Y|X}(1|x) + g_m(0) W_{Y|X}(0|x)
\]
\[
= \rho_m(P_{X|S=0}; W_{Y|X}) g_m(0) W_{Y|X}(1|x) + \rho_m(P_{X|S=0}; W_{Y|X}) g_m(0)
\]
\[\square\]

D. Details for Example 1

When \( P_{S|X}(1|x) = (1 + \exp(\theta_0 + \langle \theta, x \rangle))^{-1} \), then
\[
f_i(x) = \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)} - \mathbb{E} \left[ \log \frac{P_{X|S=0}(x)}{P_{X|S=1}(x)} \right]_{S = 0}
\]
\[\square\]

E. Proof of Lemma 1

Proof. Suppose that \( g_m(0) = a \geq 0 \) and \( g_m(1) = b \). Then
\[
\mathbb{E} \left[ g_m(Y) | S = 0 \right] = a(1 - p) + b p = 0 \quad \text{which implies that} \quad b = \frac{(1-p) a}{p}.
\]
Since \( g_m(Y)^2 | S = 0 \right] = a^2 (1 - p) + b^2 p = a^2 (1 - p) + a^2 \frac{(1-p)^2}{p} = 1, \text{ then} a = \sqrt{\frac{p}{1-p}}.
\]
Next,
\[
f_m(x) = \mathbb{E} \left[ g_m(Y) | X = x, S = 0 \right] = \rho_m(P_{X|S=0}; W_{Y|X}) g_m(1) W_{Y|X}(1|x) + g_m(0) W_{Y|X}(0|x)
\]
\[
= \rho_m(P_{X|S=0}; W_{Y|X}) g_m(0) W_{Y|X}(1|x) + \rho_m(P_{X|S=0}; W_{Y|X}) g_m(0)
\]
\[\square\]