Avoiding Discrimination with Counterfactual Distributions

Hao Wang HAO_WANG@G.HARVARD.EDU

Berk Ustun BERK@SEAS.HARVARD.EDU

Flavio P. Calmon FLAVIO@SEAS.HARVARD.EDU

John A. Paulson School of Engineering and Applied Sciences
Harvard University

Abstract

When a classification model is used to make predictions on individuals, it may be undesirable or illegal for the performance of the model to change with respect to a sensitive attribute (e.g., race or gender). In this paper, we describe a distributional paradigm to evaluate and mitigate such disparities in model performance. Given a black-box classifier that performs unevenly across groups, we consider a *counterfactual distribution* of input variables that will eliminate the performance gap. We characterize properties of counterfactual distributions for common fairness criteria. We then present an information-theoretic approach to efficiently recover counterfactual distributions for a black-box classifier given a sample of points from its target population. We describe how counterfactual distributions can be used to avoid discrimination between protected groups by (i) identifying proxy variables to omit in training and (ii) building a preprocessor that can mitigate discrimination. We validate these use cases and illustrate their benefits through experiments on real-world datasets.

1. Introduction

A machine learning model has *disparate impact* when its performance changes across groups defined by *sensitive (protected) attributes*, such as race and gender. Recent work has shown that a large class of models can have a significant performance variation between protected groups even when they do not use sensitive attributes as input features. For example, facial recognition models have been reported to consistently misclassify African-American women (Buolamwini and Gebru, 2018), and recidivism prediction instruments may, on average, assign higher risk scores to young minority males (Angwin et al., 2016).

These disparities have motivated a plethora of research on fairness in machine learning, focusing on how disparate impact arises (Chen et al., 2018), how it can be measured (Žliobaitė, 2017; Pierson et al., 2017; Simoiu et al., 2017; Kilbertus et al., 2017; Kusner et al., 2017; Galhotra et al., 2017), and how it can be mitigated (Feldman et al., 2015; Corbett-Davies et al., 2017; Zafar et al., 2017b; Calmon et al., 2017). In spite of these contributions, disparate impact remains difficult to avoid in many real-world applications. This is because:

- Models may be deployed on a population that does reflect the patterns contained in the training data (Sugiyama et al., 2017).
- Models may *not* be developed in-house, but procured from a third-party vendors who have the required domain and technical expertise.

These challenges are becoming increasingly pervasive due to the adoption of machine learning in new areas and a growing number of companies selling proprietary models (Diakopoulos, 2014). In such settings, disparate impact is difficult to understand, let alone mitigate: users typically have black-box access to the classifier (e.g., via a prediction API); may not have access to the training data (due to privacy or intellectual property issues); may not be able to draw conclusions from the training data (due to dataset shift).

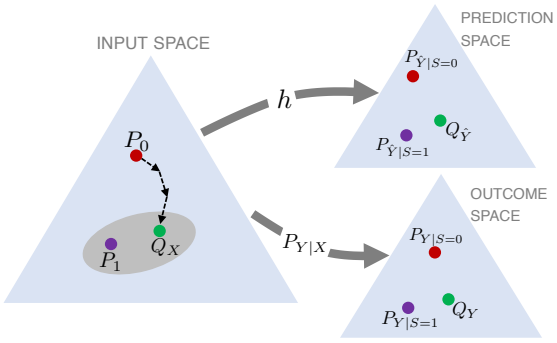

Figure 1: Illustration of disparate impact on the probability simplex for a fixed (black-box) classifier h. P_0 and P_1 denote the distribution of input variables for protected groups ($S = 0$ or 1). Disparate impact arises due to differences in the distribution of predicted outcomes ($P_{\hat{Y}|S=0}$ vs. $P_{\hat{Y}|S=1}$) and true outcomes ($P_{Y|S=0}$ vs. $P_{Y|S=1}$). A *counterfactual distribution* Q_X is a perturbation of P_0 that minimizes a measure of disparity. If the disparate impact persists under Q_X, there may be irreconcilable differences between the groups (i.e., $P_{Y|X,S=0} \neq P_{Y|X,S=1}$, see Prop. 2). The counterfactual distribution may not be unique, as illustrated by the shaded ellipse.

In this paper, we use tools from information theory and robust statistics to build a framework for evaluating and mitigating disparate impact in this challenging setting. Our object of interest is a hypothetical distribution of input variables that minimizes disparate impact in a population of interest (i.e., the *target population*). We refer to this distribution as a *counterfactual distribution*, as an analog to the counterfactual explanations of (Wachter and Mittelstadt, 2018).

As we will show, a formal study of counterfactual distributions has much to offer. Once a classifier is fixed, disparate impact can be traced back to differences between the distributions of input and output variables across protected groups (cf. Figure 1). Informally, a counterfactual distribution can be found by continuously perturbing the distribution of input features over a target population until a given discrimination metric is minimized. The resulting counterfactual distribution reveals insights on the sources of disparate impact that are tailored to that target population. Naturally, the counterfactual distribution also informs pre-processing methods for reducing disparity without requiring training of a new model.

The main contributions of this paper are:

1. We introduce the counterfactual distribution framework to evaluate and mitigate performance disparities in classification models.

2. We design a practical procedure to recover counterfactual distributions for common fairness criteria given (i) a black-box classifier and (ii) a sample of data from the target population on which it will be deployed (Section 3). Our tools recover a counterfactual distribution using a descent procedure in the simplex of probability distributions (Algorithm 1). We prove that influence functions can be used to compute a gradient in this setting (Proposition 3), and derive closed-form estimators that enable efficient computation (Proposition 5,6).

3. We validate our procedure by recovering counterfactual distributions for classifiers trained with real-world datasets (Section 4 and Appendix D). We use our results to describe ways in which counterfactual distributions can help us avoid disparate impact in real-world applications, namely by building a data preprocessor for the target population and explaining discrimination by proxy through contrastive analyses.

The approach in (Datta et al., 2016) and (Feldman et al., 2015) is related to ours, in the sense that they also propose proxy identification and pre-processing methods for mitigating the disparate

impact of black box models. We believe counterfactual distributions provide an alternative set of tools to evaluate and mitigate discrimination, offering new insights into disparate impact by leveraging tools from information theory and robust statistics. We provide a more detailed discussion of related work in Section 5.

2. Framework

In this section, we formally define counterfactual distributions and their properties.

Preliminaries

We consider a standard classification task where the goal is to predict a binary label $Y \in \{0, 1\}$ using a vector of d random variables $X = (X_1, \ldots, X_d) \in \mathcal{X}$ with distribution P_X. We assume a fixed classifier $h : \mathcal{X} \to [0, 1]$ where:

- $h(\boldsymbol{x}) \in \{0, 1\}$ if the classifier outputs a predicted label (e.g., SVM);
- $h(\boldsymbol{x}) \in [0, 1]$ if the classifier outputs a predicted probability (e.g., logistic regression).

We aim to characterize differences in the performance of the classifier with respect to a *sensitive attribute* $S \in \{0, 1\}$. We assume that the sensitive attribute S is not used as an input to the classifier (as this would violate legal constraints in applications such as hiring and credit scoring, Barocas and Selbst, 2016). We refer to individuals where $S = 0$ as the *minority* group, and individuals where $S = 1$ as the *majority* group. We denote the distributions of input variables for the minority and majority groups as $P_0 \triangleq P_{X|S=0}$ and $P_1 \triangleq P_{X|S=1}$, respectively. Likewise, we let $P_{\hat{Y}|S=0}(1) \triangleq \mathbb{E}\left[h(X)|S = 0\right]$, $P_{\hat{Y}|S=1}(1) \triangleq \mathbb{E}\left[h(X)|S = 1\right]$.

Discrimination Metrics

We measure the performance disparity between groups in terms of a *discrimination metric*. We use $\mathsf{M}(P_0, P_1)$ to denote these metrics since they can be expressed in terms of P_0 and P_1 (see Appendix A). We present examples of discrimination metrics for common fairness criteria in Table 1. Our framework can be generalized to other metrics in the literature (see e.g., Romei and Ruggieri, 2014; Žliobaitė, 2017, for a list) by adapting the derivations in Appendix C or using linear combinations of these metrics (see Section 3).

Definition 1. Given the classification model h and distributions $P_{Y|X,S}$ and P_S, a discrimination metric is a mapping $\mathsf{M} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$.

While it is possible to reduce disparate impact by simultaneously perturbing the input distributions for the minority group (P_0) and the majority group, we restrict our attention to changes that affect the minority group. This is due to an implicit assumption that the minority group achieves less favorable performance, and that we would rather address a performance disparate by having the minority group perform better rather than the majority group perform worse. For the metrics in Table 1, this implies $\mathsf{M}(P_0, P_1) \geq 0$ for all P_0, P_1.

Counterfactual Distributions

A *counterfactual distribution* is a hypothetical probability distribution of input variables for the minority group that minimizes a specific discrimination metric.

Definition 2. For a given discrimination metric $\mathsf{M}(P_0, P_1)$, a counterfactual distribution is probability distribution of input variables X such that:

$$Q_X \in \underset{Q'_X \in \mathcal{P}}{\operatorname{argmin}} \left|\mathsf{M}(Q'_X, P_1)\right|, \tag{1}$$

Performance Metric	Discrimination Metric
Distribution Alignment (DA)	$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}\|S=0}\|P_{\hat{Y}\|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(P_0\|P_1)$
Calibration Error (CAL)	$\mathbb{E}\left[\left\|P_{Y\|X,S=0}(1\|X) - h(X)\right\| \middle\| S=0\right] - \mathbb{E}\left[\left\|P_{Y\|X,S=1}(1\|X) - h(X)\right\| \middle\| S=1\right]$
False Discovery Rate (FDR)	$\Pr(Y=0\|\hat{Y}=1,S=0) - \Pr(Y=0\|\hat{Y}=1,S=1)$
False Negative Rate (FNR)	$\Pr(\hat{Y}=0\|Y=1,S=0) - \Pr(\hat{Y}=0\|Y=1,S=1)$
False Positive Rate (FPR)	$\Pr(\hat{Y}=1\|Y=0,S=0) - \Pr(\hat{Y}=1\|Y=0,S=1)$

Table 1: Discrimination metrics $\mathsf{M}(P_0, P_1)$ for common fairness criteria. Distribution Alignment (DA) is a new metric related to the divergence in output distributions, which measures the statistical indistinguishability of two populations (see Appendix A). DA with $\lambda = 0$ measures the parity of predicted outcomes (similar to statistical parity in Corbett-Davies et al., 2017).

where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$.

Proposition 1 shows that counterfactual distributions always exist using a geometric argument.

Proposition 1. *For a given discrimination metric* $\mathsf{M}(P_0, P_1)$ *in Table 1, the set of counterfactual distributions is non-empty, closed, and convex.*

The uniqueness of a counterfactual distribution depends on our choice of discrimination metric. For a metric such as DA with $\lambda > 0$, there exists only one counterfactual distribution P_1. For other discrimination metrics of interest, the counterfactual distributions may not be unique (see Example 2 in the Appendix).

The distribution of input variables for the majority group P_1 is not necessarily a counterfactual distribution when $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

Example 1. Consider a classification problem in which the input variables $X = (X_1, X_2) \in \{-1, 1\}^2$ are drawn from a distribution $\Pr(X_i = 1|S = j) = p_{i,j}$ where $(p_{1,0}, p_{2,0}) = (0.9, 0.2)$ and $(p_{1,1}, p_{2,1}) = (0.1, 0.5)$, and the outcome variables are drawn from the conditional distributions:

$$P_{Y|X,S=0}(1|\boldsymbol{x}) = \mathsf{logistic}(2x_1 + x_2),$$
$$P_{Y|X,S=1}(1|\boldsymbol{x}) = \mathsf{logistic}(x_1 + 2x_2),$$

Given a Bayes optimal classifier for the majority group is $h(\boldsymbol{x}) = \mathbb{1}[x_2 = 1]$, the discrimination metric over a target population in terms of CAL is $\mathsf{M}(P_0, P_1) = 39.5\%$. Here, we have that $\mathsf{M}(P_1, P_1) = 23.1\% > 0$ for the majority distribution, while $\mathsf{M}(Q_X, P_1) = 0.0\%$ for the counterfactual distribution:

$$Q_X(-1,-1) = Q_X(1,1) = 0.42,$$
$$Q_X(-1,1) = Q_X(1,-1) = 0.08.$$

Example 1 shows that counterfactual distributions provide a tool to detect irreconcilable differences in the conditional distributions across groups. We formalize this in Proposition 2.

Proposition 2. *If* $\mathsf{M}(Q_X, P_1) > 0$ *where* Q_X *is a counterfactual distribution for a discrimination metric in Table 1, then* $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

This result illustrates how a counterfactual distribution can be used to detect cases where a classifier has an irreconcilable performance disparity between groups (i.e., a disparity that cannot be addressed by distribution of input variables for the minority group). The result complements a recent set of impossibility results on inevitable trade-offs between groups (see e.g., Lipton et al., 2018), and provides a sufficient testable condition that can motivate to train different classifiers for different groups (see e.g., the methods of Dwork et al., 2018; Zafar et al., 2017b).

3. Methodology

In this section, we present an information-theoretic methodology to recover counterfactual distributions. We first describe how influence functions provide a natural descent direction in our setting. We then present closed-form expressions of these functions that can be computed using data points from the target population. We end with a descent procedure that recovers a counterfactual distributions by combining these elements.

Influence Functions

The influence function (Huber, 2011; Koh and Liang, 2017) measures the change of a discrimination metric if a sample $\boldsymbol{x} \in \mathcal{X}$ from P_0 is removed (or added) in a very large dataset.

Definition 3. The influence function $\psi : \mathcal{X} \to \mathbb{R}$ is:

$$\psi(\boldsymbol{x}) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}, P_1\right) - \mathsf{M}(P_0, P_1)}{\epsilon} \tag{2}$$

where $\delta_{\boldsymbol{x}}(\mathbf{z}) \triangleq \mathbb{1}[\boldsymbol{x} = \mathbf{z}]$ is the Dirac delta function at $\boldsymbol{x}$.

We investigate how the discrimination metric in Table 1 decreases when the distribution P_0 is slightly perturbed. The local perturbation of a distribution is defined as follows.

Definition 4. The perturbed distribution $\widetilde{P}_0$ is

$$\widetilde{P}_0(\boldsymbol{x}) \triangleq P_0(\boldsymbol{x})(1 + \epsilon f(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in \mathcal{X} \tag{3}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a perturbation function from the class of all functions with zero mean and unit variance w.r.t. P_0, and $\epsilon > 0$ is a positive scaling constant chosen so that $\widetilde{P}_0$ is a valid probability distribution.

Here, $f(\boldsymbol{x})$ represents a direction in the probability simplex while ϵ represents the magnitude of perturbation. We show that the negatively normalized influence function indicates the direction of steepest descent of the discrimination metric.

Proposition 3. *For a given discrimination metric* $\mathsf{M}(P_0, P_1)$*, we have that*

$$\operatorname*{argmin}_{f(\boldsymbol{x})} \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_0, P_1) - \mathsf{M}(P_0, P_1)}{\epsilon} = \frac{-\psi(\boldsymbol{x})}{\sqrt{\mathbb{E}\left[\psi(X)^2 | S = 0\right]}}, \tag{4}$$

for any influence function $\psi : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}\left[\psi(X)^2 | S = 0\right] \neq 0$.

Perturbing the distribution P_0 along the negative influence function results in the largest local decrease of the discrimination metric.

When discrimination is measured using a combination of metrics (see e.g., Zafar et al., 2017a), Proposition 4 shows that the influence function for the compound metric is also a linear combination of the influence functions for each metric.

Proposition 4. *Given a compound discrimination metric which is a linear combination of* K *different discrimination metrics:* $\mathsf{M}(P_0, P_1) = \sum_{i=1}^{K} \lambda_i \mathsf{M}_i(P_0, P_1)$*, the influence function can be computed by*

$$\psi(\boldsymbol{x}) = \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x}). \tag{5}$$

Proposition 4 allows us to consider discrimination measures beyond the ones in Table 1. For instance, one can recover a counterfactual distribution for equalized odds (Hardt et al., 2016) by using a linear combination of influence functions for FPR and FNR. The result is also of particular interest given the sufficient condition in Proposition 2 and results on the impossibility of simultaneously satisfying multiple fairness criteria (see e.g., Chouldechova, 2017; Kleinberg et al., 2016; Pleiss et al., 2017). In this case, we would expect that the counterfactual distribution would not be able to resolve discrimination between mutually exclusive fairness objectives.

Computing Influence Functions

We present closed-form expressions of influence functions for discrimination metrics in Table 1. The expressions depend on three functions:

- $h(\boldsymbol{x})$, a fixed classifier that we wish to audit;
- $P_{S|X}(1|\boldsymbol{x})$, a conditional distribution of the sensitive attribute given features;
- $P_{Y|X,S=0}(1|\boldsymbol{x})$, a conditional distribution of the outcome given features for the minority group.

We estimate the last two models using an *auditing dataset* $\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^n$. Given this dataset, we: (i) train a classifier to predict the group membership, $P_{S|X}(1|\boldsymbol{x})$; and (ii) train a classifier to predict the outcome for individuals from the minority group, $P_{Y|X,S=0}(1|\boldsymbol{x})$. With these terms, we compute influence functions in closed-form.

Proposition 5. *The influence function under DA can be expressed as*

$$\psi(\boldsymbol{x}) = A \cdot h(\boldsymbol{x}) + \lambda \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - B, \tag{6}$$

where A and B are constants such that,

$$A \triangleq \log \frac{P_{\hat{Y}|S=0}(1) P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1) P_{\hat{Y}|S=0}(0)},$$
$$B \triangleq A \cdot P_{\hat{Y}|S=0}(1) + \lambda \mathbb{E}\left[\log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \middle| S = 0\right].$$

DA stands out as a discrimination metric since its influence function can be computed using only the classifier $h(\boldsymbol{x})$ and $P_{S|X}(1|\boldsymbol{x})$. In contrast, computing influence functions for other metrics in Table 1 require $P_{Y|X,S=0}(1|\boldsymbol{x})$. We state the expression of the influence function under CAL in the next proposition and present similar expressions for other metrics in Appendix C.

Proposition 6. *The influence function under CAL can be expressed as*

$$\psi(\boldsymbol{x}) = d(\boldsymbol{x}) - \mathbb{E}\left[d(X)|S = 0\right], \tag{7}$$

where $d(\boldsymbol{x}) \triangleq \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - h(\boldsymbol{x})\right|$.

Generalization Bounds

When the closed-form expressions of influence functions are estimated empirically from data, they can be subject to estimation error. In what follows, we show that the estimation error can be bounded in terms of the difference between the conditional distributions $P_{S|X}(1|\boldsymbol{x})$, $P_{Y|X,S=0}(1|\boldsymbol{x})$ and their estimators $\widehat{P}_{S|X}(1|\boldsymbol{x})$, $\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})$.

Proposition 7. *Let $\widehat{\psi}(\boldsymbol{x})$ and $\psi(\boldsymbol{x})$ be the estimated influence function and the true influence function, respectively. If the given discrimination metric is DA,*

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_p \leq O\left(\left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p\right).$$

For all other discrimination metrics in Table 1,

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_p \leq O\left(\left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p\right).$$

Here, $\|f(\boldsymbol{x}) - g(\boldsymbol{x})\|_p \triangleq \left(\mathbb{E}\left[|f(X) - g(X)|^p | S = 0\right]\right)^{1/p}$ denotes the ℓ_p-norm for $p \geq 1$.

In particular, when the estimators are empirical conditional distributions, we can obtain the following probabilistic upper bounds.

Corollary 1. *If $\widehat{P}_{S|X}$ and $\widehat{P}_{Y|X,S=0}$ are the empirical conditional distributions obtained from n i.i.d. samples, then, with probability at least $1 - \beta$,*

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_1 \leq O\left(\sqrt{n^{-1}\left(|\mathcal{X}| - \log\beta\right)}\right). \tag{8}$$

Recovering Counterfactual Distributions

In Algorithm 1, we present a descent procedure to approximate a counterfactual distribution for a given discrimination metric $\mathsf{M}(\cdot)$ using a classifier and a sample of points from the distribution $P_{X,Y,S}$ (split in terms of an auditing and hold out dataset).

At each iteration, the influence function $\psi(\boldsymbol{x})$ for $\mathsf{M}(\cdot)$ is computed over the auditing dataset, indicating the "direction" that the distribution over the minority group should be perturbed in order to reduce disparate impact. Since perturbing a distribution is equivalent to resampling, the entries in the holdout dataset corresponding to the minority population are resampled with weights $1 - \epsilon\psi(\boldsymbol{x})$, with ϵ corresponding to the step size. The resulting resampled dataset mimics one drawn from the perturbed distribution that locally minimizes $\mathsf{M}(\cdot)$. Discrimination is then assesed over this perturbed dataset, and the procedure repeats until $\mathsf{M}(\cdot)$ ceases to decrease. Note that this (greedy) algorithm is nothing more than (stochastic) gradient descent in the space of distributions over $\mathcal{X}$, with the resampling at each iteration corresponding to a gradient step.

In Figure 2, we show the progress of Algorithm 1 when recovering a counterfactual distribution for a synthetic dataset (see Appendix D.1 for details). The procedure efficiently converges to a counterfactual distribution.

4. Use Cases and Demonstrations

In this section, we describe how counterfactual distributions can be used to control disparate impact by helping understand discrimination by proxy through contrastive analyses, and by building a data preprocessor that can mitigate performance disparities. We support both use cases through experiments in which we recover counterfactual distributions for classifiers trained on real-world datasets.

Setup

We aim to recover counterfactual distributions for different discrimination metrics in Table 1 and classifiers trained on real world datasets. In particular, we consider processed versions of the `adult` dataset from the UCI ML Repository and the ProPublica `compas` dataset (Angwin et al., 2016).

Given each dataset, we use: 30% of samples to train a hypothetical classifier $h(\boldsymbol{x})$ that for the sake of an audit; 50% of samples to recover a counterfactual distribution via Algorithm 1; and 20% as a hold-out set to evaluate the performance of a preprocessor discussed shortly. Our

Algorithm 1 Distributional Descent

Input

$h : \mathcal{X} \to [0, 1]$ *classification model*
$\mathsf{M}(\cdot)$ *discrimination metric*
$\epsilon > 0$ *step size*
$P_{S|X}(1|\boldsymbol{x})$ *group membership model*
$P_{Y|X,S=0}(1|\boldsymbol{x})$ *outcome model for minority*
$\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{n}$ *auditing dataset*
$\mathcal{D}^{\text{holdout}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=n+1}^{m}$ *holdout dataset*

Initialize

$c(\boldsymbol{x}) \leftarrow 1$ *sampling weights*
$\mathcal{D} \leftarrow \mathcal{D}^{\text{audit}}$
$\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathcal{D}^{\text{holdout}}\right)$
repeat
  $\psi(\boldsymbol{x}) \leftarrow$ compute influence function for all $\boldsymbol{x} \in \mathcal{D}$
  $c(\boldsymbol{x}) \leftarrow (1 - \epsilon\psi(\boldsymbol{x}))c(\boldsymbol{x})$
  $\mathcal{D} \leftarrow \mathsf{Resample}(\mathcal{D}, 1 - \epsilon\psi(\boldsymbol{x}))$ *resample points*
  $\mathsf{M}_{\text{old}} \leftarrow \mathsf{M}_{\text{new}}$
  $\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathsf{Resample}\left(\mathcal{D}^{\text{holdout}}, c(\boldsymbol{x})\right)\right)$
until $\mathsf{M}_{\text{new}} \geq \mathsf{M}_{\text{old}}$
return: $c(\boldsymbol{x})$ *$c(\boldsymbol{x})$ is an aggregate perturbation*

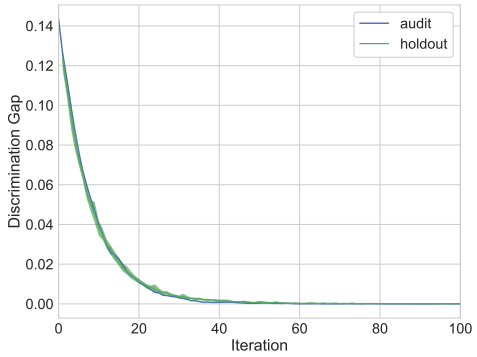

Figure 2: Values of DA for auditing dataset (blue) and holdout dataset (green), respectively, with each iteration in distributional descent for a synthetic dataset. Here, the procedure converges to a counterfactual distribution in 40 iterations. We show additional steps for the sake of illustration.

- $P_{S|X}(1|\boldsymbol{x})$, a model to estimate the distribution of the sensitive attribute from the features.
- $P_{Y|X,S=0}(1|\boldsymbol{x})$, a model to estimate the outcome distribution for individuals in the minority group $S = 0$ in the auditing dataset.

With these models in hand, we can estimate the value of influence functions through the closed-form expressions in Section 3. We fit all three classifiers using ℓ_2-logistic regression using a standard nested 10-CV setup to tune parameters and estimate performance. We provide additional details related to the datasets and models in Appendix D.

4.1 Removing Discrimination

Given a counterfactual distribution Q_X, we can mitigate disparate impact in a population of interest by construct a *preprocessor* to map each sample from P_0 to a new sample in Q_X. In particular, the

			No Preprocessing			With Preprocessor		Change in Performance	
Dataset	Metric	Minority Group	Majority Value	Minority Value	Disc. Gap	Minority Value	Disc. Gap New	Minority AUC Before	Minority AUC After
adult	CAL	Male	0.066	0.152	0.086	0.196	0.130	0.826	0.721
adult	FPR	Male	0.016	0.105	0.089	0.019	0.002	0.826	0.724
adult	FNR	Female	0.508	0.653	0.144	0.500	-0.008	0.893	0.857
adult	$DA_{0.0}$	Male			0.119		0.000	0.826	0.729
adult	$DA_{0.1}$	Male			0.206		0.000	0.826	0.683
compas	CAL	Non-white	0.206	0.210	0.004	0.215	0.009	0.733	0.717
compas	FPR	Non-white	0.152	0.269	0.116	0.129	-0.023	0.733	0.671
compas	FNR	White	0.377	0.567	0.190	0.537	0.160	0.713	0.703
compas	$DA_{0.0}$	Non-white			0.057		0.000	0.733	0.653
compas	$DA_{0.1}$	Non-white			0.131		0.000	0.733	0.615

Table 2: Changes in disparate impact and performance of a classification model when using a randomized preprocessor built using a counterfactual distribution. We report the value of performance metrics for various discrimination criteria in Table 1. All metrics are computed using a hold-out sample that is not used to train the model or build the preprocessor.

preprocessor can be obtained by solving an optimal transport problem of the form:

$$\min_{\gamma \in \Gamma(P_0, Q_X)} \int \mathsf{cost}(\mathbf{x}, \widetilde{\mathbf{x}}) d\gamma(\mathbf{x}, \widetilde{\mathbf{x}}). \tag{9}$$

Here, $\mathsf{cost}(\cdot, \cdot)$ denotes a user-specified cost function, and $\Gamma(P_0, Q_X)$ denotes the set of all couplings of P_0 and Q_X (i.e., a joint distribution on $\mathcal{X} \times \mathcal{X}$ with marginals P_0 and Q_X).

Given a coupling that achieves the minimal cost as γ^*, we can construct a randomized preprocessor that, given a sample $\boldsymbol{x}$, will return a perturbed sample $\widetilde{\boldsymbol{x}} = T(\boldsymbol{x})$ with probability $\gamma^*(\boldsymbol{x}, \widetilde{\boldsymbol{x}})/P_0(\boldsymbol{x})$. In practice, the optimal solution to (9) can be efficiently obtained via linear programming (LP) or other specialized techniques (see e.g., Johndrow and Lum, 2017).

In Table 2, we show the effectiveness when we use a counterfactual distribution to build a randomized preprocessor for the classifiers trained using `adult` and `compas`. We build the preprocessor by solving (9) for a counterfactual distribution recovered using Algorithm 1. As shown, the approach reduce disparate impact in the protected group, while having a minor effect on accuracy for decision points across the full ROC curve.

4.2 Understanding Discrimination

Counterfactual distributions provide a flexible tool for understanding discrimination through contrastive analysis (Wachter et al., 2017). As shown in Table 3, the change between the observed and counterfactual distributions can be directly computed and analyzed. Alternatively, individual and aggregate measures of the difference between P_0 and Q_X can be used to either identify prototypical samples (see e.g, Bien and Tibshirani, 2011; Kim et al., 2016), or score features in terms of their ability to discriminate by proxy in the target population. **In what follows, we describe these approaches and provide examples in the Appendix.**

Identifying Proxy Scores. Counterfactual distributions can be used to score proxy variables (see e.g., Datta et al., 2016; Adler et al., 2018).

Definition 5. Let the input to the classifier be given by $X = (X_1, \ldots, X_d)$. For a given discrimination metric and a counterfactual distribution Q_X, the proxy score for feature X_j is defined as

$$\gamma_j \triangleq \mathsf{dist}\left(P_{X_j}, Q_{X_j}\right), \tag{10}$$

where P_{X_j} and Q_{X_j} are the marginal distributions of X_j w.r.t. P_0 and Q_X, respectively, and $\mathsf{dist}(\cdot, \cdot)$ is a generic distance metric (e.g., the KL-divergence or $|P_{X_i}(1) - Q_{X_i}(1)|$ when the features are binary).

	Majority	Minority	Counterfactual Distribution		
	P_1	P_0	DA	FNR	CAL
Married	61.6	17.5	31.9	15.7	43.2
Immigrant	10.4	10.5	10.0	10.9	10.4
HighestDegree_is_HS	32.4	32.1	26.4	31.1	27.3
HighestDegree_is_AS	7.4	7.6	7.7	7.5	9.1
HighestDegree_is_BS	17.2	13.7	18.7	14.0	18.0
HighestDegree_is_MSorPhD	7.2	5.4	8.5	4.3	9.6
AnyCapitalLoss	4.5	3.2	6.4	3.3	6.1
Age_leq_30	31.1	40.1	33.6	40.6	27.5
WorkHrsPerWeek_lt_40	17.6	38.5	34.8	38.0	32.4
JobType_is_WhiteCollar	18.5	33.6	35.4	34.2	35.8
JobType_is_BlueCollar	33.8	4.6	3.7	4.5	3.8
JobType_is_Specialized	21.9	22.9	26.8	22.5	29.0
JobType_is_ArmedOrProtective	2.9	0.9	1.1	1.0	1.2
Industry_is_Private	69.2	70.2	66.3	71.0	65.7
Industry_is_Government	12.2	15.8	18.3	15.2	19.0
Industry_is_SelfEmployed	13.9	5.2	7.5	5.3	6.8

Table 3: Counterfactual distributions obtained using Algorithm 1 for a classifier on `adult`.

Prototypes. Given a counterfactual distribution Q_X, one can identify prototypes to highlight the characteristics that must change the least or the most in order to mitigate disparities (see e.g, Bien and Tibshirani, 2011; Kim et al., 2016). Prototypes can be chosen in terms of the value of $Q_X(\boldsymbol{x})/P_0(\boldsymbol{x})$. These examples can be chosen to reflect to maximize or minimize this score (Kim et al., 2016).

Local Influence. Our interpretation of influence functions motivates their use as a way to score proxy features and prototypes. Since influence functions represent the change in discrimination due to the addition or removal of a point from a dataset (by definition), prototypes that minimize or maximize the influence function $\phi(\mathbf{x})$ result in the largest / smallest local reduction in discrimination. Since an influence function reflects the direction of steepest descent with respect to a discrimination metric (see Proposition 3), proxy features can be scored in terms of the value of $\psi(X_1|X_{2,\ldots,d})$ (see Definition 6 in Appendix for details). The benefit of using influence functions in this setting is that they are unique and can be computed directly. However, they reflect local information with respect to the discrimination. Even as these metrics may not appear to differ significantly, small differences in these initial directions can lead to different counterfactual distributions.

5. Discussion and Related Work

In what follows, we describe limitations and extensions of our approach, and we frame our results within the context of existing work.

Limitations

Collecting & Predicting Protected Attributes: Our approach requires collecting data on sensitive attributes, such as race or gender, which may infringe privacy (though this is in general unavoidable, see e.g., Žliobaitė and Custers, 2016). It also requires training a model to detect membership in the minority class, which can be problematic if the model is used for other purposes (see Wachter and Mittelstadt, 2018, for a discussion).

Convergence Guarantees: While we find that our procedure consistently recovers counterfactual distribution in our experiments, our procedure is missing formal convergence guarantees. In practice, we expect convergence given that stochastic gradient descent still converges with noisy gradients (even in the presence of biased gradients Chen and Luss, 2018). One can potentially validate convergence

empirically (e.g., by computing the value of the discrimination metric). Nevertheless, a formal characterization of convergence would help guide specifications for the inputs to Algorithm 1 (e.g., the sample size of the auditing dataset). The bounds that we present in Proposition 7 and Corollary 1 provide a step in this direction.

Randomized Preprocessing: For the sake of generality, we have presented a way to mitigate discrimination using randomized preprocessors. While randomization is a common approach in the machine learning literature (see e.g., the randomized classifiers of Hardt et al., 2016; Agarwal et al., 2018), randomized preprocessors may not be practical in applications such as loan approval since an applicant could change a different predicted outcome by applying multiple times. Given we have considered counterfactual distributions for the minority group, potential costs of randomization are only incurred by individuals in the minority. Some effects of randomization can be mitigated by using deterministic mappings (e.g., sampling a deterministic mapping from the randomizer to deploy for a fixed window). One could also use a deterministic mapping similar to the majority vote classifier (see e.g., Germain et al., 2009), albeit at the cost of increasing discrimination

Extensions

We list next several future research directions.

Other Supervised Learning Models: The proposed framework can be adapted to other supervised learning models, such as ordinal regression and multiclass classification, so long as a well-defined discrimination metric is specified.

Hypothesis Test for Irreconcilable Differences: One can adapt the result in Prosition 2 to design a hypothesis test for discrimination under covariate shift $H_0 : M(Q_X, P_1) > 0$. This could be used to inform when to train different classifiers for different groups as in (Dwork et al., 2018; Lipton et al., 2018).

Multiple Sensitive Groups: Our approach can be directly extended to handle problems with multiple sensitive attributes by using a one-vs-all approach. This may provide an interesting alternative to identify subgroups in which discrimination is most pronounced (see e.g, Chouldechova and G'Sell, 2017; Zhang and Neill, 2016). This extension can benefit from further study given that the choice of reference group will affect the reliability with which one can estimate counterfactual distributions and the resulting ability to mitigate discrimination.

References

Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

Solon Barocas and Andrew Selbst. Big data's disparate impact. 2016.

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, pages 2403–2424, 2011.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

Jie Chen and Ronny Luss. Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*, 2018.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy*, pages 598–617. IEEE, 2016.

Nicholas Diakopoulos. Algorithmic-accountability: the investigation of black boxes. *Tow Center for Digital Journalism*, 2014.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark DM Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360. ACM, 2009.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, 2016.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity? *arXiv preprint arXiv:1711.07076*, 2018.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. *arXiv preprint arXiv:1702.08536*, 2017.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5684–5693, 2017.

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Masashi Sugiyama, Neil D Lawrence, Anton Schwaighofer, et al. *Dataset shift in machine learning*. The MIT Press, 2017.

Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai. 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology (Forthcoming)*, 2017.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017b.

Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.

Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.

Indrė Žliobaitė and Bart Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201, 2016.

Appendix A. Further Discussions of Metrics

Distribution Alignment (DA)

We motivate the choice of KL-divergence based on the following binary hypothesis testing framework.

We consider n i.i.d.samples x^n drawn from an unknown distribution Q such that $Q \in \{Q_1, Q_2\}$. To determine the source distribution, we construct a decision rule that will choose between the hypotheses $H_1 : Q = Q_1$ and $H_2 : Q = Q_2$. The Neyman–Pearson lemma states that the optimal decision rule is given by the two acceptance regions, $\mathcal{A}_1^{(n)} = \{x^n \mid \log \frac{Q_1(x^n)}{Q_2(x^n)} > T\}$ and $\mathcal{A}_2^{(n)} = (\mathcal{A}_1^{(n)})^c$. A Type I (resp. Type II) error occurs when x^n is drawn from Q_1 (resp. Q_2) but $x^n \in \mathcal{A}_2^{(n)}$ (resp. $x^n \in \mathcal{A}_1^{(n)}$). Let $\beta_1^{(n)}$ (respectively, $\beta_2^{(n)}$) be the probability of Type I (resp. Type II) error.

When n is finite, there is a trade-off between Type I and Type II errors. From the Chernoff-Stein lemma (Cover and Thomas, 2012), for a desired $\beta_1^{(n)} \leq \delta$ (where $0 < \delta < 0.5$), the Type II error exponent of minimal $\beta_2^{(n)}$, denoted as $\beta_2^{(n)}(\delta)$, is $\lim_{n\to\infty} -\frac{1}{n} \log \beta_2^{(n)}(\delta) = \mathrm{D}_{\mathrm{KL}}(Q_1 \| Q_2)$. Hence, we use KL-divergence to measure the statistical indistinguishability of two populations in order to control error exponents.

Factorization of Joint Distribution

Following from the framework, we have the graphical model in Figure 3 which corresponds to a factorization of the joint distribution $P_{S,X,Y,\hat{Y}}$:

$$P_{S,X,Y,\hat{Y}} = P_{\hat{Y}|X} P_{Y|X,S} P_S P_{X|S}. \tag{11}$$

We view the classifier $h(\boldsymbol{x})$ as the conditional distribution $P_{\hat{Y}|X}(1|\boldsymbol{x})$ (i.e., $h(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})$).

Next, we show that given $P_{\hat{Y}|X}$, $P_{Y|X,S}$, and P_S, all discrimination metrics in Table 1 can be cast in terms of P_0 and P_1.

1. **DA.**

$$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0} \| P_{\hat{Y}|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(P_0 \| P_1) = \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X} \circ P_0 \| P_{\hat{Y}|X} \circ P_1) + \lambda \mathrm{D}_{\mathrm{KL}}(P_0 \| P_1). \tag{12}$$

2. **CAL.**

$$\begin{aligned} &\mathbb{E}\left[\left|P_{Y|X,S=0}(1|X) - h(X)\right| \middle| S = 0\right] - \mathbb{E}\left[\left|P_{Y|X,S=1}(1|X) - h(X)\right| \middle| S = 1\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - h(\boldsymbol{x})\right| P_0(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{Y|X,S=1}(1|\boldsymbol{x}) - h(\boldsymbol{x})\right| P_1(\boldsymbol{x}). \end{aligned} \tag{13}$$

3. **FDR.**

$$\begin{aligned} &\Pr(Y = 0|\hat{Y} = 1, S = 0) - \Pr(Y = 0|\hat{Y} = 1, S = 1) \\ &= \frac{\Pr(Y = 0, \hat{Y} = 1, S = 0)}{\Pr(\hat{Y} = 1, S = 0)} - \frac{\Pr(Y = 0, \hat{Y} = 1, S = 1)}{\Pr(\hat{Y} = 1, S = 1)} \\ &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=0}(0|\boldsymbol{x}) P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_0(\boldsymbol{x})} - \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=1}(0|\boldsymbol{x}) P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_1(\boldsymbol{x})}. \end{aligned} \tag{14}$$

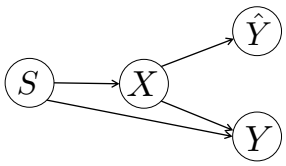

Figure 3: Graphical model of the framework.

4. **FNR.**

$$\begin{aligned}&\Pr(\hat{Y}=0|Y=1,S=0)-\Pr(\hat{Y}=0|Y=1,S=1)\\&=\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=0}(1|\boldsymbol{x})P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=0}(1|\boldsymbol{x})P_0(\boldsymbol{x})}-\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=1}(1|\boldsymbol{x})P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=1}(1|\boldsymbol{x})P_1(\boldsymbol{x})}.\end{aligned}\tag{15}$$

5. **FPR.**

$$\begin{aligned}&\Pr(\hat{Y}=1|Y=0,S=0)-\Pr(\hat{Y}=1|Y=0,S=1)\\&=\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=0}(0|\boldsymbol{x})P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=0}(0|\boldsymbol{x})P_0(\boldsymbol{x})}-\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=1}(0|\boldsymbol{x})P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=1}(0|\boldsymbol{x})P_1(\boldsymbol{x})}.\end{aligned}\tag{16}$$

Appendix B. Omitted Proofs

B.1 Proof of Proposition 1

Proof. In what follows, we show, for different discrimination metrics listed in Table 1, the following set of counterfactual distributions is non-empty, closed and convex.

$$\left\{Q_X\in\mathcal{P}\;\middle|\;\mathsf{M}(Q_X,P_1)=\inf_{Q'_X\in\mathcal{P}}|\mathsf{M}(Q'_X,P_1)|\right\},$$

where $\mathcal{P}$ is the set of probability distribution over $\mathcal{X}$. We assume that:

$$P_{Y|X,S=0}(1|\boldsymbol{x})>0,\;\; h(\boldsymbol{x})>0,\;\;\forall x\in\mathcal{X}.\tag{17}$$

1. **DA.** When $\lambda>0$, there is only one counterfactual distribution: P_1. This is because:

   (a) KL-divergence is non-negative:

   $$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X}\circ Q'_X\|P_{\hat{Y}|X}\circ P_1)+\lambda\mathrm{D}_{\mathrm{KL}}(Q'_X\|P_1)\geq 0;$$

   (b) If and only if $Q'_X=P_1$ (Cover and Thomas, 2012),

   $$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X}\circ Q'_X\|P_{\hat{Y}|X}\circ P_1)=0 \text{ and } \mathrm{D}_{\mathrm{KL}}(Q'_X\|P_1)=0.$$

   On the other hand, when $\lambda=0$, there may be multiple counterfactual distributions besides P_1. In this case, the set of counterfactual distributions is closed and convex following from standard continuity and convexity results (Cover and Thomas, 2012).

2. **CAL.** Note that

   $$\mathbb{E}\left[\left|P_{Y|X,S=0}(1|X)-h(X)\right|\middle|S=0\right]=\sum_{\boldsymbol{x}\in\mathcal{X}}d(\boldsymbol{x})P_0(\boldsymbol{x}),$$

   where $d(\boldsymbol{x})\triangleq\left|P_{Y|X,S=0}(1|\boldsymbol{x})-h(\boldsymbol{x})\right|$. Then

   $$|\mathsf{M}(Q'_X,P_1)|=\left|\sum_{\boldsymbol{x}\in\mathcal{X}}d(\boldsymbol{x})Q'_X(\boldsymbol{x})-\mathbb{E}\left[\left|P_{Y|X,S=1}(1|X)-h(X)\right|\middle|S=1\right]\right|,$$

   which implies, for fixed h, $P_{Y|X,S}$, P_1, the mapping $Q'_X\to|\mathsf{M}(Q'_X,P_1)|$ is continuous. Therefore, the set of counterfactual distributions is closed and non-empty.

We define

$$m_{\mathrm{CAL}} \triangleq \max\left\{\min_{Q'_X \in \mathcal{P}}\left\{\sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x})Q'_X(\boldsymbol{x})\right\}, \mathbb{E}\left[\left|P_{Y|X,S=1}(1|X) - h(X)\right| \middle| S=1\right]\right\}.$$

Then counterfactual distributions, following from Definition 2, are distributions which satisfy

$$\sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x})Q_X(\boldsymbol{x}) = m_{\mathrm{CAL}}.$$

For any two counterfactual distributions Q_1 and Q_2 and $\mu \in [0,1]$, $\mu Q_1 + (1-\mu)Q_2$ is a feasible probability distribution. Also of note,

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x})(\mu Q_1(\boldsymbol{x}) + (1-\mu)Q_2(\boldsymbol{x})) \\
&= \mu \sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x})Q_1(\boldsymbol{x}) + (1-\mu)\sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x})Q_2(\boldsymbol{x}) \\
&= m_{\mathrm{CAL}}.
\end{aligned}$$

Hence, $\mu Q_1 + (1-\mu)Q_2$ is also a counterfactual distribution. Therefore, the counterfactual distributions form a convex set.

3. **Class-Based Error Metrics.** Here we only show the proof for the FNR. Similar analysis also holds for the FDR and the FPR. Note that

$$\Pr(\hat{Y}=0|Y=1,S=0) = \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} r_1(\boldsymbol{x})P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} r_2(\boldsymbol{x})P_0(\boldsymbol{x})},$$

where $r_1(\boldsymbol{x}) \triangleq (1-h(\boldsymbol{x}))P_{Y|X,S=0}(1|\boldsymbol{x})$ and $r_2(\boldsymbol{x}) \triangleq P_{Y|X,S=0}(1|\boldsymbol{x})$. Then

$$|\mathsf{M}(Q'_X, P_1)| = \left|\frac{\sum_{\boldsymbol{x}\in\mathcal{X}} r_1(\boldsymbol{x})Q'_X(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} r_2(\boldsymbol{x})Q'_X(\boldsymbol{x})} - \Pr(\hat{Y}=0|Y=1,S=1)\right|,$$

which implies, for fixed h, $P_{Y|X,S}$, P_1, the mapping $Q'_X \to |\mathsf{M}(Q'_X, P_1)|$ is continuous. Therefore, the set of counterfactual distributions is closed and non-empty.

We define

$$m_{\mathrm{FNR}} \triangleq \max\left\{\min_{Q'_X \in \mathcal{P}}\left\{\frac{\sum_{\boldsymbol{x}\in\mathcal{X}} r_1(\boldsymbol{x})Q'_X(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} r_2(\boldsymbol{x})Q'_X(\boldsymbol{x})}\right\}, \Pr(\hat{Y}=0|Y=1,S=1)\right\}.$$

Then counterfactual distributions, following from Definition 2, are distributions which satisfy

$$\frac{\sum_{\boldsymbol{x}\in\mathcal{X}} r_1(\boldsymbol{x})Q_X(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} r_2(\boldsymbol{x})Q_X(\boldsymbol{x})} = m_{\mathrm{FNR}}.$$

This is equivalent to

$$\sum_{\boldsymbol{x}\in\mathcal{X}} (r_1(\boldsymbol{x}) - m_{\mathrm{FNR}} r_2(\boldsymbol{x}))Q_X(\boldsymbol{x}) = 0.$$

For any two counterfactual distributions Q_1 and Q_2 and $\mu \in [0,1]$, $\mu Q_1 + (1-\mu)Q_2$ is a feasible distribution. Also of note,

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathcal{X}} (r_1(\boldsymbol{x}) - m_{\mathrm{FNR}} r_2(\boldsymbol{x}))(\mu Q_1(\boldsymbol{x}) + (1-\mu)Q_2(\boldsymbol{x})) \\
&= \mu \sum_{\boldsymbol{x}\in\mathcal{X}} (r_1(\boldsymbol{x}) - m_{\mathrm{FNR}} r_2(\boldsymbol{x}))Q_1(\boldsymbol{x}) + (1-\mu)\sum_{\boldsymbol{x}\in\mathcal{X}} (r_1(\boldsymbol{x}) - m_{\mathrm{FNR}} r_2(\boldsymbol{x}))Q_2(\boldsymbol{x}) \\
&= 0.
\end{aligned}$$

Hence, $\mu Q_1 + (1-\mu)Q_2$ is also a counterfactual distribution. Therefore, the counterfactual distributions form a convex set.

□

B.2 Example of Counterfactual Distributions

We show that the counterfactual distributions are not always unique.

Example 2. We use DA with $\lambda = 0$ as a discrimination metric. Specifically, its definition is

$$\mathsf{M}(P_0, P_1) = \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0} \| P_{\hat{Y}|S=1}). \tag{18}$$

We choose $X|S=0 \sim \mathsf{Bernoulli}(0.1)$ and $X|S=1 \sim \mathsf{Bernoulli}(0.2)$. The classifier is chosen as $h(0) = h(1) = 0.2$ (i.e., $P_{\hat{Y}|X}(1|0) = P_{\hat{Y}|X}(1|1) = 0.2$). In this case, any Bernoulli distribution, including P_0 and P_1, over $\{0,1\}$ is a counterfactual distribution.

B.3 Proof of Proposition 2

Proof. First, the counterfactual distributions under DA always achieve zero of the discrimination metric. Hence, $\mathsf{M}(Q_X, P_1)$ happens only if the discrimination metric is not DA. We assume that $P_{Y|X,S=0} = P_{Y|X,S=1}$ and $\mathsf{M}(Q_X, P_1) > 0$. In particular, $|\mathsf{M}(P_1, P_1)| \geq \mathsf{M}(Q_X, P_1) > 0$. Note that the discrimination metrics in Table 1 except DA are the form of the discrepancies of performance metrics between two groups. Here the performance metric for each group only depends on $P_{Y|X,S=i}$, $P_{X|S=i}$, and $P_{\hat{Y}|X}$. If we assume that $P_{Y|X,S=0} = P_{Y|X,S=1}$ and set the distribution of minority group as P_1, then the performance metrics achieve the same values for two groups. Hence, $\mathsf{M}(P_1, P_1) = 0$ which contradicts the assumption, so $P_{Y|X,S=0} \neq P_{Y|X,S=1}$. □

B.4 Proof of Proposition 3

Proof. First, we define

$$\Delta(f) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_0, P_1) - \mathsf{M}(P_0, P_1)}{\epsilon}, \tag{19}$$

where $\widetilde{P}_0(\boldsymbol{x})$ is the perturbed distribution defined in (3). Then we prove that

$$\Delta(f) = \mathbb{E}\left[f(X)\psi(X) | S = 0\right].$$

Note that an alternative way (see e.g., Huber, 2011) to define influence functions is in terms of the Gâteaux derivative:

$$\sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) P_0(\boldsymbol{x}) = 0,$$

and

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathsf{M}\left((1-\epsilon)P_0 + \epsilon Q, P_1\right) - \mathsf{M}\left(P_0, P_1\right)\right) = \sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) Q(\boldsymbol{x}), \ \forall Q \in \mathcal{P}.$$

In particular, we can choose

$$Q(\boldsymbol{x}) = \left(\frac{1}{M_U} f(\boldsymbol{x}) + 1\right) P_0(\boldsymbol{x}),$$

where $M_U \triangleq \sup\{|f(\boldsymbol{x})| \mid \boldsymbol{x} \in \mathcal{X}\} + 1$. Then

$$(1-\epsilon)P_0(\boldsymbol{x}) + \epsilon Q(\boldsymbol{x}) = P_0(\boldsymbol{x}) + \frac{\epsilon}{M_U} f(\boldsymbol{x}) P_0(\boldsymbol{x}).$$

For simplicity, we use $P_0 + \epsilon f P_0$ and $P_0 + \frac{\epsilon}{M_U} f P_0$ to represent $P_0(\boldsymbol{x}) + \epsilon f(\boldsymbol{x}) P_0(\boldsymbol{x})$ and $P_0(\boldsymbol{x}) + \frac{\epsilon}{M_U} f(\boldsymbol{x}) P_0(\boldsymbol{x})$, respectively. Then

$$\begin{aligned}
\Delta(f) &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathsf{M}(P_0 + \epsilon f P_0, P_1) - \mathsf{M}(P_0, P_1) \right) \\
&= \lim_{\epsilon \to 0} \frac{M_U}{\epsilon} \left(\mathsf{M}\left(P_0 + \frac{\epsilon}{M_U} f P_0, P_1 \right) - \mathsf{M}(P_0, P_1) \right) \\
&= M_U \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathsf{M}((1-\epsilon)P_0 + \epsilon Q, P_1) - \mathsf{M}(P_0, P_1) \right) \\
&= M_U \sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) Q(\boldsymbol{x}) \\
&= M_U \sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) \left(\frac{1}{M_U} f(\boldsymbol{x}) + 1 \right) P_0(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) f(\boldsymbol{x}) P_0(\boldsymbol{x}) \\
&= \mathbb{E}\left[f(X) \psi(X) | S = 0 \right].
\end{aligned}$$

Following from Cauchy-Schwarz inequality,

$$\mathbb{E}\left[f(X)\psi(X) | S = 0 \right] \geq -\sqrt{\mathbb{E}\left[f(X)^2 | S = 0 \right]} \sqrt{\mathbb{E}\left[\psi(X)^2 | S = 0 \right]} = -\sqrt{\mathbb{E}\left[\psi(X)^2 | S = 0 \right]}.$$

Here the equality can be achieved by choosing

$$f(\boldsymbol{x}) = \frac{-\psi(\boldsymbol{x})}{\sqrt{\mathbb{E}\left[\psi(X)^2 | S = 0 \right]}}.$$

□

B.5 Proof of Proposition 4

Proof. When the discrimination metric is a linear combination of K different discrimination metrics:

$$\mathsf{M}(P_0, P_1) = \sum_{i=1}^{K} \lambda_i \mathsf{M}_i(P_0, P_1),$$

the influence function, following from Definition 3, is

$$\begin{aligned}
\psi(\boldsymbol{x}) &= \lim_{\epsilon \to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon \delta_{\boldsymbol{x}}, P_1 \right) - \mathsf{M}(P_0, P_1)}{\epsilon} && (20) \\
&= \sum_{i=1}^{K} \lambda_i \lim_{\epsilon \to 0} \frac{\mathsf{M}_i\left((1-\epsilon)P_0 + \epsilon \delta_{\boldsymbol{x}}, P_1 \right) - \mathsf{M}_i(P_0, P_1)}{\epsilon} && (21) \\
&= \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x}). && (22)
\end{aligned}$$

□

B.6 Proof of Proposition 7 and Corollary 1

Proof. We denote $\widehat{P}$ and $\widehat{\Pr}$ as estimated probability distribution and probability, respectively. Then we assume that

$$\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_p \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \quad (23)$$

$$\left|\widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \quad (24)$$

$$\left|\widehat{\Pr}(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p, \quad (25)$$

where $\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_p \triangleq \left(\sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{X|S=0}(\boldsymbol{x}) - P_{X|S=0}(\boldsymbol{x})\right|^p\right)^{1/p}$. We make similar assumptions for $\widehat{P}_{S|X}(1|\boldsymbol{x})$ (i.e., the $\mathcal{L}_p$ distance between $\widehat{P}_{S|X}(1|\boldsymbol{x})$ and $P_{S|X}(1|\boldsymbol{x})$ upper bounds the left-hand side of (23), (24), (25)). These assumptions are reasonable in practice since estimating conditional distribution is usually harder than estimating marginal distribution which is harder than estimating the distribution of Bernoulli random variable. The closed-form expressions of influence functions under different discrimination metrics are given in Proposition 8.

1. **CAL.** The influence function under calibration error is

$$\begin{aligned}\psi(\boldsymbol{x}) &= d(\boldsymbol{x}) - \mathbb{E}\left[d(X)|S=0\right] \\ &= \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| - \sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}).\end{aligned}$$

The estimated influence function under calibration error is

$$\widehat{\psi}(\boldsymbol{x}) = \left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| - \sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \widehat{P}_{X|S=0}(\boldsymbol{x}).$$

By the triangle inequality, we have

$$\begin{aligned}&\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| - \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \\ &\leq \left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right|. \end{aligned} \quad (26)$$

Similarly, we have

$$\begin{aligned}&\left|\sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \widehat{P}_{X|S=0}(\boldsymbol{x})\right| \\ &\leq \sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{X|S=0}(\boldsymbol{x}) - \widehat{P}_{X|S=0}(\boldsymbol{x})\right|. \end{aligned} \quad (27)$$

Following (26) and (27), we know, for calibration error,

$$
\begin{aligned}
&\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\|_p \\
&\leq \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p \\
&\quad + \left\|\sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x})\right\|_p + \left\|\sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{X|S=0}(\boldsymbol{x}) - \widehat{P}_{X|S=0}(\boldsymbol{x})\right|\right\|_p \\
&= \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p \\
&\quad + \sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}} \left|P_{X|S=0}(\boldsymbol{x}) - \widehat{P}_{X|S=0}(\boldsymbol{x})\right| \\
&= \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p \\
&\quad + \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_1 + \left\|P_{X|S=0} - \widehat{P}_{X|S=0}\right\|_1 \\
&\leq 2\left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p + \left\|P_{X|S=0} - \widehat{P}_{X|S=0}\right\|_1 \\
&\lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p .
\end{aligned}
$$

2. **Class-Based Error Metrics.** Next, we present a proof of the generalization bound for FNR. Similar proofs hold for other class-based error metrics such as FDR and FPR.

The influence function under FNR is

$$
\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[r_2(X)|S=0\right] r_1(\boldsymbol{x}) - \mathbb{E}\left[r_1(X)|S=0\right] r_2(\boldsymbol{x})}{\Pr(Y=1|S=0)^2},
$$

where $r_1(\boldsymbol{x}) = P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=0}(1|\boldsymbol{x})$ and $r_2(\boldsymbol{x}) = P_{Y|X,S=0}(1|\boldsymbol{x})$. Note that

$$
\begin{aligned}
\mathbb{E}\left[r_2(X)|S=0\right] &= \Pr(Y=1|S=0), \\
\mathbb{E}\left[r_1(X)|S=0\right] &= \Pr(\hat{Y}=0, Y=1|S=0).
\end{aligned}
$$

Hence, the influence function under FNR has the following equivalent expression.

$$
\begin{aligned}
\psi(\boldsymbol{x}) &= \frac{\Pr(Y=1|S=0)P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0, Y=1|S=0)}{\Pr(Y=1|S=0)^2} P_{Y|X,S=0}(1|\boldsymbol{x}) \\
&= \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} P_{Y|X,S=0}(1|\boldsymbol{x}). \qquad (28)
\end{aligned}
$$

The estimated influence function under FNR is

$$
\widehat{\psi}(\boldsymbol{x}) = \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}). \qquad (29)
$$

Following from (28), (29) and the triangle inequality, we have, for FNR,

$$
\begin{aligned}
&\|\psi(\boldsymbol{x}) - \widehat{\psi}(\boldsymbol{x})\|_p \\
&\leq \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} (P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})) \right\|_p \\
&\quad + \left\| \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \left(\frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} \right.\right. \\
&\qquad \left.\left. - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right) \right\|_p \\
&\leq \left\| \frac{1}{\Pr(Y=1|S=0)} (P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})) \right\|_p \\
&\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p \\
&\lesssim \left\| P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \right\|_p \\
&\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p . \qquad (30)
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
&\left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p \\
&\leq \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\widehat{\Pr}(Y=1|S=0)} \right\|_p + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{\widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right| \\
&\leq \left| \frac{\widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\
&\quad + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\
&\lesssim \left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| \\
&\quad + \left| \Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0) \right| \\
&\leq 2 \left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| + \left| \widehat{\Pr}(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=0) \right| . \qquad (31)
\end{aligned}
$$

Combining (30) and (31) with the assumptions (24) and (25), we have, for FNR,

$$
\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\|_p \lesssim \left\| P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \right\|_p .
$$

3. **DA.** The influence function under DA is

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

Since $h(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})$ is a given classifier, estimating

$$\left(\log \frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)}\right)\left(P_{\hat{Y}|X}(1|\boldsymbol{x}) - P_{\hat{Y}|S=0}(1)\right)$$

is more reliable than estimating

$$\begin{aligned}\psi_r(\boldsymbol{x}) &\triangleq \log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log\frac{1-P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle|S=0\right] \\ &= \log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\mathcal{X}} P_{X|S=0}(\boldsymbol{x})\log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}. \end{aligned} \tag{32}$$

Next, we bound the generalization error of estimating $\psi_r(\boldsymbol{x})$. Its estimator is

$$\widehat{\psi}_r(\boldsymbol{x}) = \log\frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x}\in\mathcal{X}} \widehat{P}_{X|S=0}(\boldsymbol{x})\log\frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})}. \tag{33}$$

Note that, for $a, b > 0$,

$$\left|\log\frac{a}{b}\right| \le \frac{|a-b|}{\min\{a,b\}}. \tag{34}$$

Then

$$\begin{aligned}&\left|\log\frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}\right| \\ &\le |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})|\left(\frac{1}{\min\{\widehat{P}_{S|X}(1|\boldsymbol{x}), P_{S|X}(1|\boldsymbol{x})\}} + \frac{1}{\min\{1-\widehat{P}_{S|X}(1|\boldsymbol{x}), 1-P_{S|X}(1|\boldsymbol{x})\}}\right) \\ &\le |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})|\frac{2}{m_X}, \end{aligned} \tag{35}$$

where m_X is a constant number:

$$m_X \triangleq \min\left\{\left\{\widehat{P}_{S|X}(1|\boldsymbol{x})|\boldsymbol{x}\in\mathcal{X}\right\} \cup \left\{P_{S|X}(1|\boldsymbol{x})|\boldsymbol{x}\in\mathcal{X}\right\} \cup \left\{1-\widehat{P}_{S|X}(1|\boldsymbol{x})|\boldsymbol{x}\in\mathcal{X}\right\} \cup \left\{1-P_{S|X}(1|\boldsymbol{x})|\boldsymbol{x}\in\mathcal{X}\right\}\right\}.$$

Also of note, for any $\boldsymbol{x}\in\mathcal{X}$,

$$\left|\log\frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})}\right| \le \frac{\left|1-2\widehat{P}_{S|X}(1|\boldsymbol{x})\right|}{\min\left\{\widehat{P}_{S|X}(1|\boldsymbol{x}), 1-\widehat{P}_{S|X}(1|\boldsymbol{x})\right\}} \le \frac{1}{m_X}. \tag{36}$$

Combining (32) and (33) with (35) and (36), we have

$$\begin{aligned}
&\left|\widehat{\psi}_r(\boldsymbol{x}) - \psi_r(\boldsymbol{x})\right| \\
&\leq \frac{2}{m_X}\left|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right| + \frac{1}{m_X}\sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{X|S=0}(\boldsymbol{x}) - P_{X|S=0}(\boldsymbol{x})\right| \\
&\quad + \frac{2}{m_X}\sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) \\
&= \frac{2}{m_X}\left|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right| + \frac{1}{m_X}\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_1 + \frac{2}{m_X}\left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|\widehat{\psi}_r(\boldsymbol{x}) - \psi_r(\boldsymbol{x})\right\|_p \\
&\leq \frac{2}{m_X}\left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p + \frac{1}{m_X}\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_1 + \frac{2}{m_X}\left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_1.
\end{aligned}$$

Based on the assumption: $\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_1 \lesssim \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_1$, we have

$$\begin{aligned}
\left\|\widehat{\psi}_r(\boldsymbol{x}) - \psi_r(\boldsymbol{x})\right\|_p &\lesssim \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p + \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_1 \\
&\lesssim \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p.
\end{aligned}$$

Hence, for DA,

$$\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\|_p \lesssim \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p.$$

If $\widehat{P}_{Y|X,S=0} = \frac{\widehat{P}_{Y,X|S=0}}{\widehat{P}_{X|S=0}}$ is the empirical conditional distribution obtained from m i.i.d. samples, then, following from (39), with probability at least $1-\beta$,

$$\left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \leq \sqrt{\frac{2}{m}\left(2|\mathcal{X}| - \log\beta\right)}. \tag{37}$$

Therefore, for the discrimination metrics in Table 1 except DA, with probability at least $1-\beta$,

$$\begin{aligned}
\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_1 &\lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_1 \\
&\lesssim \left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \\
&\lesssim \sqrt{\frac{1}{m}\left(|\mathcal{X}| - \log\beta\right)}.
\end{aligned}$$

Here the second inequality holds true because

$$\begin{aligned}
&\left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_1 \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} P_{X|S=0}(\boldsymbol{x})\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right| \\
&\leq \left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 + \sum_{\boldsymbol{x}\in\mathcal{X}}\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\left|\widehat{P}_{X|S=0}(\boldsymbol{x}) - P_{X|S=0}(\boldsymbol{x})\right| \\
&\leq \left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 + \left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_1 \lesssim \left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1.
\end{aligned}$$

Corollary 1 follows from Proposition 7 and the following large deviation results by (Weissman et al., 2003).

For all $\epsilon > 0$,

$$\Pr\left(\|\widehat{P} - P\|_1 \geq \epsilon\right) \leq (2^M - 2)\exp\left(-n\bar{\phi}(\pi_P)\epsilon^2/4\right),$$

where P is a probability distribution on the set $[M]$, $\widehat{P}$ is the empirical distribution obtained from n i.i.d. samples, $\pi_P \triangleq \max_{\mathcal{M}\subseteq[M]} \min(P(\mathcal{M}), 1 - P(\mathcal{M}))$,

$$\bar{\phi}(p) \triangleq \begin{cases} \frac{1}{1-2p}\log\frac{1-p}{p} & p \in [0, 1/2), \\ 2 & p = 1/2, \end{cases}$$

and $\|\widehat{P} - P\|_1 \triangleq \sum_{x\in\mathcal{X}} |\widehat{P}(x) - P(x)|$. Note that $\bar{\phi}(\pi_P) \geq 2$ which implies that

$$\Pr\left(\|\widehat{P} - P\|_1 \geq \epsilon\right) \leq \exp(M)\exp(-n\epsilon^2/2). \tag{38}$$

Hence, by taking $P = P_{Y,X|S=0}$, $M = |\mathcal{Y}||\mathcal{X}| = 2|\mathcal{X}|$ and $\epsilon = \sqrt{\frac{2}{n}(M - \log\beta)}$, Inequality (38) implies that, with probability at least $1 - \beta$,

$$\left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \leq \sqrt{\frac{2}{n}(2|\mathcal{X}| - \log\beta)}, \tag{39}$$

where $\widehat{P}_{Y,X|S=0}$ is the empirical distribution obtained from n i.i.d. samples. Similarly, with probability at least $1 - \beta$,

$$\left\|\widehat{P}_{S,X} - P_{S,X}\right\|_1 \leq \sqrt{\frac{2}{n}(2|\mathcal{X}| - \log\beta)}. \tag{40}$$

□

Appendix C. Computing Influence Functions

C.1 Closed-Form Expressions of Influence Functions

We give the closed-form expressions of influence functions under different discrimination metrics in this section. Here we view the classifier $h(\boldsymbol{x})$ as a conditional distribution $P_{\hat{Y}|X}(1|\boldsymbol{x})$.

Proposition 8. *Influence functions for the discrimination metrics in Table 1 can be expressed as*

- *DA:*

$$\begin{aligned}\psi(\boldsymbol{x}) = &\left(\log\frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)}\right)\left(P_{\hat{Y}|X}(1|\boldsymbol{x}) - P_{\hat{Y}|S=0}(1)\right) \\ &+ \lambda\left(\log\frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log\frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle| S = 0\right]\right);\end{aligned} \tag{41}$$

- *CAL:*

$$\psi(\boldsymbol{x}) = d(\boldsymbol{x}) - \mathbb{E}\left[d(X)|S = 0\right], \tag{42}$$

  where $d(\boldsymbol{x}) = \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right|$*;*

- *FDR:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[t_2(X)|S = 0\right]t_1(\boldsymbol{x}) - \mathbb{E}\left[t_1(X)|S = 0\right]t_2(\boldsymbol{x})}{\Pr(\hat{Y} = 1|S = 0)^2}, \tag{43}$$

  where $t_1(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=0}(0|\boldsymbol{x})$ *and* $t_2(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})$*;*

- *FNR:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[r_2(X)|S=0\right] r_1(\boldsymbol{x}) - \mathbb{E}\left[r_1(X)|S=0\right] r_2(\boldsymbol{x})}{\Pr(Y=1|S=0)^2}, \tag{44}$$

where $r_1(\boldsymbol{x}) = P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=0}(1|\boldsymbol{x})$ *and* $r_2(\boldsymbol{x}) = P_{Y|X,S=0}(1|\boldsymbol{x})$*;*

- *FPR:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[s_2(X)|S=0\right] s_1(\boldsymbol{x}) - \mathbb{E}\left[s_1(X)|S=0\right] s_2(\boldsymbol{x})}{\Pr(Y=0|S=0)^2}, \tag{45}$$

where $s_1(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=0}(0|\boldsymbol{x})$ *and* $s_2(\boldsymbol{x}) = P_{Y|X,S=0}(0|\boldsymbol{x})$*.*

Proof. **Influence function for DA.** Following from the definition of influence functions, we start with computing $\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\|P_1)$.

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\|P_1) &= \sum_{\boldsymbol{x}'\in\mathcal{X}} ((1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}')) \log \frac{(1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}')}{P_1(\boldsymbol{x}')} \\
&= \sum_{\boldsymbol{x}'\in\mathcal{X}} (P_0(\boldsymbol{x}') + \epsilon(\delta_{\boldsymbol{x}}(\boldsymbol{x}') - P_0(\boldsymbol{x}'))) \\
&\qquad \times \left(\log \frac{P_0(\boldsymbol{x}')}{P_1(\boldsymbol{x}')} + \log\left(1 + \frac{\epsilon(\delta_{\boldsymbol{x}}(\boldsymbol{x}') - P_0(\boldsymbol{x}'))}{P_0(\boldsymbol{x}')}\right)\right) \\
&= \sum_{\boldsymbol{x}'\in\mathcal{X}} (P_0(\boldsymbol{x}') + \epsilon(\delta_{\boldsymbol{x}}(\boldsymbol{x}') - P_0(\boldsymbol{x}'))) \\
&\qquad \times \left(\log \frac{P_0(\boldsymbol{x}')}{P_1(\boldsymbol{x}')} + \epsilon\frac{\delta_{\boldsymbol{x}}(\boldsymbol{x}') - P_0(\boldsymbol{x}')}{P_0(\boldsymbol{x}')} + O(\epsilon^2)\right) \\
&= \mathrm{D}_{\mathrm{KL}}(P_0\|P_1) + \epsilon \sum_{\boldsymbol{x}'\in\mathcal{X}} (\delta_{\boldsymbol{x}}(\boldsymbol{x}') - P_0(\boldsymbol{x}')) \log \frac{P_0(\boldsymbol{x}')}{P_1(\boldsymbol{x}')} + O(\epsilon^2) \\
&= \mathrm{D}_{\mathrm{KL}}(P_0\|P_1) + \epsilon\left(\log \frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_0(X)}{P_1(X)}\middle| S=0\right]\right) + O(\epsilon^2).
\end{aligned}$$

Hence,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\|P_1) - \mathrm{D}_{\mathrm{KL}}(P_0\|P_1)\right) = \log \frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_0(X)}{P_1(X)}\middle| S=0\right]. \tag{46}$$

Similarly, we have

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_{\hat{Y}|S=0} + \epsilon P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}}\|P_{\hat{Y}|S=1}) - \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0}\|P_{\hat{Y}|S=1})\right) \\
&= \sum_{y\in\{0,1\}} ((P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}})(y) - P_{\hat{Y}|S=0}(y)) \log \frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} \\
&= \sum_{y\in\{0,1\}} \log \frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} P_{\hat{Y}|X}(y|\boldsymbol{x}) - \mathbb{E}\left[\log \frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle| S=0\right].
\end{aligned} \tag{47}$$

Combining (46) with (47), we have

$$\begin{aligned}
\psi(\boldsymbol{x}) = &\sum_{y\in\{0,1\}} \log \frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} P_{\hat{Y}|X}(y|\boldsymbol{x}) - \mathbb{E}\left[\log \frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle| S=0\right] \\
&+ \lambda\left(\log \frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_0(X)}{P_1(X)}\middle| S=0\right]\right).
\end{aligned}$$

Note that

$$\log \frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} = \log \frac{P_{X,S}(\boldsymbol{x},0)}{P_{X,S}(\boldsymbol{x},1)} + \log \frac{P_S(1)}{P_S(0)} = \log \frac{P_{S|X}(0|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} + \log \frac{P_S(1)}{P_S(0)}.$$

Hence,

$$\begin{aligned}\log \frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_0(X)}{P_1(X)}\middle| S=0\right] &= \log \frac{P_{S|X}(0|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{P_{S|X}(0|X)}{P_{S|X}(1|X)}\middle| S=0\right] \\ &= \log \frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{1-P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle| S=0\right].\end{aligned}$$

Next,

$$\begin{aligned}&\sum_{y\in\{0,1\}} \log \frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} P_{\hat{Y}|X}(y|\boldsymbol{x}) - \mathbb{E}\left[\log \frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle| S=0\right] \\ &= \log \frac{P_{\hat{Y}|S=0}(1)}{P_{\hat{Y}|S=1}(1)} P_{\hat{Y}|X}(1|\boldsymbol{x}) + \log \frac{P_{\hat{Y}|S=0}(0)}{P_{\hat{Y}|S=1}(0)}(1-P_{\hat{Y}|X}(1|\boldsymbol{x})) \\ &\quad - \mathbb{E}\left[\log \frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle| S=0\right] \\ &= \log \frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)} P_{\hat{Y}|X}(1|\boldsymbol{x}) \\ &\quad + \log \frac{P_{\hat{Y}|S=0}(0)}{P_{\hat{Y}|S=1}(0)} - \log \frac{P_{\hat{Y}|S=0}(0)}{P_{\hat{Y}|S=1}(0)} P_{\hat{Y}|S=0}(0) - \log \frac{P_{\hat{Y}|S=0}(1)}{P_{\hat{Y}|S=1}(1)} P_{\hat{Y}|S=0}(1) \\ &= \left(\log \frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)}\right) P_{\hat{Y}|X}(1|\boldsymbol{x}) - \left(\log \frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)}\right) P_{\hat{Y}|S=0}(1).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\psi(\boldsymbol{x}) = &\left(\log \frac{P_{\hat{Y}|S=0}(1)P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1)P_{\hat{Y}|S=0}(0)}\right)\left(P_{\hat{Y}|X}(1|\boldsymbol{x}) - P_{\hat{Y}|S=0}(1)\right) \\ &+ \lambda\left(\log \frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{1-P_{S|X}(1|X)}{P_{S|X}(1|X)}\middle| S=0\right]\right).\end{aligned}$$

Influence function for CAL. Based on the definition of $d(\boldsymbol{x})$, we have

$$\mathbb{E}\left[\left|P_{Y|X,S=0}(1|X) - P_{\hat{Y}|X}(1|X)\right| \middle| S=0\right] = \sum_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x}) P_0(\boldsymbol{x}).$$

When we perturb the distribution P_0, the function $d(\boldsymbol{x})$ and the quantity

$$\mathbb{E}\left[\left|P_{Y|X,S=1}(1|X) - P_{\hat{Y}|X}(1|X)\right| \middle| S=1\right]$$

do not change. Therefore,

$$\begin{aligned}\psi(\boldsymbol{x}) &= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(\sum_{\boldsymbol{x}'\in\mathcal{X}} d(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}')) - \sum_{\boldsymbol{x}'\in\mathcal{X}} d(\boldsymbol{x}')P_0(\boldsymbol{x}')\right) \\ &= d(\boldsymbol{x}) - \mathbb{E}\left[d(X)|S=0\right].\end{aligned}$$

Influence function for FNR. Next, we compute the influence function of FNR. Similar analysis holds for FPR and FDR.

$$\Pr(\hat{Y}=0|Y=1,S=0)=\frac{\sum_{\boldsymbol{x}'\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x}')P_{Y|X,S=0}(1|\boldsymbol{x}')P_0(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{X}}P_{Y|X,S=0}(1|\boldsymbol{x}')P_0(\boldsymbol{x}')}.$$

Then, following from the definition of $r_1(\boldsymbol{x})$ and $r_2(\boldsymbol{x})$, we have

$$\Pr(\hat{Y}=0|Y=1,S=0)=\frac{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_1(\boldsymbol{x}')P_0(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_2(\boldsymbol{x}')P_0(\boldsymbol{x}')}=\frac{\mathbb{E}\left[r_1(X)|S=0\right]}{\mathbb{E}\left[r_2(X)|S=0\right]},$$

which implies

$$\begin{aligned}
&\mathsf{M}((1-\epsilon)P_0+\epsilon\delta_{\boldsymbol{x}},P_1)\\
&=\frac{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_1(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}')+\epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}'))}{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_2(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}')+\epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}'))}-\Pr(\hat{Y}=0|Y=1,S=1)\\
&=\frac{\mathbb{E}\left[r_1(X)|S=0\right]+\epsilon\left(r_1(\boldsymbol{x})-\mathbb{E}\left[r_1(X)|S=0\right]\right)}{\mathbb{E}\left[r_2(X)|S=0\right]+\epsilon\left(r_2(\boldsymbol{x})-\mathbb{E}\left[r_2(X)|S=0\right]\right)}-\Pr(\hat{Y}=0|Y=1,S=1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\psi(\boldsymbol{x})&=\lim_{\epsilon\to0}\frac{1}{\epsilon}\left(\mathsf{M}((1-\epsilon)P_0+\epsilon\delta_{\boldsymbol{x}},P_1)-\mathsf{M}(P_0,P_1)\right)\\
&=\frac{\mathbb{E}\left[r_2(X)|S=0\right]r_1(\boldsymbol{x})-\mathbb{E}\left[r_1(X)|S=0\right]r_2(\boldsymbol{x})}{\mathbb{E}\left[r_2(X)|S=0\right]^2}\\
&=\frac{\mathbb{E}\left[r_2(X)|S=0\right]r_1(\boldsymbol{x})-\mathbb{E}\left[r_1(X)|S=0\right]r_2(\boldsymbol{x})}{\Pr(Y=1|S=0)^2}.
\end{aligned}$$

□

C.2 Estimating Influence Functions

We show how to estimate influence functions given an auditing dataset from the target population. Here the input is

- $\mathcal{D}^{\text{audit}}=\{(\boldsymbol{x}_i,y_i,s_i)\}_{i=1}^n$: an auditing dataset;
- $h(\boldsymbol{x})$: a fixed classifier that we wish to audit;
- $P_{S|X}(1|\boldsymbol{x})$: a conditional distribution of the sensitive attribute given features;
- $P_{Y|X,S=0}(1|\boldsymbol{x})$: a conditional distribution of the outcome given features for the minority group.

We divide the auditing dataset $\mathcal{D}^{\text{audit}}$ into two parts: $\mathcal{D}_0^{\text{audit}}$ and $\mathcal{D}_1^{\text{audit}}$ where $\mathcal{D}_i^{\text{audit}}$ contains all samples from the auditing dataset with $S=i$. We denote n_i as the number of samples in $\mathcal{D}_i^{\text{audit}}$. First we show how to compute influence functions from $\mathcal{D}^{\text{audit}}$:

- DA:

$$\psi(\boldsymbol{x})=\left(\log\frac{\hat{p}_0(1-\hat{p}_1)}{\hat{p}_1(1-\hat{p}_0)}\right)(h(\boldsymbol{x})-\hat{p}_0)+\lambda\left(\log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}-\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}}\log\frac{1-P_{S|X}(1|\boldsymbol{x}')}{P_{S|X}(1|\boldsymbol{x}')}\right),\tag{48}$$

  where $\hat{p}_0=\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}}h(\boldsymbol{x}')$ and $\hat{p}_1=\frac{1}{n_1}\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}}h(\boldsymbol{x}')$;

- CAL:

$$\psi(\boldsymbol{x}) = d(\boldsymbol{x}) - \frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} d(\boldsymbol{x}'), \tag{49}$$

  where $d(\boldsymbol{x}) = \left| P_{Y|X,S=0}(1|\boldsymbol{x}) - h(\boldsymbol{x}) \right|$;

- FDR:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} t_2(\boldsymbol{x}')\right) t_1(\boldsymbol{x}) - \left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} t_1(\boldsymbol{x}')\right) t_2(\boldsymbol{x})}{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} t_2(\boldsymbol{x}')\right)^2}, \tag{50}$$

  where $t_1(\boldsymbol{x}) = h(\boldsymbol{x})(1 - P_{Y|X,S=0}(1|\boldsymbol{x}))$ and $t_2(\boldsymbol{x}) = h(\boldsymbol{x})$;

- FNR:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} r_2(\boldsymbol{x}')\right) r_1(\boldsymbol{x}) - \left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} r_1(\boldsymbol{x}')\right) r_2(\boldsymbol{x})}{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} r_2(\boldsymbol{x}')\right)^2}, \tag{51}$$

  where $r_1(\boldsymbol{x}) = (1 - h(\boldsymbol{x}))P_{Y|X,S=0}(1|\boldsymbol{x})$ and $r_2(\boldsymbol{x}) = P_{Y|X,S=0}(1|\boldsymbol{x})$;

- FPR:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} s_2(\boldsymbol{x}')\right) s_1(\boldsymbol{x}) - \left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} s_1(\boldsymbol{x}')\right) s_2(\boldsymbol{x})}{\left(\frac{1}{n_0} \sum_{\boldsymbol{x}' \in \mathcal{D}_0^{\text{audit}}} s_2(\boldsymbol{x}')\right)^2}, \tag{52}$$

  where $s_1(\boldsymbol{x}) = h(\boldsymbol{x})(1 - P_{Y|X,S=0}(1|\boldsymbol{x}))$ and $s_2(\boldsymbol{x}) = 1 - P_{Y|X,S=0}(1|\boldsymbol{x})$.

Appendix D. Supporting Material for Experimental Results

D.1 Synthetic Data

Setup

We consider a simple experiment to show that the preprocessor mitigates discrimination while removing a single proxy variable does not. We consider a setting where $X = (X_1, X_2, X_3) \in \{-1, 1\}^3$ and choose the joint distribution matrices of (X_1, X_2) for $S = 0$ and $S = 1$ as

$$\mathbf{P}_0 = \begin{pmatrix} 0.60 & 0.00 \\ 0.25 & 0.15 \end{pmatrix}, \mathbf{P}_1 = \begin{pmatrix} 0.05 & 0.00 \\ 0.20 & 0.75 \end{pmatrix}. \tag{53}$$

Then we choose X_3 to be independent of (X_1, X_2) with $\Pr(X_3 = 1|S = i) = 0.3$ for $i = 0, 1$. We draw the values of Y according to $P_{Y|X,S=i}(1|\boldsymbol{x}) = \mathsf{logistic}(6x_1x_2 + x_3)$ for $i = 0, 1$, and fit a logistic regression using 50k samples.

Results

The value of DA with $\lambda = 0$ is 14.0%. In this case, both X_1 and X_2 are proxy variable. We remove X_1 from dataset and retrain a logistic regression as a classifier. It turns out that the value of DA becomes larger: 24.8%. This is because the pair (X_1, X_2) is a joint proxy and, consequently, removing one of them could not reduce discrimination.

Next, we apply Algorithm 1 and the proposed preprocessor to decrease discrimination. For the sake of example, we randomly draw 12.5k new samples for the auditing dataset and 12.5k samples for the holdout dataset, and apply the descent procedure in Algorithm 1 under DA. At each step, the influence function is computed on the auditing dataset, and applied to both the auditing and the holdout set. We show the values of DA with each iteration in Figure 2. Then we use the preprocessor to map samples from $S = 0$ to new samples and DA becomes 0.0%.

D.2 Synthetic Data

We apply Algorithm 1 to a toy example in order to show the descent of the discrimination metric for each iteration in Figure 4. Here, $X = (X_1, X_2, X_3)$ where X_i is a binary random variable with $\Pr(X_i = 1|S = 0) = p_i$ with $(p_1, p_2, p_3) = (0.9, 0.2, 0.2)$, $\Pr(X_i = 1|S = 1) = q_i$ with $(q_1, q_2, q_3) = (0.1, 0.5, 0.5)$, and $P_S(1) = 0.5$. We draw the values of Y according to $P_{Y|X,S=0}(1|\boldsymbol{x}) = P_{Y|X,S=1}(1|\boldsymbol{x}) = \mathsf{logistic}(5x_1 - 2x_2 - 2x_3)$, and fit a logistic regression over 50k samples. We randomly draw 12.5k samples for the auditing dataset and 12.5k samples for the holdout dataset, and apply the descent procedure in Algorithm 1 for the FPR metric. At each step, the influence function is computed on the auditing dataset, and applied to both the auditing and the holdout set. TAs shown in Figure 4, the procedure converges to a counterfactual distribution after around 40 iterations (we show additional steps for the sake of illustration). In practice, a stopping rule can be designed to stop the descent procedure based on number of iterations or a target discrimination gap value.

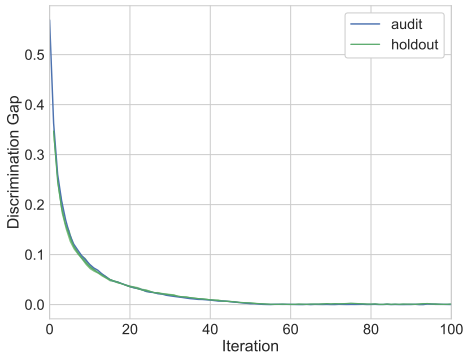

Figure 4: Values of FPR for auditing dataset (blue) and holdout dataset (green), respectively, with each iteration in corrective distributional descent for a synthetic dataset.

Then we use the proposed preprocessor to map samples from $S = 0$ to new samples. Then the value of FPR decreases from 29.1% to 4.1%.

D.3 Real-World Datasets

We show additional experimental results on real-world datasets in this section.

Understanding Discrimination

As discussed in Section 4.2, influence functions can also be used to score proxy features and prototypes. We give their formal definitions as follows.

Definition 6. Let the input to the classifier be given by $X = (X_1, \ldots, X_d)$. For a given discrimination metric, the proxy score for feature X_1 is defined as

$$\gamma_1 \triangleq \sum_{x_2, \cdots, x_d} \delta_\psi(x_2, \cdots, x_d) \Pr(X_2 = x_2, \cdots, X_d = x_d | S = 0),$$

where the function $\delta_\psi(x_2, \cdots, x_d)$ measures the change of the influence function ψ with respect to the first feature. The proxy score for the remaining variables $X_2, \ldots, X_d$ is defined equivalently. It can also be generalized to measure the variation of the influence functions with respect to more than one given feature.

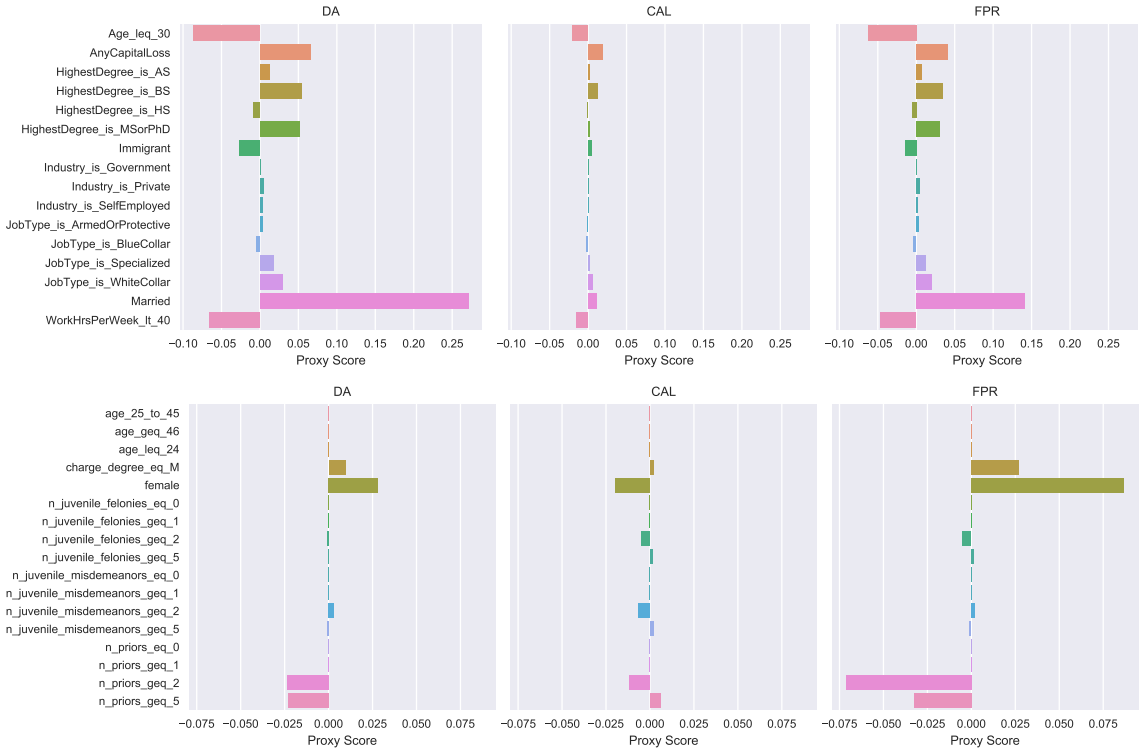

Figure 5: Proxy scores for variables in the `adult` dataset (top) and `compas` dataset (bottom). We show scores computed using the influence functions for DA (left), CAL (middle) and FPR (right).

For example, one can choose $\delta_\psi(x_2, \cdots, x_d) = \max_{x_1, x_1' \in \mathcal{X}_1} |\psi(x_1, \cdots, x_d) - \psi(x_1', \cdots, x_d)|$ with $\mathcal{X}_1$ the support set of X_1 or, alternatively, $\delta_\psi(x_2, \cdots, x_d) = \psi(1, \cdots, x_d) - \psi(0, \cdots, x_d)$ when the features are binary.

We compute the values of proxy scores for all input variables and show the results in Figure 5.

Next, we show that influence functions can be used for identifying prototypes. Specifically, we compute the argmax of the influence functions and show the results in Table 4 and Table 5.

	DA	FNR	CAL
Married	0	1	0
Immigrant	0	0	1
HighestDegree_is_HS	0	0	1
HighestDegree_is_AS	0	0	0
HighestDegree_is_BS	1	0	0
HighestDegree_is_MSorPhD	0	1	0
AnyCapitalLoss	0	0	0
Age_leq_30	1	0	1
WorkHrsPerWeek_lt_40	0	1	1
JobType_is_WhiteCollar	0	0	0
JobType_is_BlueCollar	1	0	0
JobType_is_Specialized	0	0	0
JobType_is_ArmedOrProtective	0	0	0
Industry_is_Private	1	0	1
Industry_is_Government	0	0	0
Industry_is_SelfEmployed	0	0	0

Table 4: We show the features of minority entries in `adult` dataset that attain the maximum values of the influence function.

	DA	CAL	FPR
age_leq_24	1	0	0
age_25_to_45	0	1	1
age_geq_46	0	0	0
female	0	0	0
n_priors_eq_0	0	0	0
n_priors_geq_1	1	1	1
n_priors_geq_2	1	1	1
n_priors_geq_5	1	0	1
n_juvenile_misdemeanors_eq_0	0	1	1
n_juvenile_misdemeanors_geq_1	1	0	0
n_juvenile_misdemeanors_geq_2	0	0	0
n_juvenile_misdemeanors_geq_5	0	0	0
n_juvenile_felonies_eq_0	0	0	0
n_juvenile_felonies_geq_1	1	1	1
n_juvenile_felonies_geq_2	1	1	1
n_juvenile_felonies_geq_5	1	0	0
charge_degree_eq_M	0	0	0

Table 5: We show the features of minority entries in `compas` dataset that attain the maximum values of the influence function.