Auditing Discrimination with Influence Functions

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

Disparate impact in machine learning can occur when the distribution of the input variables of a predictive model changes significantly conditioned on a sensitive attribute (e.g., race or gender). In this paper, we introduce a new framework to audit disparate impact of a predictive model. Our framework aims to determine features that act as "proxies" of a sensitive attribute and cause the disparity observed at the model's output. We identify these proxies by characterizing the impact of small perturbations of the input distribution on some common discrimination metrics, which can be expressed in terms of the influence functions from robust statistics. We illustrate the value and technical feasibility of our framework through numerical experiments on synthetic and real-world datasets.

1. Introduction

The performance of machine learning models changes when the distribution of the input data is perturbed. In applications of individual-level consequence, these statistical variations may lead to uneven performance over different populations, resulting in a *disparate impact* across demographic groups. More specifically, disparate impact occurs when a sensitive attribute (e.g., race, gender) is omitted from the model, but still affects its predictions through correlations with input features (e.g., income, education level).

If a given feature is heavily weighed in the predictive model's decisions, and its distribution changes significantly between a majority and a minority group, then this feature acts as a potential "proxy" for discrimination. Consider, for example, a logistic regression model to predict recidivism that assigns a large weight to the number of prior convictions. A larger prevalence of individuals with prior convictions in the minority population may then result in a difference in the output distributions with respect to the majority group. Consequently, a small change in the distribution of this group would help to equalize the output scores between the majority and minority populations. One can repeatedly perturb input distributions to reach a *counterfactual distribution* under which there is less discrimination (see Figure 1 for further illustration).

In this paper, we use tools from information theory and robust statistics to systematically quantify which features (or combinations thereof) may be driving the disparity observed at the output of a fixed predictive model. We achieve this by characterizing the impact of small perturbations of the input distribution under different discrimination metrics. For a fixed model, features that have their distribution changed the most are potential proxies for discrimination. Such proxy features will depend on the discrimination metric, the underlying distribution, and the predictive model used. Our main contributions are three-fold:

1. We provide a general theoretical framework to audit predictive models that may produce disparate impact. Our framework quantifies how small perturbations of the input distribution over the minority (or majority) group affects a given discrimination metric. The direction of perturbation that decreases a given discrimination metric the most is called the *correction function*, which, in

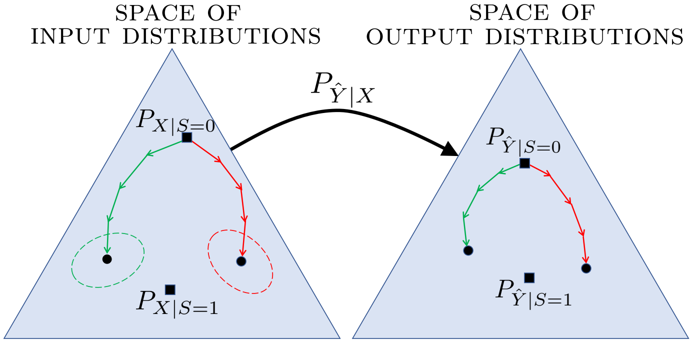

Figure 1: Illustration of correction path on the probability simplex. The distributions within the green and the red ellipses are counterfactual distributions under different discrimination metrics. The green line and red line represent paths to the counterfactual distributions.

turn, assigns a score for each element in the input space in terms of its influence on the disparate impact.

In Proposition 2, we demonstrate that the correction functions from an information-theoretic view of fairness are *entirely equivalent* (up to a scaling factor) to influence functions in the statistics literature (see e.g., Hampel et al., 2011; Huber, 2011).

2. We prove that influence functions (and the corresponding correction functions) for some common discrimination metrics can be cast in closed-form in terms of (i) the predictive model being audited, (ii) an equivalent predictive model trained only over the minority group, and (iii) a model that predicts membership in the sensitive group[1]. Given that the last two components are reasonably accurate, influence functions can be computed directly using the expressions derived in Proposition 4. Generalization results for influence functions computed using this method are presented in Proposition 5.

3. We validate our framework through numerical experiments derived from synthetic and real-world datasets.

Related Work

Disparate impact in prediction and classification has been documented in applications as varied as loan approval (Hunt, 2005), recidivism prediction (Angwin et al., 2016) and facial recognition (Buolamwini and Gebru, 2018). Disparate impact can be quantified using discrimination and fairness metrics that capture its many facets, such as predictive parity (Corbett-Davies et al., 2017), error rate imbalance (Chouldechova, 2017), similarity of output distributions (Feldman et al., 2015), and calibration error (Pleiss et al., 2017). Different tools, such as statistical tests (Simoiu et al., 2017; Pierson et al., 2017) and causal reasoning (Kilbertus et al., 2017; Kusner et al., 2017; Galhotra et al., 2017), are used to detect disparate impact. Another fairness definition is the notion of individual fairness (Dwork et al., 2012) that requires that any two individuals with similar background should be treated similarly.

Influence functions have been studied both in robust statistics (see e.g., Hampel et al., 2011; Huber, 2011) and in semiparametric statistics (see e.g., Robins et al., 2009). Recently, influence functions have been used to understand the effect of training data on black-box predictions (Koh and Liang, 2017), to analyze the robustness of machine learning models (Christmann and Steinwart,

[1] We understand that the last component may be problematic (e.g., a predictor for race or gender) and we do not condone such models being trained outside of the context considered here.

Performance Metric	Discrimination Metric
Distribution Alignment (DA)	$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}\|S=0} \\| P_{\hat{Y}\|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(P_{X\|S=0} \\| P_{X\|S=1})$
Calibration Error (CAL)	$\mathbb{E}\left[\left\|P_{Y\|X,S=0}(1\|X) - P_{\hat{Y}\|X}(1\|X)\right\| \,\middle\|\, S=0\right]$ $-\mathbb{E}\left[\left\|P_{Y\|X,S=1}(1\|X) - P_{\hat{Y}\|X}(1\|X)\right\| \,\middle\|\, S=1\right]$
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 for different performance objectives. We assume $S = 0$ has the less favorable value of the performance. Each metric corresponds to a different set of counterfactual distributions in Figure 1. Our framework generalizes to other metrics considered in the literature (see e.g., Berk et al., 2017, for a list). Distribution Alignment (DA) is a new metric related to the divergence in output distributions. It measures the statistical indistinguishability of two populations (see Appendix C for connections between DA and binary hypothesis testing).

2004) and to do fast cross-validation (Liu et al., 2014). The influence of local perturbations on probability distributions has also been studied in information theory (see e.g., Borade and Zheng, 2008). Auditing a black-box model to quantify the influence of a feature (or a group of features) has been studied using different tools, such as perturbing the test point (Adler et al., 2018) and learning a simpler model locally around the prediction (Ribeiro et al., 2016).

2. Framework

We consider the task of auditing a fixed predictive model $P_{\hat{Y}|X}(1|\boldsymbol{x})$, which takes as input a vector of d random variables $X = (X_1, \cdots, X_d) \in \mathcal{X}$ with distribution P_X and outputs a belief that the predicted binary output $\hat{Y}$ is 1. For example:

- If $P_{\hat{Y}|X}(1|\boldsymbol{x})$ outputs a predicted class directly (e.g., SVM), then $P_{\hat{Y}|X}(1|\boldsymbol{x}) \in \{0,1\}$;
- If $P_{\hat{Y}|X}(1|\boldsymbol{x})$ outputs a predicted probability (e.g., logistic regression), then $P_{\hat{Y}|X}(1|\boldsymbol{x}) \in [0,1]$.

We aim to characterize differences in the output distribution of $P_{\hat{Y}|X}(1|\boldsymbol{x})$ with respect to a *sensitive attribute* S. We focus on the case where the sensitive attribute is binary $S \in \{0,1\}$, and use $P_{X|S=0}, P_{X|S=1}$ and $P_{\hat{Y}|S=0}(y) = \mathbb{E}\left[P_{\hat{Y}|X}(y|X)|S=0\right]$, $P_{\hat{Y}|S=1}(y) = \mathbb{E}\left[P_{\hat{Y}|X}(y|X)|S=1\right]$ to denote the conditional distributions of inputs and predicted outputs, respectively.

We assume that $P_{\hat{Y}|X}(1|\boldsymbol{x})$ does not use the sensitive attribute S, as doing so would violate legal constraints in applications such as hiring and credit scoring (see e.g., Barocas and Selbst, 2016). In this setting, the Markov condition $S \to X \to \hat{Y}$ ensures that $P_{\hat{Y}|S=0} = P_{\hat{Y}|X} \circ P_{X|S=0}$ and $P_{\hat{Y}|S=1} = P_{\hat{Y}|X} \circ P_{X|S=1}$. Thus, disparate impact may occur when $P_{X|S=0} \neq P_{X|S=1}$.

Lastly, we denote the true underlying outcome variable by Y, potentially accessible only through samples.

Given the distributions $P_{\hat{Y}|X}$, $P_{Y|X,S}$, and P_S, many measures of discrimination can be expressed in terms of $P_{X|S=0}$ and $P_{X|S=1}$, as shown in Table 1. We denote the discrimination metrics in the right-hand column of Table 1 by $\mathsf{M}(P_{X|S=0}, P_{X|S=1})$.

2.1 Counterfactual Distributions

The *counterfactual distributions* are distributions which minimize the given discrimination metric.

Definition 1. For a given discrimination metric $\mathsf{M}(P_{X|S=0}, P_{X|S=1})$, the counterfactual distributions are distributions contained in the following set

$$\left\{Q_X \;\middle|\; \mathsf{M}(Q_X, P_{X|S=1}) = \inf_{\mathsf{supp}(Q'_X)\subseteq\mathsf{supp}(P_{X|S=0})} \mathsf{M}(Q'_X, P_{X|S=1}), \mathsf{supp}(Q_X) \subseteq \mathsf{supp}(P_{X|S=0})\right\}. \quad (1)$$

The counterfactual distributions are global optimal distributions which minimize the given discrimination metric. The following proposition characterizes the properties of the set of counterfactual distributions (see Appendix A for the proof).

Proposition 1. *For a given discrimination metric $\mathsf{M}(P_{X|S=0}, P_{X|S=1})$ in Table 1, the set of counterfactual distributions, defined in (1), is a non-empty, closed, convex set.*

In what follows, we take a step back from the global optimality and introduce correction/influence functions which capture the local behavior of the discrimination metrics. We will show how to iteratively perturb the dataset using influence functions to recover a counterfactual distribution in Section 4.1.

2.2 Influence Functions

We investigate next how the discrimination metrics displayed in Table 1 decrease when the distribution $P_{X|S=0}$ is slightly perturbed. The local perturbation of a distribution is defined as follows.

Definition 2. The perturbed distribution $\widetilde{P}_{X|S=0}$ is defined as

$$\widetilde{P}_{X|S=0}(\boldsymbol{x}) \triangleq P_{X|S=0}(\boldsymbol{x})(1 + \epsilon f(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in \mathcal{X} \quad (2)$$

where $f : \mathcal{X} \to \mathbb{R}$ is a perturbation function that belongs to the class of all functions with zero mean and unit variance

$$\mathcal{L}_{S=0} \triangleq \left\{f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}\left[f(X)|S=0\right] = 0, \mathbb{E}\left[f(X)^2|S=0\right] = 1\right\}, \quad (3)$$

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

The *correction function* is a perturbation function of steepest descent with respect to a given discrimination metric.

Definition 3. The correction function $\phi : \mathcal{X} \to \mathbb{R}$ is defined as

$$\phi(\boldsymbol{x}) \triangleq \operatorname*{argmin}_{f(\boldsymbol{x})\in\mathcal{L}_{S=0}} \lim_{\epsilon\to 0} \frac{\mathsf{M}(\widetilde{P}_{X|S=0}, P_{X|S=1}) - \mathsf{M}(P_{X|S=0}, P_{X|S=1})}{\epsilon}. \quad (4)$$

Perturbing the distribution $P_{X|S=0}$ along the correction function results in the largest local decrease of the discrimination metric. The correction function can be viewed as a normalized *influence function* (Huber, 2011; Koh and Liang, 2017) up to a sign difference as shown in Proposition 2 (see Appendix A for the proof).

Definition 4. The influence function $\psi : \mathcal{X} \to \mathbb{R}$ is defined as

$$\psi(\boldsymbol{x}) \triangleq \lim_{\epsilon\to 0} \frac{\mathsf{M}\left((1-\epsilon)P_{X|S=0} + \epsilon\delta_{\boldsymbol{x}}, P_{X|S=1}\right) - \mathsf{M}(P_{X|S=0}, P_{X|S=1})}{\epsilon}, \quad (5)$$

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

Proposition 2. *For a given discrimination metric* $\mathsf{M}(P_{X|S=0}, P_{X|S=1})$*, we have that*

$$\phi(\boldsymbol{x}) = \frac{-\psi(\boldsymbol{x})}{\sqrt{\mathrm{Var}\left[\psi(X)|S=0\right]}}, \tag{6}$$

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

The influence function approximates the change of a discrimination metric if a sample from $P_{X|S=0}$ is removed (or added) in a very large dataset. Correction functions and influence functions are conceptually different: influence functions capture the effect of samples while correction functions indicate the direction of steepest descent in the probability simplex. Nevertheless, Proposition 2 shows that these two functions are equivalent (up to a scaling factor).

2.3 Choice of Metrics

A combination of multiple metrics can be used to assess discrimination (see e.g., Zafar et al., 2017). This is of particular interest since there is a trade-off between different metrics (Kleinberg et al., 2016; Chouldechova, 2017), and it may be impossible to satisfy multiple notions of fairness simultaneously. Hence, a linear combination of metrics can be used as a single compound metric to understand the trade-offs between potentially competing fairness objectives. The following proposition demonstrates that, in this case, the resulting influence function (and correction function) for the compound metric is also a linear combination of the influence functions for each individual metric.

Proposition 3. *Given a compound discrimination metric which is a linear combination of* K *different discrimination metrics:* $\mathsf{M}(P_{X|S=0}, P_{X|S=1}) = \sum_{i=1}^{K} \lambda_i \mathsf{M}_i(P_{X|S=0}, P_{X|S=1})$*, the influence function can be computed by* $\psi(\boldsymbol{x}) = \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x})$*. Furthermore, the correction function is*

$$\phi(\boldsymbol{x}) = \sum_{i=1}^{K} \frac{-\lambda_i}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^{K} \lambda_i \psi_i(X)|S=0\right]}} \psi_i(\boldsymbol{x}). \tag{7}$$

3. Computing Influence Functions

In what follows, we present the closed-form expressions of influence functions for discrimination metrics in Table 1. The expressions depend on three functions:

- $P_{\hat{Y}|X}(1|\boldsymbol{x})$, a fixed predictive model 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 and $S=0$.

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

Proposition 4. *Influence functions for the discrimination metrics in Table 1 can be expressed as*

- *Distribution Alignment:*

$$\begin{aligned}\psi(\boldsymbol{x}) = &\left(\log \frac{P_{\hat{Y}|S=0}(1) P_{\hat{Y}|S=1}(0)}{P_{\hat{Y}|S=1}(1) P_{\hat{Y}|S=0}(0)}\right) \left(P_{\hat{Y}|X}(1|\boldsymbol{x}) - P_{\hat{Y}|S=0}(1)\right) \\ &+ \lambda \left(\log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \middle| S=0\right]\right);\end{aligned} \tag{8}$$

- *Calibration Error:*

$$\psi(\boldsymbol{x}) = h_C(\boldsymbol{x}) - \mathbb{E}\left[h_C(X)|S=0\right], \tag{9}$$

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

- *False Discovery Rate:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[t_2(X)|S=0\right] t_1(\boldsymbol{x}) - \mathbb{E}\left[t_1(X)|S=0\right] t_2(\boldsymbol{x})}{\Pr(\hat{Y}=1|S=0)^2}, \tag{10}$$

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

- *False Negative Rate:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[r_2(X)|S=0\right] r_1(\boldsymbol{x}) - \mathbb{E}\left[r_1(X)|S=0\right] r_2(\boldsymbol{x})}{\Pr(Y=1|S=0)^2}, \tag{11}$$

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

- *False Positive Rate:*

$$\psi(\boldsymbol{x}) = \frac{\mathbb{E}\left[s_2(X)|S=0\right] s_1(\boldsymbol{x}) - \mathbb{E}\left[s_1(X)|S=0\right] s_2(\boldsymbol{x})}{\Pr(Y=0|S=0)^2}, \tag{12}$$

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

A proof of the previous proposition, together with expressions for estimating influence functions directly from $\mathcal{D}^{\text{audit}}$, is presented in Appendix B.

4. Optimization and Generalization

4.1 Beyond Local Perturbations

Influence functions produce local perturbations that correspond to the steepest reduction of a given discrimination metric. The closed-form expressions from the previous section can be used to estimate influence functions directly from an auditing dataset. In Algorithm 1, we iteratively perturb the dataset using influence functions to recover a counterfactual distribution. Specifically, for each iteration, we compute the influence function $\psi(\boldsymbol{x})$ and resample points for the minority group $S=0$ in the auditing dataset with weights $1-\epsilon\psi(\boldsymbol{x})$.

We apply this algorithm to a toy example in order to show the descent of the discrimination metric for each iteration in Figure 2. 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 $P_{Y|X,S=0}(\boldsymbol{x}) = P_{Y|X,S=1}(\boldsymbol{x}) = \mathsf{logit}(5x_1 - 2x_2 - 2x_3)$, and fit a logistic regression over 50k samples. Naturally, the model accurately recovers the coefficients for each x_i. For the sake of example, we randomly draw fresh 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. As shown in Figure 2, the procedure converges to a non-discriminatory counterfactual distribution in around 20 iterations (we show further 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. The descent procedure on real-world data is illustrated in Section 5.

4.2 Generalization Bounds

The closed-form expressions of influence functions depend on the predictive model and the conditional distributions $P_{S|X}(1|\boldsymbol{x})$ and $P_{Y|X,S=0}(1|\boldsymbol{x})$. Here we let the predictive model for the outcome be

Algorithm 1 Corrective Distributional Descent

```
Input:
    $\epsilon > 0$    ▷ step size
    $\mathsf{M}(\cdot)$    ▷ discrimination metric
    $P_{\hat{Y}|X}(1|\boldsymbol{x})$    ▷ predictive model for the outcome
    $P_{S|X}(1|\boldsymbol{x})$    ▷ conditional distribution of the sensitive attribute given features
    $P_{Y|X,S=0}(1|\boldsymbol{x})$    ▷ conditional distribution of the outcome given features and $S = 0$
    $\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{n}$    ▷ auditing dataset
    $\mathcal{D}^{\text{holdout}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=n+1}^{m}$    ▷ holdout dataset
Initialize:
$c(\boldsymbol{x}) \leftarrow 1$
$\mathcal{D} \leftarrow \mathcal{D}^{\text{audit}}$
$\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathcal{D}^{\text{holdout}}\right)$    ▷ compute the discrimination metric on holdout dataset
repeat
    $\psi(\boldsymbol{x}) \leftarrow$ compute the influence function from $\mathcal{D}$ using Proposition 4
    $\mathcal{D} \leftarrow \mathsf{Resample}(\mathcal{D}, 1 - \epsilon\psi(\boldsymbol{x}))$
    $c(\boldsymbol{x}) \leftarrow (1 - \epsilon\psi(\boldsymbol{x}))c(\boldsymbol{x})$
    $\mathsf{M}_{\text{old}} \leftarrow \mathsf{M}_{\text{new}}$
    $\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathsf{Resample}\left(\mathcal{D}^{\text{holdout}}, c(\boldsymbol{x})\right)\right)$
until $\mathsf{M}_{\text{new}} \geq \mathsf{M}_{\text{old}}$
return: $c(\boldsymbol{x})$    ▷ $c(\boldsymbol{x})$ is an aggregate multiplication of perturbations
```

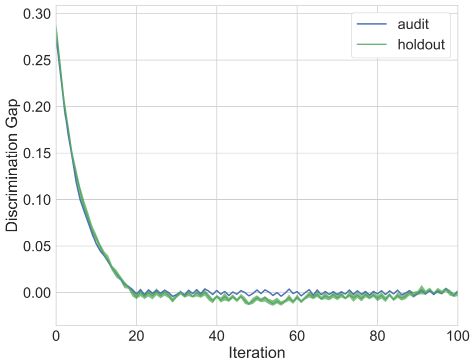

Figure 2: Values of FPR with each iteration in corrective distributional descent for a synthetic dataset.

fixed, and estimate the conditional distributions using classifiers trained over an auditing dataset. These classifiers may be imprecise due to, for example, a limited number of samples in the auditing data, or perhaps due to limited capacity of the classification models used.

We show in the following proposition, that the difference between the estimated influence functions and the true influence functions is fully determined by how accurately the capacity of estimating the conditional distributions $P_{S|X}(1|\boldsymbol{x})$ and $P_{Y|X,S=0}(1|\boldsymbol{x})$ (see Appendix A for the proof).

Proposition 5. *Let $\widehat{\psi}(\boldsymbol{x})$ and $\psi(\boldsymbol{x})$ be the estimated influence function and the true influence function, respectively. Then*

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_p \lesssim \begin{cases} \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_p & DA, \\ \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p & \text{other metrics in Table 1}, \end{cases} \tag{13}$$

where we define $\|f(\boldsymbol{x}) - g(\boldsymbol{x})\|_p \triangleq \left(\mathbb{E}\left[|f(X) - g(X)|^p|S=0\right]\right)^{1/p}$ and let $p \geq 1$. Furthermore, if $\widehat{P}_{S|X}$ and $\widehat{P}_{Y|X,S=0}$ are the empirical conditional distributions obtained from m i.i.d. samples, then, with probability at least $1-\beta$,

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_1 \lesssim \sqrt{\frac{1}{m}\left(|\mathcal{X}| - \log\beta\right)}. \tag{14}$$

5. Numerical Experiments

Setup

We consider the `Adult` dataset from the UCI repository (Kohavi, 1996). We use a processed version of the dataset with $n = 32,561$ records and $d = 16$ binary variables. Here, $y_i = 1$ if individual i earns over $\$50k$ with $\Pr(y = +1) = 24.1\%$. The sensitive attribute is gender, for $S = 0$ if *Female* and $S = 1$ if *Male*.

We train a hypothetical predictive model $P_{\hat{Y}|X}(1|\boldsymbol{x})$ using 30% of the dataset. We then use the remaining 70% sample to train the required models for the auditing procedure, namely

- $P_{S|X}(1|\boldsymbol{x})$, 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.

For illustration purposes, we fit all three classifiers using vanilla logistic regression and use a standard nested 10-CV setup to estimate their performance. The mean 10-CV test AUC for the models are: 0.866 ± 0.01 for $P_{\hat{Y}|X}(1|\boldsymbol{x})$; 0.854 ± 0.011 for $P_{Y|X,S=0}(1|\boldsymbol{x})$; and 0.849 ± 0.036 for $P_{S|X}(1|\boldsymbol{x})$.

Although $P_{\hat{Y}|X}(1|\boldsymbol{x})$ does not use S as an input, it has significant disparate impact over the groups. Specifically, the mean predicted outcome $P_{\hat{Y}|X}(1|\boldsymbol{x})$ is 12.5% for females and 29.4% for males. The values of DA, CAL and FPR are 0.845 %, 13.1%, and 16.8% respectively.

Evaluating Proxy Variables

The influence function assigns a score for all possible outcomes of X (i.e., every point in $\mathcal{X}$). We can estimate the influence of a single feature by using an aggregate *proxy score*. The proxy score captures the variation of the influence function with respect to a given feature and is defined next.

Definition 5. 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.

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.

In Figure 3, we show the values of proxy scores from Definition 5 for all input variables in the dataset.

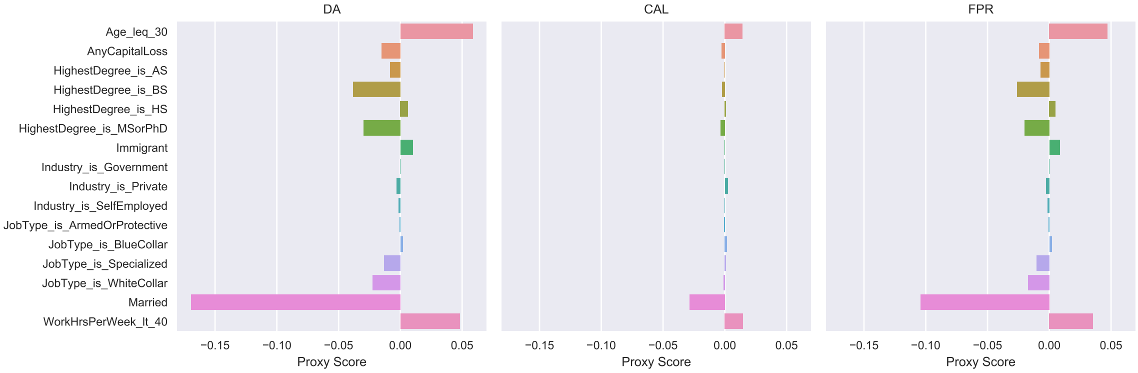

Figure 3: Proxy scores for variables in the `Adult` dataset. We show scores computed using the influence functions for DA (left), CAL (middle) and FPR (right).

Recovering Counterfactual Distributions

Since proxy scores are derived from influence functions, they reflect local information with respect to discrimination. Even as these metrics may not appear to differ significantly, small differences in these initial directions can lead to different counterfactual distributions. In Table 2, we show the counterfactual distributions recovered using Algorithm 1 to the input distribution for the minority class using influence functions for DA, CAL and FPR (using 20% of the auditing dataset or $n = 4558$ examples).

6. Discussion

We have proposed a theoretical framework to audit predictive models for discrimination. Our framework is based on correction/influence functions, which indicate the steepest descent for a given discrimination metric under locally small perturbations of the input distribution. When applied to real-world data, influence functions reveal potential proxies for discrimination. The closed-form expressions for influence functions allow them to be efficiently computed over an auditing dataset under different discrimination metrics. The procedure used to derive these expressions can be replicated for a range of other discrimination metrics.

	Majority	Minority	Counterfactual Distribution		
	$P_{X\|S=1}$	$P_{X\|S=0}$	DA	CAL	FPR
Married	56.4	16.7	29.0	21.9	46.3
Immigrant	8.3	8.3	10.1	8.8	12.3
HighestDegree_is_HS	29.4	35.2	30.2	34.2	24.6
HighestDegree_is_AS	9.2	8.3	7.9	8.4	8.5
HighestDegree_is_BS	14.2	16.7	23.4	17.5	26.5
HighestDegree_is_MSorPhD	6.0	5.6	8.0	5.3	10.1
AnyCapitalLoss	4.1	2.8	3.3	2.7	3.0
Age_leq_30	29.8	36.1	29.8	33.4	20.9
WorkHrsPerWeek_lt_40	19.7	40.7	34.4	37.5	24.7
JobType_is_WhiteCollar	21.6	35.2	41.2	35.6	45.2
JobType_is_BlueCollar	28.0	4.6	4.0	4.7	2.9
JobType_is_Specialized	21.6	19.4	20.4	20.0	23.5
JobType_is_ArmedOrProtective	1.4	0.9	0.8	1.0	0.7
Industry_is_Private	68.8	67.6	63.5	64.7	56.5
Industry_is_Government	11.9	13.0	14.3	12.4	13.0
Industry_is_SelfEmployed	13.8	10.2	15.0	14.0	26.3

Table 2: Counterfactual distributions obtained using Algorithm 1 on the holdout dataset.

Appendix A. Proofs

A.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.

$$\left\{Q_X \;\middle|\; \mathsf{M}(Q_X, P_{X|S=1}) = \inf_{\mathsf{supp}(Q'_X)\subseteq\mathsf{supp}(P_{X|S=0})} \mathsf{M}(Q'_X, P_{X|S=1}), \mathsf{supp}(Q_X) \subseteq \mathsf{supp}(P_{X|S=0})\right\}.$$

We make the **assumption**:

$$P_{Y|X,S=0}(1|\boldsymbol{x}) > 0,\ \forall x \in \mathcal{X}. \tag{15}$$

1. **Distribution Alignment.** When $\lambda > 0$, there is only one counterfactual distribution: $P_{X|S=0}$. This is because:

   (a) KL-divergence is non-negative:

   $$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X} \circ Q'_X \| P_{\hat{Y}|X} \circ P_{X|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(Q'_X \| P_{X|S=1}) \geq 0;$$

   (b) If and only if $Q'_X = P_{X|S=1}$ (Cover and Thomas, 2012),

   $$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X} \circ Q'_X \| P_{\hat{Y}|X} \circ P_{X|S=1}) = 0 \text{ and } \mathrm{D}_{\mathrm{KL}}(Q'_X \| P_{X|S=1}) = 0.$$

   On the other hand, when $\lambda = 0$, there may be multiple counterfactual distributions besides $P_{X|S=1}$. In this case, the set of counterfactual distributions is closed and convex following from standard continuity and convexity results (Cover and Thomas, 2012).

2. **Calibration Error.** Note that

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

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

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

   which implies, for fixed $P_{\hat{Y}|X}$, $P_{Y|X,S}$, $P_{X|S=1}$, the mapping $Q'_X \to \mathsf{M}(Q'_X, P_{X|S=1})$ is continuous. Therefore, the set of counterfactual distributions is closed and non-empty.

   We define

   $$m_{\mathrm{CAL}} \triangleq \max\left\{\min_{\substack{\mathsf{supp}(Q'_X)\\ \subseteq\mathsf{supp}(P_{X|S=0})}} \left\{\sum_{\boldsymbol{x}\in\mathcal{X}} h_C(\boldsymbol{x}) Q'_X(\boldsymbol{x})\right\}, \mathbb{E}\left[\left|P_{Y|X,S=1}(1|X) - P_{\hat{Y}|X}(1|X)\right| \middle| S = 1\right]\right\}.$$

   Then counterfactual distributions, following from Definition 1, are distributions which satisfy

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

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,

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

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

3. **False Imbalance Rate.** Here we only give the proof for the False Negative Rate. Similar proof also holds for the False Discovery Rate and the False Positive Rate. Note that

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

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

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

which implies, for fixed $P_{\hat{Y}|X}$, $P_{Y|X,S}$, $P_{X|S=1}$, the mapping $Q'_X \to \mathsf{M}(Q'_X, P_{X|S=1})$ is continuous. Therefore, the set of counterfactual distributions is closed and non-empty.

We define

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

Then counterfactual distributions, following from Definition 1, are distributions which satisfy

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

This is equivalent to

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

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,

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

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

□

A.2 Proof of Proposition 2

Proof. First, we define

$$\Delta(f) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_{X|S=0}, P_{X|S=1}) - \mathsf{M}(P_{X|S=0}, P_{X|S=1})}{\epsilon}, \tag{16}$$

where $\widetilde{P}_{X|S=0}(\boldsymbol{x})$ is the perturbed distribution defined in (2). 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 Gateaux derivative:

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

and

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

Here Q is any distribution supported on $\mathcal{X}$. In particular, we can choose

$$Q(\boldsymbol{x}) = \left(\frac{1}{M_U} f(\boldsymbol{x}) + 1\right) P_{X|S=0}(\boldsymbol{x}),$$

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

$$(1-\epsilon)P_{X|S=0}(\boldsymbol{x}) + \epsilon Q(\boldsymbol{x}) = P_{X|S=0}(\boldsymbol{x}) + \frac{\epsilon}{M_U} f(\boldsymbol{x}) P_{X|S=0}(\boldsymbol{x}).$$

For simplicity, we use $P_{X|S=0} + \epsilon f P_{X|S=0}$ and $P_{X|S=0} + \frac{\epsilon}{M_U} f P_{X|S=0}$ to represent $P_{X|S=0}(\boldsymbol{x}) + \epsilon f(\boldsymbol{x}) P_{X|S=0}(\boldsymbol{x})$ and $P_{X|S=0}(\boldsymbol{x}) + \frac{\epsilon}{M_U} f(\boldsymbol{x}) P_{X|S=0}(\boldsymbol{x})$, respectively. Then

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

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

$$f(\boldsymbol{x}) = \frac{-\psi(\boldsymbol{x})}{\sqrt{\operatorname{Var}\left[\psi(X)|S=0\right]}}.$$

Therefore, following from the definition of correction function, we have

$$\phi(\boldsymbol{x}) = \frac{-\psi(\boldsymbol{x})}{\sqrt{\operatorname{Var}\left[\psi(X)|S=0\right]}}.$$

□

A.3 Proof of Proposition 5

Before we state the proof, let us recall a result by (Weissman et al., 2003) for the $\mathcal{L}_1$ deviation of the empirical distribution. For all $\epsilon > 0$,

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

where P is a probability distribution on the set $[M]$, $\pi_P \triangleq \max_{\mathcal{M}\subseteq[M]} \min(P(\mathcal{M}), 1-P(\mathcal{M}))$,

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

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

$$\Pr\left(\|\widehat{P} - P\|_1 \geq \epsilon\right) \leq \exp(M)\exp(-m\epsilon^2/2). \tag{17}$$

Hence, by taking $P = P_{Y,X|S=0}$, $M = |\mathcal{Y}||\mathcal{X}| = 2|\mathcal{X}|$ and $\epsilon = \sqrt{\frac{2}{m}(M - \log\beta)}$, Inequality (17) implies that, with probability at least $1-\beta$,

$$\left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \leq \sqrt{\frac{2}{m}(2|\mathcal{X}| - \log\beta)}, \tag{18}$$

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

$$\left\|\widehat{P}_{S,X} - P_{S,X}\right\|_1 \leq \sqrt{\frac{2}{m}(2|\mathcal{X}| - \log\beta)}. \tag{19}$$

Next, we prove the proposition.

Proof. We denote $\widehat{P}$ and $\widehat{\Pr}$ as estimated probability distribution and probability, respectively. Then we make the following **assumptions**:

$$\left\|\widehat{P}_{X|S=0} - P_{X|S=0}\right\|_p \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \tag{20}$$

$$\left|\widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \tag{21}$$

$$\left|\widehat{\Pr}(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p, \tag{22}$$

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 (20), (21), (22)). 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.

1. **Calibration Error.** Recall that the influence function of calibration error is

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

The estimated influence function of 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

$$\begin{aligned}&\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| - \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right|\\ &\leq \left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{Y|X,S=0}(1|\boldsymbol{x})\right|. \end{aligned} \tag{23}$$

Similarly, we have

$$\begin{aligned}&\left|\sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \widehat{P}_{X|S=0}(\boldsymbol{x})\right|\\ &\leq \sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right| P_{X|S=0}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}}\left|P_{X|S=0}(\boldsymbol{x}) - \widehat{P}_{X|S=0}(\boldsymbol{x})\right|. \end{aligned}\tag{24}$$

Following from (23) and (24), 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. **False Imbalance Rate.** Next, we prove the generalization bound for FNR. Similar proof holds for other metrics: FDR, and FPR. Recall that the influence function of 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 of FNR has the following alternative expression.

$$\begin{aligned}
\psi(\boldsymbol{x}) &= \frac{\Pr(Y=1|S=0) P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0, Y=1|S=0)}{\Pr(Y=1|S=0)^2} P_{Y|X,S=0}(1|\boldsymbol{x})\\
&= \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} P_{Y|X,S=0}(1|\boldsymbol{x}). \qquad (25)
\end{aligned}$$

The estimated influence function of FNR is

$$\widehat{\psi}(\boldsymbol{x}) = \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}). \qquad (26)$$

Following from (25), (26) and the triangle inequality, we have, for FNR,

$$\begin{aligned}
&\|\psi(\boldsymbol{x}) - \widehat{\psi}(\boldsymbol{x})\|_p\\
&\leq \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} \left(P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right) \right\|_p\\
&\quad + \left\| \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \left(\frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} \right.\right.\\
&\qquad \left.\left. - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right) \right\|_p\\
&\leq \left\| \frac{1}{\Pr(Y=1|S=0)} \left(P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})\right) \right\|_p\\
&\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p\\
&\lesssim \left\| P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \right\|_p\\
&\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p. \qquad (27)
\end{aligned}$$

Next, we have

$$\begin{aligned}
&\left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p \\
&\leq \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\widehat{\Pr}(Y=1|S=0)} \right\|_p + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{\widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right| \\
&\leq \left| \frac{\widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\
&\quad + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\
&\lesssim \left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| \\
&\quad + \left| \Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0) \right| \\
&\leq 2\left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| + \left| \widehat{\Pr}(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=0) \right|.
\end{aligned} \tag{28}$$

Combining (27) and (28) with the assumptions (21) and (22), 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. **Distribution Alignment.** Recall that the influence function of DA is

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

Since $P_{\hat{Y}|X}(1|\boldsymbol{x})$ is a given predictive model, 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

$$\begin{aligned}
\psi_r(\boldsymbol{x}) &\triangleq \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \middle| S=0 \right] \\
&= \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x} \in \mathcal{X}} P_{X|S=0}(\boldsymbol{x}) \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}.
\end{aligned} \tag{29}$$

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

$$\widehat{\psi}_r(\boldsymbol{x}) = \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x} \in \mathcal{X}} \widehat{P}_{X|S=0}(\boldsymbol{x}) \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})}. \tag{30}$$

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

$$\left|\log \frac{a}{b}\right| \leq \frac{|a-b|}{\min\{a,b\}}. \tag{31}$$

Then

$$\begin{aligned}
&\left|\log \frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \log \frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}\right| \\
&\leq |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})| \left(\frac{1}{\min\{\widehat{P}_{S|X}(1|\boldsymbol{x}), P_{S|X}(1|\boldsymbol{x})\}} + \frac{1}{\min\{1-\widehat{P}_{S|X}(1|\boldsymbol{x}), 1-P_{S|X}(1|\boldsymbol{x})\}}\right) \\
&\leq |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})| \frac{2}{m_X},
\end{aligned} \tag{32}$$

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 $\boldsymbol{x} \in \mathcal{X}$,

$$\left|\log \frac{1-\widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})}\right| \leq \frac{\left|1-2\widehat{P}_{S|X}(1|\boldsymbol{x})\right|}{\min\left\{\widehat{P}_{S|X}(1|\boldsymbol{x}), 1-\widehat{P}_{S|X}(1|\boldsymbol{x})\right\}} \leq \frac{1}{m_X}. \tag{33}$$

Combining (29) and (30) with (32) and (33), 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 (18), with probability at least $1-\beta$,

$$\left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \leq \sqrt{\frac{2}{m}\left(2|\mathcal{X}| - \log\beta\right)}. \tag{34}$$

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

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

Similarly, for DA, if $\widehat{P}_{S|X}$ is the empirical conditional distribution obtained from m i.i.d. samples, then

$$\left\|\widehat{\psi}(\boldsymbol{x}) - \psi(\boldsymbol{x})\right\|_1 \lesssim \left\|\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})\right\|_1 \lesssim \sqrt{\frac{1}{m}\left(|\mathcal{X}| - \log\beta\right)}.$$

□

Appendix B. Computing Influence Functions

B.1 Closed-Form Expressions of Influence Functions

We prove the closed-form expressions of influence functions under different discrimination metrics in this section.

Lemma 1 (Influence function of DA)**.** *For fixed $P_{\hat{Y}|X}$, $P_{Y|X,S}$, and P_S, let*

$$\mathsf{M}(P_{X|S=0}, P_{X|S=1}) = \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0}\|P_{\hat{Y}|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1}).$$

Then the influence function has the expression

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

Proof. Following from the definition of influence functions, we start with computing $\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_{X|S=0}+\epsilon\delta_{\boldsymbol{x}}\|P_{X|S=1})$.

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

Hence,

$$\begin{aligned}
&\lim_{\epsilon\to0}\frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_{X|S=0}+\epsilon\delta_{\boldsymbol{x}}\|P_{X|S=1})-\mathrm{D}_{\mathrm{KL}}(P_{X|S=0}\|P_{X|S=1})\right)\\
&=\log\frac{P_{X|S=0}(\boldsymbol{x})}{P_{X|S=1}(\boldsymbol{x})}-\mathbb{E}\left[\log\frac{P_{X|S=0}(X)}{P_{X|S=1}(X)}\middle|S=0\right].
\end{aligned}\tag{35}$$

Similarly, we have

$$\begin{aligned}
&\lim_{\epsilon\to0}\frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_{\hat{Y}|S=0}+\epsilon P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}}\|P_{\hat{Y}|S=1})-\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0}\|P_{\hat{Y}|S=1})\right)\\
&=\sum_{y\in\{0,1\}}((P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}})(y)-P_{\hat{Y}|S=0}(y))\log\frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)}\\
&=\sum_{y\in\{0,1\}}\log\frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)}P_{\hat{Y}|X}(y|\boldsymbol{x})-\mathbb{E}\left[\log\frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle|S=0\right].
\end{aligned}\tag{36}$$

Combining (35) with (36), we have

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

Note that

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

Hence,

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

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

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

□

Lemma 2 (Influence function of CAL). *For fixed $P_{\hat{Y}|X}$, $P_{Y|X,S}$, and P_S, let*

$$\begin{aligned}&\mathsf{M}(P_{X|S=0},P_{X|S=1})\\&=\mathbb{E}\left[\left|P_{Y|X,S=0}(1|X)-P_{\hat{Y}|X}(1|X)\right|\middle|S=0\right]-\mathbb{E}\left[\left|P_{Y|X,S=1}(1|X)-P_{\hat{Y}|X}(1|X)\right|\middle|S=1\right].\end{aligned}$$

Then the influence function has the expression

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

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

Proof. Based on the definition of $h_C(\boldsymbol{x})$, we have

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

When we perturb the distribution $P_{X|S=0}$, the function $h_C(\boldsymbol{x})$ and the quantity

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

do not change. Therefore,

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

□

Next, we compute the influence function of FNR. Similar analysis can be used for computing the influence functions of FPR and FDR, respectively.

Lemma 3 (Influence function of FNR)**.** *For fixed $P_{\hat{Y}|X}$, $P_{Y|X,S}$, and P_S, let*

$$\begin{aligned}&\mathsf{M}(P_{X|S=0}, P_{X|S=1})\\ &= \Pr(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=1).\end{aligned}$$

Then the influence function has the expression

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

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

Proof. First, based on the setup, the joint distribution $P_{S,X,Y,\hat{Y}}$ can be factorized as $P_{S,X,Y,\hat{Y}} = P_{\hat{Y}|X}P_{Y|X,S}P_SP_{X|S}$. Then

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

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

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

which implies

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

Therefore,

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

□

B.2 Estimating Influence Functions and Discrimination Metrics from Dataset

We show how to estimate influence functions and discrimination metrics from an auditing dataset. Here the input is

- $\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^n$: an auditing dataset;
- $P_{\hat{Y}|X}(1|\boldsymbol{x})$: a fixed predictive model 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 and $S=0$.

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 $\mathcal{D}^{\text{audit}}$:

- Distribution Alignment:

$$\begin{aligned}\psi(\boldsymbol{x}) = &\left(\log\frac{\hat{p}_0(1-\hat{p}_1)}{\hat{p}_1(1-\hat{p}_0)}\right)\left(P_{\hat{Y}|X}(1|\boldsymbol{x}) - \hat{p}_0\right)\\ &+\lambda\left(\log\frac{1-P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}}\log\frac{1-P_{S|X}(1|\boldsymbol{x}')}{P_{S|X}(1|\boldsymbol{x}')}\right),\end{aligned} \tag{37}$$

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

- Calibration Error:

$$\psi(\boldsymbol{x}) = h_C(\boldsymbol{x}) - \frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} h_C(\boldsymbol{x}'), \tag{38}$$

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

- False Discovery Rate:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} t_2(\boldsymbol{x}')\right)t_1(\boldsymbol{x}) - \left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} t_1(\boldsymbol{x}')\right)t_2(\boldsymbol{x})}{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} t_2(\boldsymbol{x}')\right)^2}, \tag{39}$$

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

- False Negative Rate:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} r_2(\boldsymbol{x}')\right) r_1(\boldsymbol{x}) - \left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} r_1(\boldsymbol{x}')\right) r_2(\boldsymbol{x})}{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} r_2(\boldsymbol{x}')\right)^2}, \tag{40}$$

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

- False Positive Rate:

$$\psi(\boldsymbol{x}) = \frac{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} s_2(\boldsymbol{x}')\right) s_1(\boldsymbol{x}) - \left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} s_1(\boldsymbol{x}')\right) s_2(\boldsymbol{x})}{\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} s_2(\boldsymbol{x}')\right)^2}, \tag{41}$$

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

Next, we show how to compute discrimination metrics from $\mathcal{D}^{\text{audit}}$:

- Distribution Alignment:

$$\begin{aligned}\mathsf{M}(\mathcal{D}^{\text{audit}}) =& \hat{p}_0 \log\frac{\hat{p}_0}{\hat{p}_1} + (1-\hat{p}_0)\log\frac{1-\hat{p}_0}{1-\hat{p}_1} \\ &+ \lambda\left(\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} \log\frac{1-P_{S|X}(1|\boldsymbol{x}')}{P_{S|X}(1|\boldsymbol{x}')} + \log\frac{p_s}{1-p_s}\right),\end{aligned} \tag{42}$$

where $\hat{p}_0 = \frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} P_{\hat{Y}|X}(1|\boldsymbol{x}')$, $\hat{p}_1 = \frac{1}{n_1}\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} P_{\hat{Y}|X}(1|\boldsymbol{x}')$, and $p_s = \frac{1}{n_0+n_1}\sum_{\boldsymbol{x}'\in\mathcal{D}^{\text{audit}}} P_{S|X}(1|\boldsymbol{x}')$;

- Calibration Error:

$$\mathsf{M}(\mathcal{D}^{\text{audit}}) = \left|\frac{1}{n_0}\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} h_{C0}(\boldsymbol{x}') - \frac{1}{n_1}\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} h_{C1}(\boldsymbol{x}')\right|, \tag{43}$$

where $h_{Ci}(\boldsymbol{x}) = \left|P_{Y|X,S=i}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right|$ for $i = 0, 1$;

- False Discovery Rate:

$$\mathsf{M}(\mathcal{D}^{\text{audit}}) = \left|\frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} t_{10}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} t_{20}(\boldsymbol{x}')} - \frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} t_{11}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} t_{21}(\boldsymbol{x}')}\right|, \tag{44}$$

where $t_{1i}(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})(1 - P_{Y|X,S=i}(1|\boldsymbol{x}))$, $t_{2i}(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})$ for $i = 0, 1$;

- False Negative Rate:

$$\mathsf{M}(\mathcal{D}^{\text{audit}}) = \left|\frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} r_{10}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} r_{20}(\boldsymbol{x}')} - \frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} r_{11}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} r_{21}(\boldsymbol{x}')}\right|, \tag{45}$$

where $r_{1i}(\boldsymbol{x}) = (1 - P_{\hat{Y}|X}(1|\boldsymbol{x}))P_{Y|X,S=i}(1|\boldsymbol{x})$, $r_{2i}(\boldsymbol{x}) = P_{Y|X,S=i}(1|\boldsymbol{x})$ for $i = 0, 1$;

- False Positive Rate:

$$\mathsf{M}(\mathcal{D}^{\text{audit}}) = \left|\frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} s_{10}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_0^{\text{audit}}} s_{20}(\boldsymbol{x}')} - \frac{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} s_{11}(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{D}_1^{\text{audit}}} s_{21}(\boldsymbol{x}')}\right|, \tag{46}$$

where $s_{1i}(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})(1 - P_{Y|X,S=i}(1|\boldsymbol{x}))$, $s_{2i}(\boldsymbol{x}) = 1 - P_{Y|X,S=i}(1|\boldsymbol{x})$ for $i = 0, 1$.

Appendix C. Further Discussions of Metrics

C.1 Distribution Alignment

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 $Q \in \{P_1, P_2\}$. To determine the source distribution, we construct a decision rule that will choose between the hypotheses $H_1 : Q = P_1$ and $H_2 : Q = P_2$. The Neyman–Pearson lemma states that the optimal decision rule is given by the two acceptance regions, $\mathcal{A}_1^{(n)} = \{x^n \mid \log \frac{P_1(x^n)}{P_2(x^n)} > T\}$ and $\mathcal{A}_2^{(n)} = (\mathcal{A}_1^{(n)})^c$. A Type I (resp. Type II) error occurs when x^n is drawn from P_1 (resp. P_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 $\lim_{n\to\infty} -\frac{1}{n} \log \beta_2^{(n)}(\delta) = \mathrm{D}_{\mathrm{KL}}(P_1 \| P_2)$. Hence, we use KL-divergence to measure the statistical indistinguishability of two populations in order to control error exponents.

C.2 Calibration Error

The standard definition of Expected Calibration Error (see e.g., Guo et al., 2017; Naeini et al., 2015) is given below.

$$\mathsf{ECE} \triangleq \sum_{p \in h(\mathcal{X})} \Pr(h(X) = p | S = 0) \left| \Pr(Y = 1 | h(X) = p, S = 0) - p \right|, \tag{47}$$

where $h(\boldsymbol{x}) \triangleq P_{\hat{Y}|X}(1|\boldsymbol{x})$ and $h(\mathcal{X}) \triangleq \{h(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. In this paper, we choose

$$\mathsf{M}(P_{X|S=0}, P_{X|S=1}) = \mathbb{E}\left[\left| P_{Y|X,S=0}(1|X) - P_{\hat{Y}|X}(1|X) \right| \middle| S = 0\right]$$

to measure calibration error. In what follows, we show our choice upper bounds the ECE:

$$\mathsf{ECE} \leq \mathbb{E}\left[\left| P_{Y|X,S=0}(1|X) - P_{\hat{Y}|X}(1|X) \right| \middle| S = 0\right].$$

Proof. For any $p \in h(\mathcal{X})$, we denote $h^{-1}(p) \triangleq \{\boldsymbol{x} \in \mathcal{X} \mid P_{\hat{Y}|X}(1|\boldsymbol{x}) = p\}$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\left|P_{Y|X,S=0}(1|X) - P_{\hat{Y}|X}(1|X)\right| \middle| S=0\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} P_{X|S=0}(\boldsymbol{x}) \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \\
&= \sum_{p\in h(\mathcal{X})} \sum_{\boldsymbol{x}\in h^{-1}(p)} P_{X|S=0}(\boldsymbol{x}) \left|P_{Y|X,S=0}(1|\boldsymbol{x}) - P_{\hat{Y}|X}(1|\boldsymbol{x})\right| \\
&= \sum_{p\in h(\mathcal{X})} \sum_{\boldsymbol{x}\in h^{-1}(p)} \left|P_{Y,X|S=0}(1,\boldsymbol{x}) - pP_{X|S=0}(\boldsymbol{x})\right| \\
&\geq \sum_{p\in h(\mathcal{X})} \left|\sum_{\boldsymbol{x}\in h^{-1}(p)} \left(P_{Y,X|S=0}(1,\boldsymbol{x}) - pP_{X|S=0}(\boldsymbol{x})\right)\right| \\
&= \sum_{p\in h(\mathcal{X})} \left|P_{Y,X|S=0}(1,\boldsymbol{x}\in h^{-1}(p)) - pP_{X|S=0}(\boldsymbol{x}\in h^{-1}(p))\right| \\
&= \sum_{p\in h(\mathcal{X})} P_{X|S=0}(\boldsymbol{x}\in h^{-1}(p)) \left|P_{Y|X,S=0}(1|\boldsymbol{x}\in h^{-1}(p)) - p\right| \\
&= \sum_{p\in h(\mathcal{X})} \Pr(h(X)=p|S=0) \left|\Pr(Y=1|h(X)=p, S=0) - p\right| \\
&= \mathsf{ECE}.
\end{aligned}$$

□

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.

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

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

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

Shashi Borade and Lizhong Zheng. Euclidean information theory. In *IEEE International Zurich Seminar on Communications*, pages 14–17. IEEE, 2008.

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.

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

Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5(Aug):1007–1034, 2004.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

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

D Bradford Hunt. Redlining. *Encyclopedia of Chicago*, 2005.

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.

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.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207, 1996.

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

Yong Liu, Shali Jiang, and Shizhong Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, pages 324–332, 2014.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907, 2015.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

James Robins, Lingling Li, Eric Tchetgen, and Aad W van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69(2-3):227–247, 2009.

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.

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, 2017.