Repairing without Retraining: 
Avoiding Disparate Impact with Counterfactual Distributions

Hao Wang 1 Berk Ustun 1 Flavio P. Calmon 1

Abstract
When the performance of a machine learning model varies over groups defined by sensitive attributes (e.g., gender or ethnicity), the performance disparity can be expressed in terms of the probability distributions of the input and output variables over each group. In this paper, we exploit this fact to reduce the disparate impact of a fixed classification model over a population of interest. Given a black-box classifier, we aim to eliminate the performance gap by perturbing the distribution of input variables for the disadvantaged group. We refer to the perturbed distribution as a counterfactual distribution, and characterize its properties for common fairness criteria. We introduce a descent algorithm to learn a counterfactual distribution from data. We then discuss how the estimated distribution can be used to build a data preprocessor that can reduce disparate impact without training a new model. We validate our approach through experiments on real-world datasets, showing that it can repair different forms of disparity without a significant drop in accuracy.

1. Introduction
A machine learning model has disparate impact when its performance changes across groups defined by sensitive attributes such as race or gender (Barocas & Selbst, 2016). Recent work has shown that models can exhibit significant performance disparities between groups (see e.g. Angwin et al., 2016; Dastin, 2018). Such disparities have led to a plethora of research on fair machine learning, focusing on how disparate impact arises (Chen et al., 2018; Datta et al., 2016), how it can be measured (Zliobaitė, 2017; Piersom et al., 2017; Simoiu 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., 2017; Calmon et al., 2017; Menon & Williamson, 2018; Canetti et al., 2019). Despite these developments, disparate impact remains difficult to avoid in a large class of real-world applications where:

- Models are procured from a third-party vendor who has the data or technical expertise required for model development (Guszcza et al., 2018).
- Models are deployed on a population where the data distribution differs from the distribution of the training data (i.e., due to dataset shift, Sugiyama et al., 2017).

In such settings, disparate impact is challenging to address, let alone understand. Users typically have black-box access to the model (e.g., via a prediction API), may not have access to the training data (e.g., due to privacy concerns or intellectual property rights), and may not be able to draw conclusions from the training data (e.g., due to distributional shifts in deployment).

In this paper, we aim to mitigate disparate impact in such settings. Our object of interest is a hypothetical distribution of input variables that minimizes disparate impact in the model’s deployment population. We refer to this distribution as a counterfactual distribution. As we will show, an information-theoretic analysis of counterfactual distributions has much to offer. Given a fixed classifier, disparate impact can be expressed in terms the distributions of input and output variables between groups (cf. Figure 1). In turn, a counterfactual distribution can be obtained by repeatedly perturbing the distribution of input variables until a specific measure of disparity is minimized over the deployment population. The counterfactual distribution can then be used to repair the model so that it no longer exhibits disparate impact in deployment.

Contributions The main contributions of this paper are:

1. We introduce a theoretical framework to mitigate performance disparities for a black-box classifiers in deployment.
2. We develop machinery to learn counterfactual distributions from data. Our tools recover a counterfactual distribution using a descent procedure in the simplex of
A rural clinic purchases a classification model to detect bone fractures in x-rays and discovers that patients with a certain physical trait have high FPR.

A bank enters a new market and discovers its credit score underperforms on customers over 60 years of age.

The tools in this paper will be able to scrutinize and repair performance disparities in all three settings — regardless of whether the model directly uses the sensitive attribute. More importantly, they will be able to achieve parity-based notions of fairness in a way that: (i) only affects one group; (ii) benefits the group it affects (on average); and (iii) incentivizes individuals to reveal their sensitive attributes at prediction time. The latter two points (i.e., do-no-harm and opt-in) are important elements of ethical treatment disparity (see e.g., Lipton et al., 2018; Ustun et al., 2019a).

Related Work  We develop a theoretical framework that is used to design methods to determine counterfactual distributions in practice. We then use counterfactual distributions to design optimal transport-based pre-processing methods for ensuring fairness. In this regard, the closest work to ours are those of Feldman et al. (2015); Johndrow & Lum (2017); Del Barrio et al. (2019), which propose methods to control specific disparate impact metrics via optimal transport. These methods differ from ours in that they (i) focus on reducing measures of disparity related to predicted outcomes; (ii) map the input variable distributions across all sensitive groups to a common distribution. More broadly, our approach differs from other model-agnostic approaches to mitigate disparate impact (e.g., pre-processing methods such as Kamiran & Calders, 2012; Calmon et al., 2017) in that it does not require access to the training data, and does not require training a new model.

The term “counterfactual distribution” is often used to describe different kinds of hypothetical effects. In the statistics and economics literature (see e.g., Balke & Pearl, 1994; DiNardo et al., 1995; Rubin, 2005; Fortin et al., 2011; Chernozhukov et al., 2013; Johansson et al., 2016; Peters et al., 2017; Fisher & Kennedy, 2018), a counterfactual distribution refers to a hypothetical distribution of an outcome variable given a specific distribution of input variables (e.g., the distribution of wages (outcome variable) for young workers if young workers had the same qualifications as older workers). The counterfactual distribution in this work describes a different kind of effect — i.e., a distribution of input variables to minimize disparate impact — and, consequently, must be derived using a different set of tools.

Use Cases  Our approach provides a way to repair classifiers in domains where treatment disparity is legal and ethical\(^1\). Illustrative use cases include:

- A hospital purchases a readmissions prediction model that satisfies equal opportunity for different genders in the training data but violates it in deployment;

\(^1\)Treatment disparity is not illegal nor unethical in all applications of machine learning where fairness is important. In medicine, for example, it is acceptable to use sensitive attributes in prediction. In lending, the Equal Credit Opportunity Act permits the use of age in credit scoring models.

\(P_0\)
\(P_1\)
\(Q_X\)
\(Q_Y\)
\(P_Y\vert X = 0\)
\(P_Y\vert X = 1\)
\(P_0\vert S = 0\)
\(P_0\vert S = 1\)
\(P_1\vert S = 0\)
\(P_1\vert S = 1\)
\(P_{\hat{Y}}\vert S = 0\)
\(P_{\hat{Y}}\vert S = 1\)
\(P_Y\vert S = 0\)
\(P_Y\vert S = 1\)
\(P_{\hat{Y}}\vert S = 0\)
\(P_{\hat{Y}}\vert S = 1\)

Figure 1: Illustration of probability distributions affecting the disparate impact of a fixed classification model \(h\). Here, \(P_0\) and \(P_1\) denote the distributions of input variables for groups where \(S = 0\) and \(S = 1\), respectively. Disparate impact is a function of the distributions of predicted outcomes \((P_{\hat{Y}}\vert S = 0, P_{\hat{Y}}\vert S = 1)\) and true outcomes \((P_Y\vert S = 0, P_Y\vert S = 1)\). A counterfactual distribution \(Q_X\) is a perturbation of \(P_0\) that minimizes a specific measure of disparity. If disparate impact persists under \(Q_X\), there may be irreconcilable differences between the groups (i.e., \(P_{\hat{Y}}\vert X, S = 0 \neq P_{\hat{Y}}\vert X, S = 1\), see Prop. 1). The counterfactual distribution may not be unique, as illustrated by the shaded ellipse.

Additional Resources  We provide a software implementation of our tools at http://github.com/ustunb/cfddist. This paper extends work that was first presented at the NeurIPS WESGAI Workshop (Wang et al., 2018b).
2. Framework

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

Preliminaries We consider a standard classification task where the goal is to predict a binary outcome variable Y ∈ {0, 1} using a vector of input variables X = (X₁, ..., X₄) ∈ X drawn from the probability distribution Pₓ. We are given a black-box classifier h : X → [0, 1]. We assume that h(x) ∈ {0, 1} if the classifier outputs a predicted outcome (e.g., SVM) and h(x) ∈ [0, 1] if it outputs a predicted probability (e.g., logistic regression).

We evaluate differences in the performance of the classifier with respect to a sensitive attribute S ∈ {0, 1} with distribution Pₛ. We refer to the subset of individuals with S = 0 and S = 1 as the target and baseline groups, respectively. We denote their distributions of input variables as P₀ ≜ Pₓ|S=0, and P₁ ≜ Pₓ|S=1. Likewise, we let P₁|S=0(1) ≜ E[h(X)|S = 0] and P₁|S=1(1) ≜ E[h(X)|S = 1]. For the sake of clarity, we assume that S is not an input variable to the model. Note, however, that our tools and results can be extended to settings where S an input variable for the classifier.

Disparity Metrics We measure the performance disparity between groups in terms of a disparity metric. Formally, a disparity metric is a mapping M : P → R where P is the set of probability distributions over X. We provide examples of M(P₁) for common fairness criteria in Table 1. Note that we write disparity metrics as M(P₁) since they can be expressed as a function of P₁ once the classifier and the distributions Pₓ|X,S, P₁, and Pₛ are fixed.²

Counterfactual Distributions A counterfactual distribution is a hypothetical probability distribution of input variables for the target group that minimizes a specific disparity metric.

Definition 1. A counterfactual distribution Qₓ is a distribution of input variables for the target group such that:

\[ Qₓ \in \arg\min_{Q′ₓ \in \mathcal{P}} |M(Q′ₓ)|, \tag{1} \]

where M(·) is a given disparity metric and \(\mathcal{P}\) is the set of probability distributions over X.

There exist several ways to resolve the performance disparity of a fixed classifier by perturbing the distributions of input variables of sensitive groups. For example, one could simultaneously perturb the input distributions for all groups to a “midpoint” distribution (see e.g., the distributions considered by Feldman et al., 2015; Johndrow & Lum, 2017; Del Barrio et al., 2019, to achieve statistical parity).

While our tools could recover such distributions, we will purposely consider a counterfactual distribution that alters the input variables for a sensitive group that attains the less favorable performance (i.e., the target group S = 0). This choice reflects our desire to resolve the performance disparity by having the target group perform better, rather than having the baseline group perform worse. As we discuss later, this choice reduces the data requirements to estimate the counterfactual distribution and the individuals who are affected by the repair (i.e., this approach only produces a preprocessor that affects individuals where S = 0).

At this point, an observant reader may wonder why a counterfactual distribution for the target group is not simply the distribution of input variables over the baseline group (i.e., \(Qₓ \equiv P₁\)). In fact, the distribution of input variables for the baseline group \(P₁\) is not necessarily a counterfactual distribution when \(P₁|X,S=0 \neq P₁|X,S=1\). We illustrate this point with the following example.

Example 1. Consider a classification task where the input variables X = (X₁, X₂) ∈ {0, 1}² are drawn from distributions such that \(Pₓ|S=s = Pₓ|1|S=s \cdot Pₓ|2|S=s\) for \(s \in \{0, 1\}\) where:

\[
\begin{align*}
\text{Pr}(X₁ = 1|S = 0) &= 0.9, \quad \text{Pr}(X₁ = 1|S = 1) = 0.2, \\
\text{Pr}(X₁ = 1|S = 1) &= 0.1, \quad \text{Pr}(X₂ = 1|S = 1) = 0.5.
\end{align*}
\]

Assume that the true outcome variables Y are drawn from the conditional distributions:

\[
\begin{align*}
P₁|X,S=0(1|x) &= \logistic(2x₁ – 2x₂), \\
P₁|X,S=1(1|x) &= \logistic(2x₁ + 4x₂ - 3). \tag{2}
\end{align*}
\]

In this case, the Bayes optimal classifier for \(S = 1\) is \(h(x) = \mathbb{I}[x₂ = 1]\). Using the difference in FPR as the disparity metric, h achieves \(M(P₁) = 25.1\%\). In this case, setting \(P₀ \leftarrow P₁\) would achieve a disparity of \(M(P₁) = 43.6\%\). In contrast, we can achieve a disparity metric of \(M(Qₓ) = 0.0\%\) for a counterfactual distribution such that:

\[
\begin{align*}
Qₓ(0,0) &= 0.50, \\
Qₓ(0,1) &= 0.09, \\
Qₓ(1,0) &= 0.41, \\
Qₓ(1,1) &= 0.00.
\end{align*}
\]

Example 1 shows that counterfactual distributions may be non-trivial when the conditional distributions of Y given X differ across groups (i.e., \(P₁|X,S=0 \neq P₁|X,S=1\)). In particular, the condition \(P₁|X,S=0 \neq P₁|X,S=1\) will always hold whenever counterfactual distributions do not completely eliminate the disparity between groups. We formalize this statement in the next proposition.

²As we show in Appendix A.1, the disparity metric can be expressed as a function of P₁ once the following objects are fixed: h, the classifier; P₁|X,S the distribution of the true outcomes given input variables and sensitive attribute; P₁, the distribution of input variables over the baseline group; and Pₛ the distribution of the sensitive attribute.
We start by formally defining the local perturbation of an input distribution. If \( M(P_0) \) represents a disparity metric in Table 1, then \( P_Y|X,S=0 \neq P_Y|X,S=1 \).

Proposition 1. If \( M(Q_X) > 0 \) where \( Q_X \) is a counterfactual distribution for a disparity metric in Table 1, then \( P_Y|X,S=0 \neq P_Y|X,S=1 \).

Proposition 1 illustrates how a counterfactual distribution can be used to detect cases where a classifier exhibits an irreconcilable performance disparity between groups – i.e., a disparity that cannot be resolved by perturbing the distributions of input variables for the target group. The result complements various impossibility results on inevitable trade-offs between groups (see e.g., Chouldechova, 2017; Kleinberg et al., 2016; Pleiss et al., 2017). It also provides a sufficient condition that can inform the need for treatment disparity (see e.g., of Kleinberg et al., 2018; Dwork et al., 2018; Lipton et al., 2018; Ustun et al., 2019a).

### Table 1: Disparity metrics

<table>
<thead>
<tr>
<th>PERFORMANCE METRIC</th>
<th>ACRONYM</th>
<th>DISPARITY METRIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Parity</td>
<td>SP</td>
<td>( \Pr(Y = 0</td>
</tr>
<tr>
<td>False Discovery Rate</td>
<td>FDR</td>
<td>( \Pr(Y = 0</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>FNR</td>
<td>( \Pr(Y = 0</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>FPR</td>
<td>( \Pr(\tilde{Y} = 1</td>
</tr>
<tr>
<td>Distribution Alignment</td>
<td>DA_{\lambda}</td>
<td>( D_{a\lambda}(P_{Y</td>
</tr>
</tbody>
</table>

Proposition 2. Given a disparity metric \( M(\cdot) \), we have that

\[
\arg\min_{f(x)} \lim_{\epsilon \to 0} \frac{M((1 - \epsilon)P_0 + \epsilon \delta_x) - M(P_0)}{\epsilon} = -\psi(x)
\]

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

Proposition 3 shows that when disparity is measured using a linear combination of metrics, the influence function for the compound metric can be expressed as a linear combination of the influence functions for its components.

3. Methodology

In this section, we present information-theoretic tools to learn counterfactual distributions from data. We first demonstrate how influence functions provide a natural “descent direction” in this setting. Next, we derive closed-form expressions for the influence functions of the disparity metrics in Table 1. Lastly, we present a descent procedure that combines these results to learn a counterfactual distribution for the deployment population.

3.1. Measuring the Descent Direction

In what follows, we describe how to reduce the value of a disparity metric by perturbing the distribution of input variables over the target group \( \tilde{P}_0 \).

We start by formally defining the local perturbation of an input distribution.

**Definition 2.** The perturbed distribution \( \tilde{P}_0 \) over the target group \( (S = 0) \) is given by

\[
\tilde{P}_0(x) = P_0(x) + \epsilon f(x), \quad \forall x \in \mathcal{X}
\]

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 \( \tilde{P}_0 \) is a valid probability distribution.

Here, \( f(x) \) represents a direction in the probability simplex while \( \epsilon \) represents the magnitude of perturbation (see e.g., Borade & Zheng, 2008; Anantharam et al., 2013; Huang et al., 2014, for other applications of local perturbations of measures in information theory).

As we will see shortly, the direction of steepest descent for disparate impact can be measured using an influence function (see e.g., Huber, 2011; Koh & Liang, 2017, for other uses in machine learning).

**Definition 3.** For a disparity metric \( M(\cdot) \), the influence function \( \psi : \mathcal{X} \to \mathbb{R} \) is given by

\[
\psi(x) = \lim_{\epsilon \to 0} \frac{M((1 - \epsilon)P_0 + \epsilon \delta_x) - M(P_0)}{\epsilon}
\]

where \( \delta_x(z) = [z = x] \) is the delta function at \( x \).

Intuitively, given a sufficiently large dataset from the deployment population, the influence function approximates the change in a disparity metric when a sample \( x \in \mathcal{X} \) from the target group is removed (or added) to the dataset.

In Proposition 2, we show that perturbing the distribution \( P_0 \) along the direction defined by \(-\psi(x)\) produces the largest local decrease of the disparity metric. That is, \(-\psi(x)\) reflects the direction of steepest descent in disparate impact.

**Proposition 3.** Distribution Alignment is a metric proposed in Wang et al. (2018a) that measures the disparity of predicted outcomes via the KL-divergence.
Proposition 3. Given any convex combination of $K$ disparity metrics $M(P_b) = \sum_{i=1}^{K} \lambda_i M_i(P_b)$, the influence function of the compound disparity metric $M(P_b)$ has the form:

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

Proposition 3 allows us to consider a larger class of disparity measures than those shown in Table 1. For instance, one can recover a counterfactual distribution to achieve equalized odds (Hardt et al., 2016) by using a convex combination of influence functions for FPR and FNR.

3.2. Computing Influence Functions

We now present closed-form expressions for the influence functions of disparity metrics shown in Table 1. The expressions will be cast in terms of three classifiers:

- $h(x)$: the black-box classifier that we aim to repair;
- $\hat{s}(x)$: a classifier that uses the same input variables as $h$, but aims to predict the probability that an individual belongs to the baseline group, $P_{S|X}(1|x)$.
- $\hat{y}_0(x)$: a classifier that uses the same input variables as $h$, but aims to predict the true outcome for individuals in the target group, $P_Y|X,S=0(1|x)$.

Given $h(x)$, we would train $\hat{s}(x)$ and $\hat{y}_0(x)$ using an auditing dataset $D_{aud} = \{(x_i, y_i, s_i)\}_{i=1}^{n}$ drawn from the deployment population. With these three models in hand, we can then compute influence functions using closed-form expressions shown in the following proposition.

Proposition 4. The influence functions for the disparity metrics in Table 1 can be expressed as

$$
\psi^{FP}(x) = -h(x) + \mu_0,
$$

$$
\psi^{FPR}(x) = h(x)(1-\hat{y}_0(x)) - \nu_{0,1} h(x),
$$

$$
\psi^{FN}(x) = (1-h(x))\hat{y}_0(x) - \gamma_{0,1} \hat{y}_0(x),
$$

$$
\psi^{FP}(x) = h(x)(1-\hat{y}_0(x)) - \gamma_{1,0} (1-\hat{y}_0(x)),
$$

$$
\psi^{DA}(x) = \log \frac{\mu_0}{\mu_1} \frac{(1-\mu_0)}{\mu_1} h(x) + \lambda \log \frac{1-\hat{s}(x)}{\hat{s}(x)}
$$

$$
\mu_{a,b} \triangleq \Pr(Y = 1 | S = s),
\mu_0 \triangleq \Pr(Y = 1),
\gamma_{a,b} \triangleq \Pr(\hat{Y} = a | Y = b, S = 0),
\nu_{a,b} \triangleq \Pr(Y = a) \cdot \hat{y}_0(x) = 1, S = 0).
$$

3.3. Learning Counterfactual Distributions from Data

So far we have shown that influence functions can be used to evaluate the direction of steepest descent of a disparity metric (Proposition 2), and that the value of an influence function can be estimated using data from the deployment population (Proposition 4).

Considering these results, one would expect that disparity could be minimized by repeatedly (i) perturbing the distribution in the direction of steepest descent (5), and (ii) estimating the influence function at the new, perturbed distribution. Repeating these steps, we would recover an approximate solution to (1) — i.e., an approximate counterfactual distribution.

In Algorithm 1, we formalize this intuition by presenting a descent procedure to recover a counterfactual distribution for a given disparity metric $M(\cdot)$. The procedure is analogous to stochastic gradient descent in the space of distributions over $\mathcal{X}$, where the resampling at each iteration corresponds to a gradient step.

**Algorithm 1 Distributional Descent.**

**Input:**

$h : \mathcal{X} \rightarrow [0, 1]$ classification model

$\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{n}$ data from deployment population

$M(\cdot)$ disparity metric

$\epsilon > 0$ step size

**Initialize**

$I_0 \leftarrow \{ i = 1, \ldots, n | s_i = 0 \}$

$I_D \leftarrow \{ i = 1, \ldots, n | s_i \neq 0 \}$

$w_0 \leftarrow \{w_i\}_{i=1}^{n} \cdot \{ w_i = 1.0 \}$ initial weights

$M \leftarrow M(\mathcal{D}_0 \cup D_1)$ disparity metric

**repeat**

$M_{old} \leftarrow M$

$\psi_i \leftarrow \psi(x_i)$ for $i \in I_0$ compute $\psi(x_i)$ for points $\mathcal{D}_0$

$w_i \leftarrow (1-\epsilon \psi_i) \cdot w_i$ for $i \in I_0$

$\mathcal{D}_0 \leftarrow \text{RESAMPLE}(\mathcal{D}_0, w_0)$ samples from counterfactual distribution

$M \leftarrow M(\mathcal{D}_0 \cup D_1)$ evaluate disparity metric

**until** $M \geq M_{old}$

**return:** $\mathcal{D}_0 \cup D_1$ samples from counterfactual distribution

**procedure** RESAMPLE($\mathcal{D}$, $w$)

**return** $\mathcal{D}$ points sampled from $\mathcal{D}$ using the weights $w$

**end procedure**

Given a classifier $h$ and a dataset $\{(x_i, y_i, s_i)\}_{i=1}^{n}$ from the distribution $P_{X,Y,S}$, the procedure outputs a dataset drawn from a counterfactual distribution.

The procedure pairs each point with a sampling weight $w_i$, which is initialized as $w_i = 1.0$. At each iteration, it first computes the value of the influence function $\psi(x)$. Next, it updates the values of each sampling weight for each point in the target group as $(1- \epsilon \psi(x_i)) \cdot w_i$, where $\epsilon$ is a
user-specified step size parameter. The updated sampling weights represent the direction in which the distribution for the target group should be perturbed to reduce \( M(\cdot) \). The data points from the target group are then resampled with their sampling weights. The set of resampled points mimics one drawn from the perturbed distribution.

The procedure determines if the classifier still has disparate impact at the end of each iteration by computing the value of \( M(\cdot) \) on the set of resampled points. These steps are repeated until \( M(\cdot) \) ceases to decrease. Once the procedure stops, it outputs: (i) dataset drawn from a counterfactual distribution; (ii) a set of sampling weights for each point from the target group, which can be used to draw samples from the counterfactual distribution.

In Figure 2, we show the progress (and convergence) of Algorithm 1 when recovering a counterfactual distribution for a synthetic dataset described in Appendix B.

4. Model Repair

In this section, we describe how to use counterfactual distributions to repair classifiers that exhibit disparate impact.

Preprocessor Given a classifier \( h(x) \), we aim to mitigate disparate impact by constructing a preprocessor \( T : X \to X \) that alters the features of the target group. Thus, the repaired classifier \( \tilde{h}(x) \) will operate as:

\[
\tilde{h}(x) = \begin{cases} 
    h(T(x)) & \text{if } s = 0, \\
    h(x) & \text{otherwise.}
\end{cases}
\] (7)

The preprocessor is a (potentially randomized) mapping that transforms the distribution of samples over the target population into the counterfactual distribution, i.e., given a random variable \( X \) drawn from the target population distribution, the distribution of \( T(X) \) will approximate counterfactual distribution.

Optimal Transport We produce the preprocessor by solving an optimal transport problem. To this end, we require the following inputs:

- \( D_0 \), which represents the original samples for the target group. We assume \( D_0 \) contains \( n_0 \) samples, of which \( m \) are distinct: \( \{ x_1, \ldots, x_m \} \).
- \( \tilde{D}_0 \), which represents the samples drawn from the counterfactual distribution (i.e., the data produced via resampling in Algorithm 1). We assume that \( \tilde{D}_0 \) contains \( \tilde{n}_0 \) samples, of which \( \tilde{m} \) are distinct: \( \{ \tilde{x}_1, \ldots, \tilde{x}_{\tilde{m}} \} \).

With these samples at hand, we formulate an optimal transport problem of the form:

\[
\begin{align}
& \min_{\gamma_{ij} \in \mathbb{R}^+} \sum_{i=1}^m \sum_{j=1}^{\tilde{m}} C_{ij} \gamma_{ij} \quad (8a) \\
& \text{s.t.} \quad \sum_{j=1}^{\tilde{m}} \gamma_{ij} = \delta_{x_i} \quad i = 1, \ldots, m \\
& \qquad \quad \sum_{i=1}^m \gamma_{ij} = \delta_{\tilde{x}_j} \quad j = 1, \ldots, \tilde{m}. \quad (8c)
\end{align}
\] (8b)

Here, \( C_{ij} \) represents the cost of altering the input variables from \( x_i \) to \( \tilde{x}_j \) given a user-specified cost function that we will discuss shortly; \( p, q \) are the empirical estimates of \( P_0 \) and \( Q_X \), respectively:

\[
p_i = \frac{1}{n_0} \sum_{x \in D_0} \delta_{x_i}(x), \quad q_j = \frac{1}{n_0} \sum_{\tilde{x} \in \tilde{D}_0} \delta_{\tilde{x}_j}(\tilde{x});
\]

The optimal transport problem in (8) is a standard linear program that aims to find a coupling of \( p \) and \( q, \gamma \) (see e.g., Peyré & Cuturi, 2017; Villani, 2008). Formally, a coupling is a joint probability distribution with marginal distributions specified by \( p \) and \( q \). Given the minimal-cost coupling \( \gamma^* \), one can construct a (randomized) preprocessor \( T(\cdot) \) which takes a sample \( x_i \) and returns an altered sample \( \tilde{x}_j \) with probability \( \gamma_{ij}^*/p_i \).

We note that the linear programming formulation in (8) is designed for settings with discrete input distributions. In settings when the distributions \( P_0 \) and \( Q_X \) are continuous, an analogous optimal transport problem can be formulated and solved with other approaches (see e.g., Benamou & Brenier, 2000; Angenent et al., 2003).
Avoiding Disparate Impact with Counterfactual Distributions

Choice of Cost Function The cost function $C_{ij}$ controls how samples of the target group are perturbed. By default, one could use a standard distance metric such as the $L_2$-norm (e.g., Feldman et al., 2015; Johndrow & Lum, 2017; Del Barrio et al., 2019). However, one could also consider additional criteria to fine-tune the mapping specified by $T(\cdot)$. For example, one can specify a cost function that avoids specific kinds of change by setting the value of $C_{ij}$ to a large constant so as to penalize undesirable mappings (e.g., a mapping that would alter immutable attributes such as marital status Ustun et al., 2019b).

Customization via Constraints Users can also fine-tune the behavior of the preprocessor by adding custom constraints to the feasible region of optimal transport problems as in (8). For example, one can impose constraints on individual fairness to ensure that the repaired classifier will “treat similar individuals similarly” (see e.g., Dwork et al., 2012). This behavior could be induced by including constraints of the form:

$$\frac{1}{2} \sum_{j=1}^{m} \left| \frac{\gamma_{ij}}{p_i} - \frac{\gamma_{lj}}{p_l} \right| \leq d(x_i, x_l) \quad \text{for all } i, l \in [m].$$

Here, the LHS is the total-variation distance (Cover & Thomas, 2012) between the distributions of $T(x_i)$ and $T(x_l)$, and $d(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a distance metric that reflects the similarity between samples.

5. Experiments

In this section, we demonstrate how counterfactual distributions can be used to avoid disparate impact for classifiers on real-world datasets. We include all datasets and scripts to reproduce our results at http://github.com/ustunb/ctfdist.

Setup We aim to recover counterfactual distributions for different disparity metrics in Table 1. To this end, we consider processed versions of the adult dataset (Bache & Lichman, 2013) and the ProPublica compas dataset (Angwin et al., 2016).

For each dataset, we use:

- 30% of samples to train a classifier $h(x)$ to repair;
- 50% of samples to recover a counterfactual distribution via Algorithm 1;
- 20% of samples as a hold-out set to evaluate the performance of the repaired model.

We use $\ell_2$-logistic regression to train a classifier $h(x)$ as well as the classifiers $\hat{y}_0(x)$ and $\hat{s}(x)$ that we use to estimate the influence functions in Algorithm 1. We tune the parameters and estimate the performance of each classifiers using a standard 10-fold CV setup.

Our setup assumes that the data used to train and repair the model are drawn from the same distribution, which may not be the case in settings with dataset shift. Our setup also differs from real-world settings in that we use 70% of the samples in each dataset to estimate the counterfactual distribution. In practice, however, we would use all available samples since we would be given the classifier to repair.

Discussion In Table 2, we show the effectiveness of preprocessors built to mitigate different kinds of disparity for the classifiers trained on the adult and compas datasets.

We build each preprocessor as follows. We first resample data from the target population according to Algorithm 1. This outputs a dataset of samples drawn from the counterfactual distribution. Next, we use the resampled dataset to produce an empirical estimate of the counterfactual distribution $Q_X$. This distribution is then used to obtain the preprocessor by solving a version of (8) with the cost function $C_{ij} = \|x_i - x_j\|_2^2$.

As shown, the approach reduces disparate impact in the target group, while having a minor effect on test accuracy at various decision points across the full ROC curve. Counterfactual distributions provide a way to scrutinize this mapping in greater detail. As shown in Table 3, one can visualize the differences between the observed distribution and counterfactual distribution to understand how the input variable distributions are altered to reduce disparity.

This kind of constrastive analysis may be helpful in understanding the factors that produce performance disparities in the first place. For example, the differences between the observed distribution $P_0$ and the counterfactual distribution $Q_X$ could be used to identify prototypical samples (see e.g., Bien & Tibshirani, 2011; Kim et al., 2016), or to score features in terms of their ability to produce disparities in the deployment population (Datta et al., 2016; 2017; Adler et al., 2018).

6. Conclusion

We have introduced a new distributional paradigm to understand and mitigate disparate impact. Our framework is based on counterfactual distributions, which can be efficiently computed given a fixed model and data from a population of interest. The tools in this work apply to binary classification models. However, our approach can be extended to handle other kinds of supervised learning models.

Limitations Our approach requires collecting data on sensitive attribute, which may infringe privacy (though this is difficult to avoid as discussed in Žliobaitė & Custers, 2016).
Avoiding Disparate Impact with Counterfactual Distributions

<table>
<thead>
<tr>
<th>DATASET</th>
<th>METRIC</th>
<th>TARGET GROUP</th>
<th>ORIGINAL MODEL</th>
<th>REPAIRED MODEL</th>
<th>TARGET GROUP AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>BASELINE GROUP</td>
<td>DISC. GAP</td>
<td>BEFORE REPAIR</td>
</tr>
<tr>
<td>adult</td>
<td>SP</td>
<td>Female</td>
<td>0.696</td>
<td>0.874</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>FNR</td>
<td>Female</td>
<td>0.478</td>
<td>0.639</td>
<td>0.161</td>
</tr>
<tr>
<td>adult</td>
<td>FPR</td>
<td>Male</td>
<td>0.021</td>
<td>0.119</td>
<td>0.098</td>
</tr>
<tr>
<td>compas</td>
<td>SP</td>
<td>White</td>
<td>0.514</td>
<td>0.594</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>FNR</td>
<td>White</td>
<td>0.350</td>
<td>0.487</td>
<td>0.137</td>
</tr>
<tr>
<td>compas</td>
<td>FPR</td>
<td>Non-white</td>
<td>0.190</td>
<td>0.278</td>
<td>0.087</td>
</tr>
</tbody>
</table>

Table 2: Change in disparate impact for classification models for adult and compas when paired with a randomized preprocessor built to mitigate different kinds of disparity. Each row shows the value of a specific performance metric for the classifier over the target and baseline groups (e.g., SP, FNR, and FPR). The target group is defined as the group that attains the less favorable value of the performance metric. The preprocessor aims to reduce to difference in performance metric by randomly perturbing the input variables for individuals in the target group. We also include AUC to show the change in performance due to the randomized preprocessor. All values are computed using a hold-out sample that is not used to train the model or build the preprocessor.

<table>
<thead>
<tr>
<th></th>
<th>OBSERVED</th>
<th>COUNTERFACTUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Married</td>
<td>18%</td>
<td>63%</td>
</tr>
<tr>
<td>Immigrant</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>HighestDegree_is_HS</td>
<td>32%</td>
<td>32%</td>
</tr>
<tr>
<td>HighestDegree_is_AS</td>
<td>7%</td>
<td>8%</td>
</tr>
<tr>
<td>HighestDegree_is_BS</td>
<td>15%</td>
<td>18%</td>
</tr>
<tr>
<td>HighestDegree_is_MSorPhD</td>
<td>6%</td>
<td>7%</td>
</tr>
<tr>
<td>AnyCapitalLoss</td>
<td>3%</td>
<td>5%</td>
</tr>
<tr>
<td>Age &lt;= 30</td>
<td>39%</td>
<td>29%</td>
</tr>
<tr>
<td>WorkHrsPerWeek&lt;40</td>
<td>38%</td>
<td>17%</td>
</tr>
<tr>
<td>JobType_is_WhiteCollar</td>
<td>34%</td>
<td>19%</td>
</tr>
<tr>
<td>JobType_is_BlueCollar</td>
<td>5%</td>
<td>34%</td>
</tr>
<tr>
<td>JobType_is_Specialized</td>
<td>23%</td>
<td>21%</td>
</tr>
<tr>
<td>JobType_is_ArmedOrProtective</td>
<td>1%</td>
<td>2%</td>
</tr>
<tr>
<td>Industry_is_Private</td>
<td>73%</td>
<td>69%</td>
</tr>
<tr>
<td>Industry_is_Government</td>
<td>15%</td>
<td>12%</td>
</tr>
<tr>
<td>Industry_is_SelfEmployed</td>
<td>5%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Table 3: Counterfactual distributions produced using Algorithm 1 for a classifier on adult. We observe that different metrics produce different counterfactual distributions. By comparing the distribution of the target group with the counterfactual distribution, we can evaluate how the repaired classifier will perturb their features to reduce disparity.

Our approach also aims to mitigate performance using a randomized preprocessor. While randomization is a common technique to reduce disparate impact in the literature (see e.g., Hardt et al., 2016; Agarwal et al., 2018), it may not be practical in applications such as loan approval since an applicant could achieve a different predicted outcome by applying multiple times. Some effects of randomization can be mitigated by heuristic strategies. Given that we have considered counterfactual distributions that improve the performance for the target group, however, our approach does have a benefit in that randomization will only apply to individuals in the target group and only be applied in a way that will improve their outcomes.
Acknowledgements
This material is based upon work supported by the National Science Foundation under Grant No. CCF-18-45852.

References


Avoiding Disparate Impact with Counterfactual Distributions


A. Omitted Proofs

A.1. Factorization of Joint Distribution

In this section, we show that the disparity metrics in Table 1 can be expressed in terms of $P_0$ when $P_{Y|X}$, $P_{Y|X,S}$, $P_1$, and $P_S$ are given.

We observe that since our classifier is fixed, the joint distribution $P_{S,X,Y,\hat{Y}}$ is characterized by the graphical model shown in Figure 3. Accordingly, we can express $P_{S,X,Y,\hat{Y}}$ as:

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

Note that $h(x) = P_{\hat{Y}|X}(1|x)$. In what follows, we use these observations to express each of the disparity metrics in Table 1 as $M(P_0)$ (i.e., a function of $P_0$).

1. DA.

$$ D_{KL}(P_{\hat{Y}|S=0}||P_{\hat{Y}|S=1}) + \lambda D_{KL}(P_0||P_1) = D_{KL}(P_{\hat{Y}|X} \circ P_0||P_{\hat{Y}|X} \circ P_1) + \lambda D_{KL}(P_0||P_1). $$

2. SP.

$$ \Pr(\hat{Y} = 0|S = 0) - \Pr(\hat{Y} = 0|S = 1) = \mathbb{E}[(1 - h(X))|S = 0] - \mathbb{E}[(1 - h(X))|S = 1] $$

$$ = \sum_{x \in \mathcal{X}} h(x) P_0(x) + \sum_{x \in \mathcal{X}} h(x) P_1(x). $$

3. FDR.

$$ \Pr(Y = 0|\hat{Y} = 1, S = 0) - \Pr(Y = 0|\hat{Y} = 1, S = 1) $$

$$ = \frac{\Pr(Y = 0, \hat{Y} = 1, S = 0)}{\Pr(\hat{Y} = 1, S = 1)} - \frac{\Pr(Y = 0, \hat{Y} = 1, S = 1)}{\Pr(Y = 1, S = 1)} $$

$$ = \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=0}(0|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_1(x) - \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=1}(0|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_1(x). $$

4. FNR.

$$ \Pr(\hat{Y} = 0|Y = 1, S = 0) - \Pr(\hat{Y} = 0|Y = 1, S = 1) $$

$$ = \sum_{x \in \mathcal{X}} P_{Y|X}(0|x) P_{Y|X,S=0}(1|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=0}(1|x) - \sum_{x \in \mathcal{X}} P_{Y|X}(0|x) P_{Y|X,S=1}(1|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=1}(1|x). $$

5. FPR.

$$ \Pr(\hat{Y} = 1|Y = 0, S = 0) - \Pr(\hat{Y} = 1|Y = 0, S = 1) $$

$$ = \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=0}(0|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=0}(0|x) P_0(x) - \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=1}(0|x) P_0(x) \sum_{x \in \mathcal{X}} P_{Y|X}(1|x) P_{Y|X,S=1}(0|x) P_0(x). $$

Figure 3: Graphical model of the framework.
A.2. Example of Counterfactual Distributions

We show that the counterfactual distributions are not always unique.

**Example 2.** We use SP as a disparity metric and set \( X|S = 0 \sim \text{Bernoulli}(0.1) \), \( X|S = 1 \sim \text{Bernoulli}(0.2) \). The classifier is chosen as \( h(0) = h(1) = 0.2 \). In this case, any Bernoulli distribution, including \( P_0 \) and \( P_1 \), over \( \{0, 1\} \) is a counterfactual distribution.

A.3. Proof of Proposition 1

**Proof.** First, the counterfactual distributions under DA or SP always achieve zero of the disparity metric. Hence, \( M(Q_X) > 0 \) happens only if the disparity metric is neither DA nor SP. We assume that \( P_{Y|X,S=0} = P_{Y|X,S=1} \) and \( M(Q_X) > 0 \). In particular, \( |M(P_1)| \geq M(Q_X) > 0 \). Note that the disparity metrics in Table 1 except DA are the form of the discrepancies of performance metrics between two groups. Here the performance metrics for each group only depend on \( P_{Y|X,S=i}, P_{X|S=i} \), and \( P_{Y|X} \). If we assume that \( P_{Y|X,S=0} = P_{Y|X,S=1} \) and set the distribution of target group as \( P_1 \), then the performance metrics achieve the same values for two groups. Hence, \( M(P_1) = 0 \) which contradicts the assumption, so \( P_{Y|X,S=0} \neq P_{Y|X,S=1} \).

A.4. Proof of Proposition 2

**Proof.** First, we define

\[
\Delta(f) \triangleq \lim_{\epsilon \to 0} \frac{M(\tilde{P}_0) - M(P_0)}{\epsilon},
\]

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

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

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

\[
\sum_{x \in \mathcal{X}} \psi(x)P_0(x) = 0,
\]

and

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

In particular, we can choose \( Q(x) = \left( \frac{1}{M_U}f(x) + 1 \right)P_0(x) \), where \( M_U \triangleq \sup\{|f(x)| \mid x \in \mathcal{X}\} + 1 \). Then

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

For simplicity, we use \( P_0 + \epsilon fP_0 \) and \( P_0 + \frac{\epsilon}{M_U}fP_0 \) to represent \( P_0(x) + \epsilon f(x)P_0(x) \) and \( P_0(x) + \frac{\epsilon}{M_U}f(x)P_0(x) \), respectively. Then

\[
\Delta(f) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( M(P_0 + \epsilon fP_0) - M(P_0) \right)
= \lim_{\epsilon \to 0} \frac{M_U}{\epsilon} \left( M \left( P_0 + \frac{\epsilon}{M_U}fP_0 \right) - M(P_0) \right)
= M_U \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( M((1 - \epsilon)P_0 + \epsilon Q) - M(P_0) \right)
= M_U \sum_{x \in \mathcal{X}} \psi(x)Q(x)
= M_U \sum_{x \in \mathcal{X}} \psi(x) \left( \frac{1}{M_U}f(x) + 1 \right) P_0(x)
= \sum_{x \in \mathcal{X}} \psi(x)f(x)P_0(x)
= \mathbb{E} [f(X)\psi(X)|S = 0].
\]
Following from Cauchy-Schwarz inequality,

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

Here the equality can be achieved by choosing

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

A.5. Proof of Proposition 3

Proof. When the disparity metric is a linear combination of \( K \) different disparity metrics:

\[ M(P_0) = \sum_{i=1}^{K} \lambda_i M_i(P_0), \]

the influence function, following from Definition 3, is

\[ \psi(x) = \lim_{\epsilon \to 0} \frac{M((1 - \epsilon)P_0 + \epsilon \delta_x) - M(P_0)}{\epsilon} \]

(16)

\[ = \sum_{i=1}^{K} \lambda_i \lim_{\epsilon \to 0} \frac{M_i((1 - \epsilon)P_0 + \epsilon \delta_x) - M_i(P_0)}{\epsilon} \]

(17)

\[ = \sum_{i=1}^{K} \lambda_i \psi_i(x). \]

(18)

A.6. Proofs of Proposition 4

We prove the closed-form expressions of influence functions provided in Proposition 4 in this section. Again, we view the classifier \( h(x) \) as a conditional distribution \( P_{\hat{Y}|X}(1|x) \).

Proof. Influence function for SP. Recall that

\[ \Pr(\hat{Y} = 0|S = 0) = 1 - \sum_{x \in \mathcal{X}} h(x)P_0(x). \]

When we perturb the distribution \( P_0 \), the classifier \( h(x) \) and \( \Pr(\hat{Y} = 1|S = 1) \) do not change. Therefore,

\[ \psi(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( \sum_{x' \in \mathcal{X}} h(x')(1 - \epsilon)P_0(x') + \epsilon \delta_x(x') \right) - \sum_{x' \in \mathcal{X}} h(x')P_0(x') \]

\[ = -h(x) + \Pr(\hat{Y} = 1|S = 0). \]

Influence function for FNR. Next, we compute the influence function of FNR. Similar analysis holds for FPR and FDR. Due to the factorization of the joint distribution (see Appendix A.1), we have

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

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

\[ \Pr(\hat{Y} = 0|Y = 1, S = 0) = \frac{\sum_{x' \in \mathcal{X}} r_1(x')P_0(x')}{\sum_{x' \in \mathcal{X}} r_2(x')P_0(x')} = \frac{\mathbb{E} [r_1(X)|S = 0]}{\mathbb{E} [r_2(X)|S = 0]}, \]
which implies
\[
M((1-\epsilon)P_0 + \epsilon \delta_x)
\]
\[
= \frac{\sum_{x' \in \chi} r_1(x')((1-\epsilon)P_0(x') + \epsilon \delta_x(x'))}{\sum_{x' \in \chi} r_2(x')((1-\epsilon)P_0(x') + \epsilon \delta_x(x'))} - \text{Pr}(\hat{Y} = 0|Y = 1, S = 1)
\]
\[
= \frac{\mathbb{E}[r_1(X)|S = 0] + \epsilon (r_1(x) - \mathbb{E}[r_1(X)|S = 0])}{\mathbb{E}[r_2(X)|S = 0] + \epsilon (r_2(x) - \mathbb{E}[r_2(X)|S = 0])} - \text{Pr}(\hat{Y} = 0|Y = 1, S = 1).
\]

Therefore,
\[
\psi(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} (M((1-\epsilon)P_0 + \epsilon \delta_x) - M(P_0))
\]
\[
= \frac{\mathbb{E}[r_2(X)|S = 0] r_1(x) - \mathbb{E}[r_1(X)|S = 0] r_2(x)}{\mathbb{E}[r_2(X)|S = 0]^2}
\]
\[
= \frac{\text{Pr}(Y = 1|S = 0) r_1(x) - \text{Pr}(\hat{Y} = 0, Y = 1|S = 0) r_2(x)}{\text{Pr}(Y = 1|S = 0)^2}
\]
\[
= \frac{P_{Y|X}(0|x) P_{Y|X,S=0}(1|x) - \text{Pr}(\hat{Y} = 0|Y = 1, S = 0) P_{Y|X,S=0}(1|x)}{\text{Pr}(Y = 1|S = 0)}.
\]

**Influence function for DA.** We start with computing \(D_{KL}((1-\epsilon)P_0 + \epsilon \delta_x \| P_1)\):

\[
D_{KL}((1-\epsilon)P_0 + \epsilon \delta_x \| P_1) = \sum_{x' \in \chi} ((1-\epsilon)P_0(x') + \epsilon \delta_x(x')) \log \frac{(1-\epsilon)P_0(x') + \epsilon \delta_x(x')}{P_1(x')}
\]
\[
= \sum_{x' \in \chi} (P_0(x') + \epsilon (\delta_x(x') - P_0(x'))) \log \frac{P_0(x')}{P_1(x')} + \log \left( 1 + \frac{\epsilon (\delta_x(x') - P_0(x'))}{P_0(x')} \right)
\]
\[
= \sum_{x' \in \chi} (P_0(x') + \epsilon (\delta_x(x') - P_0(x'))) \log \frac{P_0(x')}{P_1(x')} + \epsilon \frac{\delta_x(x') - P_0(x')}{P_0(x')} + O(\epsilon^2)
\]
\[
= D_{KL}(P_0 \| P_1) + \epsilon \sum_{x' \in \chi} (\delta_x(x') - P_0(x')) \log \frac{P_0(x')}{P_1(x')} + O(\epsilon^2)
\]
\[
= D_{KL}(P_0 \| P_1) + \epsilon \left( \log \frac{P_0(x)}{P_1(x)} - \mathbb{E} \left[ \log \frac{P_0(X)}{P_1(X)} | S = 0 \right] \right) + O(\epsilon^2).
\]

Hence,
\[
\lim_{\epsilon \to 0} \frac{1}{\epsilon} (D_{KL}((1-\epsilon)P_0 + \epsilon \delta_x \| P_1) - D_{KL}(P_0 \| P_1)) = \log \frac{P_0(x)}{P_1(x)} - \mathbb{E} \left[ \log \frac{P_0(X)}{P_1(X)} | S = 0 \right].
\] (19)

Similarly, we have
\[
\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( D_{KL}(P_{Y|X,S=0} + \epsilon P_{Y|X} \circ \delta_x \| P_{Y|S=0}) - D_{KL}(P_{Y|S=0} \| P_{Y|S=1}) \right)
\]
\[
= \sum_{y \in \{0,1\}} (P_{Y|X,S=0}(y) - P_{Y|S=0}(y)) \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)}
\]
\[
= \sum_{y \in \{0,1\}} \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)} P_{Y|X}(y|x) - \mathbb{E} \left[ \log \frac{P_{Y|S=0}(\hat{Y})}{P_{Y|S=1}(\hat{Y})} | S = 0 \right].
\] (20)
Combining (19) with (20), we have

\[ \psi(x) = \sum_{y \in \{0, 1\}} \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)} P_{Y|X}(y|x) - \mathbb{E} \left[ \log \frac{P_{Y|S=0}(\bar{Y})}{P_{Y|S=1}(\bar{Y})} \big| S = 0 \right] \]

\[ + \lambda \left( \log \frac{P_0(x)}{P_1(x)} - \mathbb{E} \left[ \log \frac{P_0(X)}{P_1(X)} \big| S = 0 \right] \right). \]

Note that

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

Hence,

\[ \log \frac{P_0(x)}{P_1(x)} = \mathbb{E} \left[ \log \frac{P_0(X)}{P_1(X)} \big| S = 0 \right] = \mathbb{E} \left[ \log \frac{P_{S|X}(0|x)}{P_{S|X}(1|x)} \big| S = 0 \right] = \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} = \mathbb{E} \left[ \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} \big| S = 0 \right]. \]

Next,

\[ \sum_{y \in \{0, 1\}} \log \frac{P_{Y|S=0}(y)}{P_{Y|S=1}(y)} P_{Y|X}(y|x) - \mathbb{E} \left[ \log \frac{P_{Y|S=0}(\bar{Y})}{P_{Y|S=1}(\bar{Y})} \big| S = 0 \right] \]

\[ = \log \frac{P_{Y|S=0}(1)}{P_{Y|S=1}(1)} P_{Y|X}(1|x) + \log \frac{P_{Y|S=0}(0)}{P_{Y|S=1}(0)} (1 - P_{Y|X}(1|x)) \]

\[ - \mathbb{E} \left[ \log \frac{P_{Y|S=0}(\bar{Y})}{P_{Y|S=1}(\bar{Y})} \big| S = 0 \right] \]

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

\[ + \log \frac{P_{Y|S=0}(0)}{P_{Y|S=1}(0)} - \log \frac{P_{Y|S=0}(0)}{P_{Y|S=1}(0)} P_{Y|S=0}(0) - \log \frac{P_{Y|S=0}(1)}{P_{Y|S=1}(1)} P_{Y|S=0}(1) \]

\[ = \left( \log \frac{P_{Y|S=0}(1)P_{Y|S=1}(0)}{P_{Y|S=1}(1)P_{Y|S=0}(0)} \right) P_{Y|X}(1|x) - \left( \log \frac{P_{Y|S=0}(1)P_{Y|S=1}(0)}{P_{Y|S=1}(1)P_{Y|S=0}(0)} \right) P_{Y|S=0}(1). \]

Therefore, we have

\[ \psi(x) = \left( \log \frac{P_{Y|S=0}(1)P_{Y|S=1}(0)}{P_{Y|S=1}(1)P_{Y|S=0}(0)} \right) (P_{Y|X}(1|x) - P_{Y|S=0}(1)) \]

\[ + \lambda \left( \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} - \mathbb{E} \left[ \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} \big| S = 0 \right] \right). \]

A.7. Convergence

When influence functions are estimated from data, they are subject to estimation error. Next, we present a probabilistic upper bound of the estimation error in terms of the number of samples and the size of the support set.
**Proposition 5.** If \( \hat{s}(x) \) and \( \hat{g}_0(x) \) are the empirical conditional distributions obtained from \( n \) i.i.d. samples, then, with probability at least \( 1 - \beta \),

\[
\left\| \hat{\psi}(x) - \psi(x) \right\|_p \leq O \left( \sqrt{n^{-1} (|X| - \log \beta)} \right).
\]

Here, \( \|f - g\|_1 \triangleq \mathbb{E} \|f(X) - g(X)\| |S = 0 \) denotes the \( \ell_1 \)-norm.

The proof of Proposition 5 relies on the following lemma.

**Lemma 1.** Let \( \hat{\psi}(x) \) and \( \psi(x) \) be the estimated influence function and the true influence function, respectively. If the given disparity metric is \( DA_{\lambda} \),

\[
\left\| \hat{\psi}(x) - \psi(x) \right\|_p \leq O \left( \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_p \right).
\]

For all other disparity metrics in Table 1,

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

Here, \( \|f - g\|_p \triangleq (\mathbb{E} \|f(X) - g(X)\|^p |S = 0) \) denotes the \( \ell_p \)-norm for \( p \geq 1 \).

**Proof.** We denote \( \hat{P} \) and \( \hat{\Pr} \) as estimated probability distribution and probability, respectively. Then we assume that

\[
\left\| \hat{P}_{X|S=0} - P_{X|S=0} \right\|_p \lesssim \left\| \hat{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x) \right\|_p; \tag{22}
\]

\[
\left\| \hat{\Pr}(Y = 1|S = 0) - \Pr(Y = 1|S = 0) \right\|_p \lesssim \left\| \hat{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x) \right\|_p; \tag{23}
\]

\[
\left\| \hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0) - \Pr(\hat{Y} = 0|Y = 1, S = 0) \right\|_p \lesssim \left\| \hat{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x) \right\|_p. \tag{24}
\]

where \( \left\| \hat{P}_{X|S=0} - P_{X|S=0} \right\|_p \triangleq \left( \sum_{x \in X} \left| \hat{P}_{X|S=0}(x) - P_{X|S=0}(x) \right|^p \right)^{1/p} \). We make similar assumptions for \( \hat{P}_{S|X}(1|x) \) (i.e., the \( \ell_p \) distance between \( \hat{P}_{S|X}(1|x) \) and \( P_{S|X}(1|x) \) upper bounds the left-hand side of (22), (23), (24)). 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. **SP.** The influence function under SP is

\[
\psi(x) = -h(x) + \Pr(\hat{Y} = 1|S = 0).
\]

In order to compute the influence function under SP, we only need to estimate \( \Pr(\hat{Y} = 1|S = 0) \) since the classifier \( h(x) \) is given. Estimating the distribution of a Bernoulli random variable is more reliable than estimating the conditional distribution so

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

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

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

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

\[
\mathbb{E}[r_2(X)|S = 0] = \Pr(Y = 1|S = 0),
\]

\[
\mathbb{E}[r_1(X)|S = 0] = \Pr(\hat{Y} = 0, Y = 1|S = 0).
\]
Hence, the influence function under FNR has the following equivalent expression.

\[
\psi(x) = \frac{\Pr(Y = 1|S = 0) P_{Y|x}(0|x) - \Pr(\hat{Y} = 0, Y = 1|S = 0)}{\Pr(Y = 1|S = 0)^2} P_{Y|x,S=0}(1|x) \\
= \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} P_{Y|x,S=0}(1|x).
\]  

(25)

The estimated influence function under FNR is

\[
\hat{\psi}(x) = \frac{P_{Y|x}(0|x) - \hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \hat{P}_{Y|x,S=0}(1|x).
\]  

(26)

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

\[
\|\psi(x) - \hat{\psi}(x)\|_p \\
\leq \left\| \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} (P_{Y|x,S=0}(1|x) - \hat{P}_{Y|x,S=0}(1|x)) \right\|_p \\
+ \left\| \frac{\hat{P}_{Y|x,S=0}(1|x)}{\Pr(Y = 1|S = 0)} \left( \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} \right) \right\|_p \\
\leq \left\| \frac{1}{\Pr(Y = 1|S = 0)} (P_{Y|x,S=0}(1|x) - \hat{P}_{Y|x,S=0}(1|x)) \right\|_p \\
+ \left\| \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} - \frac{P_{Y|x}(0|x) - \hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right\|_p \\
\leq \left\| \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} - \frac{P_{Y|x}(0|x) - \hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right\|_p .
\]  

(27)

Next, we have

\[
\left\| \frac{P_{Y|x}(0|x) - \Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} - \frac{P_{Y|x}(0|x) - \hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right\|_p \\
\leq \left\| \frac{P_{Y|x}(0|x)}{\Pr(Y = 1|S = 0)} - \frac{P_{Y|x}(0|x)}{\Pr(Y = 1|S = 0)} \right\|_p + \left\| \frac{\Pr(\hat{Y} = 0|Y = 1, S = 0)}{\Pr(Y = 1|S = 0)} - \frac{\hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right\|_p \\
\leq \left\| \frac{\Pr(Y = 1|S = 0) - \Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} \right\|_p + \left| \frac{\Pr(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} - \frac{\hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right| \\
\leq \left| \frac{\Pr(Y = 1|S = 0) - \Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} \right| + \left| \frac{\Pr(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} - \frac{\hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right| \\
\leq 2 \left| \frac{\Pr(Y = 1|S = 0) - \Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} \right| + \left| \frac{\Pr(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\Pr(Y = 1|S = 0)} - \frac{\hat{\Pr}(\hat{Y} = 0|Y = 1, S = 0)\Pr(Y = 1|S = 0)}{\hat{\Pr}(Y = 1|S = 0)} \right|.
\]  

(28)
Combining (27) and (28) with the assumptions (23) and (24), we have, for FNR,
\[
\|\hat{\psi}(x) - \psi(x)\|_p \lesssim \|P_{Y|X,S=0}(1|x) - \hat{P}_{Y|X,S=0}(1|x)\|_p.
\]

3. **DA.** The influence function under DA is
\[
\psi(x) = \left( \log \frac{P_{Y|S=0}(1)P_{Y|S=1}(0)}{P_{Y|S=1}(1)P_{Y|S=0}(0)} \right) \left( P_{Y|X}(1|x) - P_{Y|S=0}(1) \right)
+ \lambda \left( \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} - \E \left[ \log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \right] | S = 0 \right).
\]

Since \( h(x) = P_{Y|X}(1|x) \) is a given classifier, estimating
\[
\left( \log \frac{P_{Y|S=0}(1)P_{Y|S=1}(0)}{P_{Y|S=1}(1)P_{Y|S=0}(0)} \right) \left( P_{Y|X}(1|x) - P_{Y|S=0}(1) \right)
\]
is more reliable than estimating
\[
\psi_r(x) \triangleq \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} - \E \left[ \log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \right] | S = 0
= \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} - \sum_{x \in \mathcal{X}} P_{X|S=0}(x) \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)}. \tag{29}
\]

Next, we bound the generalization error of estimating \( \psi_r(x) \). Its estimator is
\[
\hat{\psi}_r(x) = \log \frac{1 - \hat{P}_{S|X}(1|x)}{P_{S|X}(1|x)} - \sum_{x \in \mathcal{X}} \hat{P}_{X|S=0}(x) \log \frac{1 - \hat{P}_{S|X}(1|x)}{P_{S|X}(1|x)}. \tag{30}
\]

Note that, for \( a, b > 0 \),
\[
\left| \log \frac{a}{b} \right| \leq \frac{|a - b|}{\min\{a, b\}}. \tag{31}
\]

Then
\[
\left| \log \frac{1 - \hat{P}_{S|X}(1|x)}{P_{S|X}(1|x)} - \log \frac{1 - P_{S|X}(1|x)}{P_{S|X}(1|x)} \right|
\leq |\hat{P}_{S|X}(1|x) - P_{S|X}(1|x)| \left( \frac{1}{\min\{\hat{P}_{S|X}(1|x), P_{S|X}(1|x)\}} + \frac{1}{\min\{1 - \hat{P}_{S|X}(1|x), 1 - P_{S|X}(1|x)\}} \right)
\leq |\hat{P}_{S|X}(1|x) - P_{S|X}(1|x)| \frac{2}{m_{\mathcal{X}}}, \tag{32}
\]

where \( m_{\mathcal{X}} \) is a constant number:
\[
m_{\mathcal{X}} \triangleq \min \left\{ \left\{ \hat{P}_{S|X}(1|x) | x \in \mathcal{X} \right\} \cup \left\{ P_{S|X}(1|x) | x \in \mathcal{X} \right\} \cup \left\{ 1 - \hat{P}_{S|X}(1|x) | x \in \mathcal{X} \right\} \cup \left\{ 1 - P_{S|X}(1|x) | x \in \mathcal{X} \right\} \right\}.
\]

Also of note, for any \( x \in \mathcal{X} \),
\[
\left| \log \frac{1 - \hat{P}_{S|X}(1|x)}{P_{S|X}(1|x)} \right| \leq \frac{|1 - 2\hat{P}_{S|X}(1|x)|}{\min\left\{ \hat{P}_{S|X}(1|x), 1 - \hat{P}_{S|X}(1|x) \right\}} \leq \frac{1}{m_{\mathcal{X}}}. \tag{33}
\]
Combining (29) and (30) with (32) and (33), we have
\[
\left| \hat{\psi}_r(x) - \psi_r(x) \right| \\
\leq \frac{2}{m_X} \left| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right| + \frac{1}{m_X} \sum_{x \in X} \left| \hat{P}_{X|S=0}(x) - P_{X|S=0}(x) \right| \\
+ \frac{2}{m_X} \sum_{x \in X} \left| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right| P_{X|S=0}(x) \\
= \frac{2}{m_X} \left| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right| + \frac{1}{m_X} \left\| \hat{P}_{X|S=0} - P_{X|S=0} \right\|_1 + \frac{2}{m_X} \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1.
\]
Therefore,
\[
\left\| \hat{\psi}_r(x) - \psi_r(x) \right\|_p \\
\leq \frac{2}{m_X} \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_p + \frac{1}{m_X} \left\| \hat{P}_{X|S=0} - P_{X|S=0} \right\|_1 + \frac{2}{m_X} \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1.
\]
Based on the assumption: \( \left\| \hat{P}_{X|S=0} - P_{X|S=0} \right\|_1 \leq \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1 \), we have
\[
\left\| \hat{\psi}_r(x) - \psi_r(x) \right\|_p \leq \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1 + \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1
\]
Hence, for DA,
\[
\left\| \hat{\psi}(x) - \psi(x) \right\|_p \leq \left\| \hat{P}_{S|X}(1|x) - P_{S|X}(1|x) \right\|_1.
\]
Proposition 5 follows from Lemma 1 and the following large deviation results by Weissman et al. (2003). For all \( \epsilon > 0 \),
\[
\Pr \left( \left\| \hat{P} - P \right\|_1 \geq \epsilon \right) \leq \left( 2^M - 2 \right) \exp \left( -n \phi(\pi_P) \epsilon^2 / 4 \right),
\]
where \( P \) is a probability distribution on the set \([M] \), \( \hat{P} \) is the empirical distribution obtained from \( n \) i.i.d. samples, \( \pi_P \triangleq \max_{M \subseteq [M]} \min(P(M), 1 - P(M)) \),
\[
\tilde{\phi}(\pi_P) \triangleq \begin{cases} \frac{1-2p}{2p} \log \frac{1-p}{p} & p \in [0, 1/2), \\ 2 & p = 1/2, \end{cases}
\]
and \( \left\| \hat{P} - P \right\|_1 \triangleq \sum_{x \in X} |\hat{P}(x) - P(x)| \). Note that \( \tilde{\phi}(\pi_P) \geq 2 \) which implies that
\[
\Pr \left( \left\| \hat{P} - P \right\|_1 \geq \epsilon \right) \leq \exp(M) \exp(-n\epsilon^2 / 2). \tag{34}
\]
Hence, by taking \( P = P_{Y,X|S=0} \), \( M = |Y||X| = 2|X| \) and \( \epsilon = \sqrt{\frac{2}{n} (M - \log \beta)} \), Inequality (34) implies that, with probability at least \( 1 - \beta \),
\[
\left\| \hat{P}_{Y,X|S=0} - P_{Y,X|S=0} \right\|_1 \leq \sqrt{\frac{2}{n} (2|X| - \log \beta)}, \tag{35}
\]
where \( \hat{P}_{Y,X|S=0} \) is the empirical distribution obtained from \( n \) i.i.d. samples. Similarly, with probability at least \( 1 - \beta \),
\[
\left\| \hat{P}_{S,X} - P_{S,X} \right\|_1 \leq \sqrt{\frac{2}{n} (2|X| - \log \beta)}. \tag{36}
\]
Let $\tilde{P}_{Y|X,S=0} = \frac{\tilde{P}_{Y|X,S=0}}{P_{X|S=0}}$ be the empirical conditional distribution obtained from $n$ i.i.d. samples. Then, for the disparity metrics in Table 1 except DA, with probability at least $1 - \beta$,

$$\|\hat{\psi}(x) - \psi(x)\|_1 \lesssim \|\tilde{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x)\|_1 \lesssim \|\tilde{P}_{Y,X|S=0} - P_{Y,X|S=0}\|_1 \lesssim \sqrt{\frac{1}{n} (|X| - \log \beta)}.$$

Here the second inequality holds true because

$$\|\tilde{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x)\|_1 = \sum_{x \in X} P_{X|S=0}(x) \left| \tilde{P}_{Y|X,S=0}(1|x) - P_{Y|X,S=0}(1|x) \right| \leq \|\tilde{P}_{Y,X|S=0} - P_{X|S=0}\|_1 + \sum_{x \in X} \tilde{P}_{Y|X,S=0}(1|x) \left| \tilde{P}_{X|S=0}(x) - P_{X|S=0}(x) \right| \leq \|\tilde{P}_{Y,X|S=0} - P_{Y,X|S=0}\|_1 \lesssim \sqrt{\frac{1}{n} (|X| - \log \beta)}.$$

Similar analysis also holds for DA.

\[\square\]

**B. Supporting Experimental Results**

**B.1. Experiments on Synthetic Datasets**

**Descent Procedure**

*Setup:* We consider a toy problem with 3 binary variables $X = (X_1, X_2, X_3)$. We define $p_i = \Pr(X_i = 1|S = 0)$ and $q_i = \Pr(X_i = 1|S = 1)$, and assume that:

$$(p_1, p_2, p_3) = (0.9, 0.2, 0.2)$$

$$(q_1, q_2, q_3) = (0.1, 0.5, 0.5)$$

Given any value of $X$, we draw the value of $Y$ for using the same distribution for each group, namely:

$$P_{Y|X,S=0}(1|x) = P_{Y|X,S=1}(1|x) = \text{logistic}(5x_1 - 2x_2 - 2x_3).$$

We train a logistic regression model 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.

*Results:* As shown in Figure 2, 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. 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%.

**Joint Proxies**

*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

$$P_0 = \begin{pmatrix}
0.60 & 0.00 \\
0.25 & 0.15
\end{pmatrix}, P_1 = \begin{pmatrix}
0.05 & 0.00 \\
0.20 & 0.75
\end{pmatrix}.$$ (37)
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 $P_{Y|X,S=i}(1|x) = \text{logistic}(6x_1x_2 + x_3)$ for $i = 0, 1$, and fit a logistic regression using 50k samples.

**Results:** The value of $\text{DA}_0$ 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 $\text{DA}_0$ 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 $\text{DA}_0$. At each step, the influence function is computed on the auditing dataset, and applied to both the auditing and the holdout set. Then we use the preprocessor to map samples from $S = 0$ to new samples and $\text{DA}_0$ becomes 0.0%.