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 $P_{\hat{Y}|X}$, $P_{Y|X,S}$, 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 1. Accordingly, we can express $P_{S,X,Y,\hat{Y}}$ as:

$$P_{S,X,Y,\hat{Y}} = P_{\hat{Y}|X} P_{Y|X,S} P_S P_{X|S}. \tag{1}$$

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

1. **DA.**

$$\mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0} \| P_{\hat{Y}|S=1}) + \lambda \mathrm{D}_{\mathrm{KL}}(P_0 \| P_1) = \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|X} \circ P_0 \| P_{\hat{Y}|X} \circ P_1) + \lambda \mathrm{D}_{\mathrm{KL}}(P_0 \| P_1). \tag{2}$$

2. **SP.**

$$\begin{aligned} \Pr(\hat{Y}=0|S=0) - \Pr(\hat{Y}=0|S=1) &= \mathbb{E}\left[(1-h(X))|S=0\right] - \mathbb{E}\left[(1-h(X))|S=1\right] \\ &= -\sum_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) P_0(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) P_1(\boldsymbol{x}). \end{aligned} \tag{3}$$

3. **FDR.**

$$\begin{aligned} &\Pr(Y=0|\hat{Y}=1,S=0) - \Pr(Y=0|\hat{Y}=1,S=1) \\ &= \frac{\Pr(Y=0,\hat{Y}=1,S=0)}{\Pr(\hat{Y}=1,S=0)} - \frac{\Pr(Y=0,\hat{Y}=1,S=1)}{\Pr(\hat{Y}=1,S=1)} \\ &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=0}(0|\boldsymbol{x}) P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_0(\boldsymbol{x})} - \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=1}(0|\boldsymbol{x}) P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_1(\boldsymbol{x})}. \end{aligned} \tag{4}$$

4. **FNR.**

$$\begin{aligned} &\Pr(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=1) \\ &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(0|\boldsymbol{x}) P_{Y|X,S=0}(1|\boldsymbol{x}) P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{Y|X,S=0}(1|\boldsymbol{x}) P_0(\boldsymbol{x})} - \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(0|\boldsymbol{x}) P_{Y|X,S=1}(1|\boldsymbol{x}) P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{Y|X,S=1}(1|\boldsymbol{x}) P_1(\boldsymbol{x})}. \end{aligned} \tag{5}$$

5. **FPR.**

$$\begin{aligned} &\Pr(\hat{Y}=1|Y=0,S=0) - \Pr(\hat{Y}=1|Y=0,S=1) \\ &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=0}(0|\boldsymbol{x}) P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{Y|X,S=0}(0|\boldsymbol{x}) P_0(\boldsymbol{x})} - \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=1}(0|\boldsymbol{x}) P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{Y|X,S=1}(0|\boldsymbol{x}) P_1(\boldsymbol{x})}. \end{aligned} \tag{6}$$

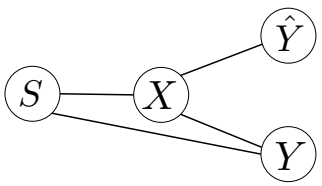

Figure 1: Graphical model of the framework.

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 $X|S=0 \sim \mathsf{Bernoulli}(0.1)$, $X|S=1 \sim \mathsf{Bernoulli}(0.2)$. 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, $\mathsf{M}(Q_X) > 0$ 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 $\mathsf{M}(Q_X) > 0$. In particular, $|\mathsf{M}(P_1)| \geq \mathsf{M}(Q_X) > 0$. 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 $P_{\hat{Y}|X}$. 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, $\mathsf{M}(P_1) = 0$ 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

$$\Delta(f) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_0) - \mathsf{M}(P_0)}{\epsilon}, \tag{7}$$

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

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

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

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

and

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

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

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

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

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

Following from Cauchy-Schwarz inequality,

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

Here the equality can be achieved by choosing

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

□

A.5. Proof of Proposition 3

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

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

the influence function is

$$\psi(\boldsymbol{x}) = \lim_{\epsilon\to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\right) - \mathsf{M}(P_0)}{\epsilon} \tag{8}$$

$$= \sum_{i=1}^{K} \lambda_i \lim_{\epsilon\to 0} \frac{\mathsf{M}_i\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\right) - \mathsf{M}_i(P_0)}{\epsilon} \tag{9}$$

$$= \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x}). \tag{10}$$

□

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 $h(\boldsymbol{x})$ as a conditional distribution $P_{\hat{Y}|X}(1|\boldsymbol{x})$.

Proof. **Influence function for SP.** Recall that

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

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

$$\begin{aligned}\psi(\boldsymbol{x}) &= \lim_{\epsilon\to 0} -\frac{1}{\epsilon}\left(\sum_{\boldsymbol{x}'\in\mathcal{X}} h(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}')) - \sum_{\boldsymbol{x}'\in\mathcal{X}} h(\boldsymbol{x}')P_0(\boldsymbol{x}')\right)\\ &= -h(\boldsymbol{x}) + \Pr(\hat{Y}=1|S=0).\end{aligned}$$

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

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

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

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

which implies

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

Therefore,

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

Influence function for DA. We start with computing $\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\|P_1)$:

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

Hence,

$$
\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\|P_1) - \mathrm{D}_{\mathrm{KL}}(P_0\|P_1)\right) = \log\frac{P_0(\boldsymbol{x})}{P_1(\boldsymbol{x})} - \mathbb{E}\left[\log\frac{P_0(X)}{P_1(X)}\middle| S=0\right]. \tag{11}
$$

Similarly, we have

$$
\begin{aligned}
&\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathrm{D}_{\mathrm{KL}}((1-\epsilon)P_{\hat{Y}|S=0} + \epsilon P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}}\|P_{\hat{Y}|S=1}) - \mathrm{D}_{\mathrm{KL}}(P_{\hat{Y}|S=0}\|P_{\hat{Y}|S=1})\right) \\
&= \sum_{y\in\{0,1\}} ((P_{\hat{Y}|X}\circ\delta_{\boldsymbol{x}})(y) - P_{\hat{Y}|S=0}(y)) \log\frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} \\
&= \sum_{y\in\{0,1\}} \log\frac{P_{\hat{Y}|S=0}(y)}{P_{\hat{Y}|S=1}(y)} P_{\hat{Y}|X}(y|\boldsymbol{x}) - \mathbb{E}\left[\log\frac{P_{\hat{Y}|S=0}(\hat{Y})}{P_{\hat{Y}|S=1}(\hat{Y})}\middle| S=0\right].
\end{aligned} \tag{12}
$$

Combining (11) with (12), we have

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

Note that

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

Hence,

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

Next,

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

Therefore, we have

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

□

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.

Theorem 1. *If $\hat{s}(\boldsymbol{x})$ and $\hat{y}_0(\boldsymbol{x})$ are the empirical conditional distributions obtained from n i.i.d. samples, then, with probability at least $1-\beta$,*

$$\left\|\widehat{\psi}(\boldsymbol{x})-\psi(\boldsymbol{x})\right\|_1 \leq O\left(\sqrt{n^{-1}\left(|\mathcal{X}|-\log\beta\right)}\right). \tag{13}$$

Here, $\|f(\boldsymbol{x})-g(\boldsymbol{x})\|_1 \triangleq \mathbb{E}\left[|f(X)-g(X)||S=0\right]$ denotes the ℓ_1-norm.

The proof of Theorem 1 relies on the following lemma.

Lemma 1. *Let $\widehat{\psi}(\boldsymbol{x})$ and $\psi(\boldsymbol{x})$ be the estimated influence function and the true influence function, respectively. If the given disparity metric is DA$_\lambda$,*

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

For all other disparity metrics in Table 1,

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

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

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

$$\left\|\widehat{P}_{X|S=0}-P_{X|S=0}\right\|_p \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})-P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \tag{14}$$

$$\left|\widehat{\Pr}(Y=1|S=0)-\Pr(Y=1|S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})-P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p; \tag{15}$$

$$\left|\widehat{\Pr}(\hat{Y}=0|Y=1,S=0)-\Pr(\hat{Y}=0|Y=1,S=0)\right| \lesssim \left\|\widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})-P_{Y|X,S=0}(1|\boldsymbol{x})\right\|_p, \tag{16}$$

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 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 (14), (15), (16)). 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

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

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

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

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

   The influence function under FNR is

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

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

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

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

$$\begin{aligned}\psi(\boldsymbol{x}) &= \frac{\Pr(Y=1|S=0)P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0, Y=1|S=0)}{\Pr(Y=1|S=0)^2} P_{Y|X,S=0}(1|\boldsymbol{x}) \\ &= \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} P_{Y|X,S=0}(1|\boldsymbol{x}). \end{aligned} \tag{17}$$

The estimated influence function under FNR is

$$\widehat{\psi}(\boldsymbol{x}) = \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}). \tag{18}$$

Following from (17), (18) and the triangle inequality, we have, under FNR,

$$\begin{aligned}&\|\psi(\boldsymbol{x}) - \widehat{\psi}(\boldsymbol{x})\|_p \\ &\le \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} (P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})) \right\|_p \\ &\quad + \left\| \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \left(\frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} \right.\right. \\ &\qquad \left.\left. - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right) \right\|_p \\ &\le \left\| \frac{1}{\Pr(Y=1|S=0)} (P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x})) \right\|_p \\ &\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p \\ &\lesssim \left\| P_{Y|X,S=0}(1|\boldsymbol{x}) - \widehat{P}_{Y|X,S=0}(1|\boldsymbol{x}) \right\|_p \\ &\quad + \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p . \end{aligned} \tag{19}$$

Next, we have

$$\begin{aligned}&\left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right\|_p \\ &\le \left\| \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\Pr(Y=1|S=0)} - \frac{P_{\hat{Y}|X}(0|\boldsymbol{x})}{\widehat{\Pr}(Y=1|S=0)} \right\|_p + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)}{\Pr(Y=1|S=0)} - \frac{\widehat{\Pr}(\hat{Y}=0|Y=1,S=0)}{\widehat{\Pr}(Y=1|S=0)} \right| \\ &\le \left| \frac{\widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\ &\quad + \left| \frac{\Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0)}{\Pr(Y=1|S=0)\widehat{\Pr}(Y=1|S=0)} \right| \\ &\lesssim \left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| \\ &\quad + \left| \Pr(\hat{Y}=0|Y=1,S=0)\widehat{\Pr}(Y=1|S=0) - \widehat{\Pr}(\hat{Y}=0|Y=1,S=0)\Pr(Y=1|S=0) \right| \\ &\le 2 \left| \widehat{\Pr}(Y=1|S=0) - \Pr(Y=1|S=0) \right| + \left| \widehat{\Pr}(\hat{Y}=0|Y=1,S=0) - \Pr(\hat{Y}=0|Y=1,S=0) \right|. \end{aligned} \tag{20}$$

Combining (19) and (20) with the assumptions (15) and (16), we have, for FNR,

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

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

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

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

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

is more reliable than estimating

$$\begin{aligned}\psi_r(\boldsymbol{x}) &\triangleq \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \mathbb{E}\left[\log \frac{1 - P_{S|X}(1|X)}{P_{S|X}(1|X)} \middle| S = 0 \right] \\ &= \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x} \in \mathcal{X}} P_{X|S=0}(\boldsymbol{x}) \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})}. \end{aligned} \tag{21}$$

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

$$\widehat{\psi}_r(\boldsymbol{x}) = \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \sum_{\boldsymbol{x} \in \mathcal{X}} \widehat{P}_{X|S=0}(\boldsymbol{x}) \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})}. \tag{22}$$

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

$$\left| \log \frac{a}{b} \right| \leq \frac{|a - b|}{\min\{a, b\}}. \tag{23}$$

Then

$$\begin{aligned}&\left| \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} - \log \frac{1 - P_{S|X}(1|\boldsymbol{x})}{P_{S|X}(1|\boldsymbol{x})} \right| \\ &\leq |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})| \left(\frac{1}{\min\{\widehat{P}_{S|X}(1|\boldsymbol{x}), P_{S|X}(1|\boldsymbol{x})\}} + \frac{1}{\min\{1 - \widehat{P}_{S|X}(1|\boldsymbol{x}), 1 - P_{S|X}(1|\boldsymbol{x})\}} \right) \\ &\leq |\widehat{P}_{S|X}(1|\boldsymbol{x}) - P_{S|X}(1|\boldsymbol{x})| \frac{2}{m_X}, \end{aligned} \tag{24}$$

where m_X is a constant number:

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

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

$$\left| \log \frac{1 - \widehat{P}_{S|X}(1|\boldsymbol{x})}{\widehat{P}_{S|X}(1|\boldsymbol{x})} \right| \leq \frac{\left| 1 - 2\widehat{P}_{S|X}(1|\boldsymbol{x}) \right|}{\min\left\{ \widehat{P}_{S|X}(1|\boldsymbol{x}), 1 - \widehat{P}_{S|X}(1|\boldsymbol{x}) \right\}} \leq \frac{1}{m_X}. \tag{25}$$

Combining (21) and (22) with (24) and (25), 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,

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

Theorem 1 follows from Lemma 1 and the following large deviation results by Weissman et al. (2003). For all $\epsilon > 0$,

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

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

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

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

$$\Pr\left(\|\widehat{P} - P\|_1 \geq \epsilon\right) \leq \exp(M)\exp(-n\epsilon^2/2). \tag{26}$$

Hence, by taking $P = P_{Y,X|S=0}$, $M = |\mathcal{Y}||\mathcal{X}| = 2|\mathcal{X}|$ and $\epsilon = \sqrt{\frac{2}{n}(M - \log\beta)}$, Inequality (26) implies that, with probability at least $1 - \beta$,

$$\left\|\widehat{P}_{Y,X|S=0} - P_{Y,X|S=0}\right\|_1 \leq \sqrt{\frac{2}{n}(2|\mathcal{X}| - \log\beta)}, \tag{27}$$

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

$$\left\|\widehat{P}_{S,X} - P_{S,X}\right\|_1 \leq \sqrt{\frac{2}{n}(2|\mathcal{X}| - \log\beta)}. \tag{28}$$

Let $\widehat{P}_{Y|X,S=0} = \frac{\widehat{P}_{Y,X|S=0}}{\widehat{P}_{X|S=0}}$ 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 $1-\beta$,

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

Here the second inequality holds true because

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

Similar analysis also holds for DA.

□

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:

$$\begin{aligned}(p_1, p_2, p_3) &= (0.9, 0.2, 0.2) \\ (q_1, q_2, q_3) &= (0.1, 0.5, 0.5)\end{aligned}$$

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

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

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

$$\mathbf{P}_0 = \begin{pmatrix}0.60 & 0.00\\ 0.25 & 0.15\end{pmatrix}, \mathbf{P}_1 = \begin{pmatrix}0.05 & 0.00\\ 0.20 & 0.75\end{pmatrix}. \tag{29}$$

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

Results: The value of DA_0 is 14.0%. In this case, both X_1 and X_2 are proxy variable. We remove X_1 from dataset and retrain a logistic regression as a classifier. It turns out that the value of DA_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 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 DA_0 becomes 0.0%.

References

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

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.