A APPENDIX

A.1 List of Random Variables Used in the Paper

Table 2: Random variables used in the paper.

Variable	Range	Meaning
X	$\mathcal{X}$	data point (i.e., features representing a data point)
A	$\{0,1\}$	true protected attribute
A_c	$\{0,1\}$	perturbed / corrupted protected attribute
Y	$\{-1,+1\}$	ground-truth label
$\widetilde{Y}$	$\{-1,+1\}$	given predictor for predicting Y
$\widehat{Y}$	$\{-1,+1\}$	EO predictor derived from $\widetilde{Y}$ and based on some protected attribute (i.e., A or A_c)
$\widehat{Y}_\text{corr}$	$\{-1,+1\}$	EO predictor derived from $\widetilde{Y}$ and based on A_c
$\widehat{Y}_\text{true}$	$\{-1,+1\}$	EO predictor derived from $\widetilde{Y}$ and based on A

A.2 Proofs

We first require a simple technical lemma.

Lemma 2. *Let $D = [0,1) \times [0,1) \times (0,1)$ and consider $F : D \rightarrow \mathbb{R}$ with*

$$F(\gamma_1, \gamma_2, p) = \frac{\gamma_1 p}{\gamma_1 p + (1-\gamma_2)(1-p)} - \frac{(1-\gamma_1)p}{(1-\gamma_1)p + \gamma_2(1-p)} + 1. \tag{7}$$

We have:

(i) $0 \leq F(\gamma_1, \gamma_2, p) \leq 2$ *for all* $(\gamma_1, \gamma_2, p) \in D$

(ii) $F(0, 0, p) = 0$ *for all* $p \in (0,1)$

(iii) $F(\gamma_1, \gamma_2, p) < 1$ *for all* $(\gamma_1, \gamma_2, p) \in D$ *with* $\gamma_1 + \gamma_2 < 1$

(iv) $F(\gamma_1, \gamma_2, p) = 1$ *for all* $(\gamma_1, \gamma_2, p) \in D$ *with* $\gamma_1 + \gamma_2 = 1$

(v) $F(\gamma_1, \gamma_2, p) = F(\gamma_2, \gamma_1, 1-p)$ *for all* $(\gamma_1, \gamma_2, p) \in D$

(vi) $\frac{\partial}{\partial \gamma_1} F(\gamma_1, \gamma_2, p) > 0$ *and* $\frac{\partial}{\partial \gamma_2} F(\gamma_1, \gamma_2, p) > 0$ *for all* $(\gamma_1, \gamma_2, p) \in D$

Proof. First note that for $(\gamma_1, \gamma_2, p) \in D$ both denominators are greater than zero and F is well-defined. Both fractions are not smaller than zero and not greater than one, which implies (i). It is trivial to show (ii). It is

$$\frac{\gamma_1 p}{\gamma_1 p + (1-\gamma_2)(1-p)} - \frac{(1-\gamma_1)p}{(1-\gamma_1)p + \gamma_2(1-p)} = \frac{p(1-p)[\gamma_1 + \gamma_2 - 1]}{\big[\gamma_1 p + (1-\gamma_2)(1-p)\big] \cdot \big[(1-\gamma_1)p + \gamma_2(1-p)\big]},$$

from which (iii), (iv) and (v) follow. Finally, it is

$$\begin{aligned}
\frac{\partial}{\partial \gamma_1} F(\gamma_1, \gamma_2, p) &= \frac{\partial}{\partial \gamma_1} \frac{p(1-p)[\gamma_1 + \gamma_2 - 1]}{\big[\gamma_1 p + (1-\gamma_2)(1-p)\big] \cdot \big[(1-\gamma_1)p + \gamma_2(1-p)\big]} \\
&= \frac{p(1-p)\Big[1 - (\gamma_1 + \gamma_2 - 1) \cdot \big\{p \cdot [(1-\gamma_1)p + \gamma_2(1-p)] - p \cdot \big[\gamma_1 p + (1-\gamma_2)(1-p)\big]\big\}\Big]}{\big[\gamma_1 p + (1-\gamma_2)(1-p)\big]^2 \cdot \big[(1-\gamma_1)p + \gamma_2(1-p)\big]^2}.
\end{aligned}$$

We have

$$\begin{aligned}\left|p\cdot[(1-\gamma_1)p+\gamma_2(1-p)]-p\cdot\left[\gamma_1 p+(1-\gamma_2)(1-p)\right]\right| &= |p|\cdot|[p(1-2\gamma_1)+(1-p)(2\gamma_2-1)]|\\ &\leq |p|\end{aligned}$$

for all $(\gamma_1,\gamma_2,p)\in D$ and hence

$$\begin{aligned}&1-(\gamma_1+\gamma_2-1)\cdot\left\{p\cdot[(1-\gamma_1)p+\gamma_2(1-p)]-p\cdot\left[\gamma_1 p+(1-\gamma_2)(1-p)\right]\right\}\geq\\ &\qquad 1-|\gamma_1+\gamma_2-1|\cdot\left|p\cdot[(1-\gamma_1)p+\gamma_2(1-p)]-p\cdot\left[\gamma_1 p+(1-\gamma_2)(1-p)\right]\right|\geq 1-p>0.\end{aligned}$$

This shows $\frac{\partial}{\partial\gamma_1}F(\gamma_1,\gamma_2,p)>0$. It follows from (v) that also $\frac{\partial}{\partial\gamma_2}F(\gamma_1,\gamma_2,p)>0$ for all $(\gamma_1,\gamma_2,p)\in D$. □

Now we can prove Theorem 1.

Proof of Theorem 1:

Let

$$\begin{aligned}\alpha_1 &:= \Pr\left[\widetilde{Y}=1\,\middle|\,Y=1,A=0\right], &\beta_1 &:= \Pr\left[\widetilde{Y}=1\,\middle|\,Y=1,A=1\right],\\ \alpha_2 &:= \Pr\left[\widetilde{Y}=1\,\middle|\,Y=-1,A=0\right], &\beta_2 &:= \Pr\left[\widetilde{Y}=1\,\middle|\,Y=-1,A=1\right].\end{aligned}\tag{8}$$

Then

$$\text{Bias}_{Y=+1}(\widetilde{Y})=|\alpha_1-\beta_1|,\quad \text{Bias}_{Y=-1}(\widetilde{Y})=|\alpha_2-\beta_2|.\tag{9}$$

When computing the probabilities $p_{-1,0},p_{-1,1},p_{1,0},p_{1,1}$ for $\widehat{Y}_{\text{corr}}$, we have to replace $\Pr\left[Y=y,A=a,\widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1\,\middle|\,Y=y,A=a\right]$ by $\Pr\left[Y=y,A_c=a,\widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1\,\middle|\,Y=y,A_c=a\right]$, respectively, in the linear program (2). Note that the assumption $\Pr\left[A_c\neq A\,|\,A=a,Y=y\right]<1$ for $y\in\{-1,+1\}$ and $a\in\{0,1\}$ implies that $\Pr[Y=y,A_c=a]>0$ for $y\in\{-1,+1\}$ and $a\in\{0,1\}$. It is

$$\Pr\left[Y=y,A_c=a,\widetilde{Y}=\tilde{y}\right]=\Pr\left[\widetilde{Y}=\tilde{y}\,\middle|\,Y=y,A_c=a\right]\cdot\Pr\left[Y=y,A_c=a\right]$$

and because of Assumptions I (a), for $a\in\{0,1\}$,

$$\begin{aligned}\Pr\left[\widetilde{Y}=1\,\middle|\,Y=1,A_c=a\right]&=\beta_1\cdot\Pr\left[A=1\,|\,Y=1,A_c=a\right]+\alpha_1\cdot\left(1-\Pr\left[A=1\,|\,Y=1,A_c=a\right]\right),\\ \Pr\left[\widetilde{Y}=1\,\middle|\,Y=-1,A_c=a\right]&=\beta_2\cdot\Pr\left[A=1\,|\,Y=-1,A_c=a\right]+\alpha_2\cdot\left(1-\Pr\left[A=1\,|\,Y=-1,A_c=a\right]\right).\end{aligned}$$

Hence, we end up with the new linear program

$$\begin{aligned}&\min_{\substack{p_{1,0},\ p_{1,1},\\ p_{-1,0},\ p_{-1,1}\in[0,1]}}\ \sum_{\substack{y\in\{-1,+1\}\\ a\in\{0,1\}}}\left\{\Pr\left[Y=-1,A_c=a,\widetilde{Y}=y\right]-\Pr\left[Y=1,A_c=a,\widetilde{Y}=y\right]\right\}\cdot p_{y,a}\\ &\text{s.t.}\quad \{\beta_1\cdot\Pr[A=1\,|\,Y=1,A_c=0]+\alpha_1\cdot(1-\Pr[A=1\,|\,Y=1,A_c=0])\}\cdot p_{1,0}\\ &\qquad\qquad +\{1-\beta_1\cdot\Pr[A=1\,|\,Y=1,A_c=0]-\alpha_1\cdot(1-\Pr[A=1\,|\,Y=1,A_c=0])\}\cdot p_{-1,0}=\\ &\qquad\{\beta_1\cdot\Pr[A=1\,|\,Y=1,A_c=1]+\alpha_1\cdot(1-\Pr[A=1\,|\,Y=1,A_c=1])\}\cdot p_{1,1}\\ &\qquad\qquad +\{1-\beta_1\cdot\Pr[A=1\,|\,Y=1,A_c=1]-\alpha_1\cdot(1-\Pr[A=1\,|\,Y=1,A_c=1])\}\cdot p_{-1,1},\\ &\qquad\{\beta_2\cdot\Pr[A=1\,|\,Y=-1,A_c=0]+\alpha_2\cdot(1-\Pr[A=1\,|\,Y=-1,A_c=0])\}\cdot p_{1,0}\\ &\qquad\qquad +\{1-\beta_2\cdot\Pr[A=1\,|\,Y=-1,A_c=0]-\alpha_2\cdot(1-\Pr[A=1\,|\,Y=-1,A_c=0])\}\cdot p_{-1,0}=\\ &\qquad\{\beta_2\cdot\Pr[A=1\,|\,Y=-1,A_c=1]+\alpha_2\cdot(1-\Pr[A=1\,|\,Y=-1,A_c=1])\}\cdot p_{1,1}\\ &\qquad\qquad +\{1-\beta_2\cdot\Pr[A=1\,|\,Y=-1,A_c=1]-\alpha_2\cdot(1-\Pr[A=1\,|\,Y=-1,A_c=1])\}\cdot p_{-1,1}.\end{aligned}\tag{10}$$

Some elementary calculations yield that the objective function $\Delta = \Delta(p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1})$ in (10) equals

$$\begin{aligned}
\Delta = \; & \Pr[Y=-1, A_c=0] \Big[(p_{1,0}-p_{-1,0}) \cdot \{\alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=0]\} + p_{-1,0}\Big] \\
& + \Pr[Y=-1, A_c=1] \Big[(p_{1,1}-p_{-1,1}) \cdot \{\alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=1]\} + p_{-1,1}\Big] \\
& - \Pr[Y=1, A_c=0] \Big[(p_{1,0}-p_{-1,0}) \cdot \{\alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=0]\} + p_{-1,0}\Big] \\
& - \Pr[Y=1, A_c=1] \Big[(p_{1,1}-p_{-1,1}) \cdot \{\alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=1]\} + p_{-1,1}\Big].
\end{aligned} \tag{11}$$

and that the constraints are equivalent to

$$\begin{aligned}
& (p_{1,0}-p_{-1,0}) \cdot \{\alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=0]\} + p_{-1,0} \\
& \qquad = (p_{1,1}-p_{-1,1}) \cdot \{\alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=1]\} + p_{-1,1}, \\
& (p_{1,0}-p_{-1,0}) \cdot \{\alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=0]\} + p_{-1,0} \\
& \qquad = (p_{1,1}-p_{-1,1}) \cdot \{\alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=1]\} + p_{-1,1}.
\end{aligned} \tag{12}$$

Let

$$e := \alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=0], \tag{13}$$

$$f := \alpha_1 + (\beta_1-\alpha_1) \cdot \Pr[A=1 \,|\, Y=1, A_c=1], \tag{14}$$

$$g := \alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=0], \tag{15}$$

$$h := \alpha_2 + (\beta_2-\alpha_2) \cdot \Pr[A=1 \,|\, Y=-1, A_c=1]. \tag{16}$$

Then the constraints are

$$\begin{aligned}
(p_{1,0}-p_{-1,0}) \cdot e + p_{-1,0} &= (p_{1,1}-p_{-1,1}) \cdot f + p_{-1,1}, \\
(p_{1,0}-p_{-1,0}) \cdot g + p_{-1,0} &= (p_{1,1}-p_{-1,1}) \cdot h + p_{-1,1}.
\end{aligned} \tag{17}$$

Because of the constraints we have

$$\begin{aligned}
\Delta &= p_{-1,0} \cdot \{\Pr[Y=-1] - \Pr[Y=1]\} + (p_{1,0}-p_{-1,0}) \cdot u \\
&= p_{-1,1} \cdot \{\Pr[Y=-1] - \Pr[Y=1]\} + (p_{1,1}-p_{-1,1}) \cdot v,
\end{aligned} \tag{18}$$

where

$$u := g \cdot \Pr[Y=-1] - e \cdot \Pr[Y=1], \qquad v := h \cdot \Pr[Y=-1] - f \cdot \Pr[Y=1]. \tag{19}$$

If $u=0$ or $v=0$, one optimal solution to (10) is $p_{1,0}=p_{1,1}=p_{-1,0}=p_{-1,1}=1$ or $p_{1,0}=p_{1,1}=p_{-1,0}=p_{-1,1}=0$, depending on whether $\Pr[Y=-1] \leq \Pr[Y=1]$ or $\Pr[Y=-1] > \Pr[Y=1]$. In this case the derived equalized odds predictor $\widehat{Y}_{\text{corr}}$ is the constant predictor $\widehat{Y}_{\text{corr}} = +1$ or $\widehat{Y}_{\text{corr}} = -1$ with $\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) = 0$, $y \in \{-1,+1\}$, and (4) is true.

So let us assume that $u \neq 0$ and $v \neq 0$. Let $\theta := \Pr[Y=-1] - \Pr[Y=1]$. Because of

$$\begin{aligned}
\Pr\Big[\widehat{Y}_{\text{corr}} = 1 \,\Big|\, Y=1, A=0\Big] &= p_{1,0} \cdot \alpha_1 + p_{-1,0} \cdot (1-\alpha_1), \\
\Pr\Big[\widehat{Y}_{\text{corr}} = 1 \,\Big|\, Y=1, A=1\Big] &= p_{1,1} \cdot \beta_1 + p_{-1,1} \cdot (1-\beta_1), \\
\Pr\Big[\widehat{Y}_{\text{corr}} = 1 \,\Big|\, Y=-1, A=0\Big] &= p_{1,0} \cdot \alpha_2 + p_{-1,0} \cdot (1-\alpha_2), \\
\Pr\Big[\widehat{Y}_{\text{corr}} = 1 \,\Big|\, Y=-1, A=1\Big] &= p_{1,1} \cdot \beta_2 + p_{-1,1} \cdot (1-\beta_2),
\end{aligned}$$

we have

$$\begin{aligned}
\text{Bias}_{Y=+1}(\widehat{Y}_{\text{corr}}) &= |\alpha_1 \cdot (p_{1,0}-p_{-1,0}) - \beta_1 \cdot (p_{1,1}-p_{-1,1}) + p_{-1,0} - p_{-1,1}|, \\
\text{Bias}_{Y=-1}(\widehat{Y}_{\text{corr}}) &= |\alpha_2 \cdot (p_{1,0}-p_{-1,0}) - \beta_2 \cdot (p_{1,1}-p_{-1,1}) + p_{-1,0} - p_{-1,1}|.
\end{aligned} \tag{20}$$

It is

$$\begin{aligned}\mathrm{Bias}_{Y=+1}(\widehat{Y}_{\mathrm{corr}}) &\overset{(18)}{=} \left|\frac{\Delta\alpha_1}{u} - \frac{\Delta\beta_1}{v} + p_{-1,0}\left(1-\frac{\theta\alpha_1}{u}\right) - p_{-1,1}\left(1-\frac{\theta\beta_1}{v}\right)\right| \\ &= \left|\frac{\Delta\alpha_1}{u} - \frac{\Delta\beta_1}{v} + p_{-1,0}\left(1-\frac{\theta e}{u}\right) - p_{-1,1}\left(1-\frac{\theta f}{v}\right) + p_{-1,0}\frac{\theta(e-\alpha_1)}{u} - p_{-1,1}\frac{\theta(f-\beta_1)}{v}\right|.\end{aligned}$$

From (17) and (18) we obtain that

$$p_{-1,0}\left(1-\frac{\theta e}{u}\right) - p_{-1,1}\left(1-\frac{\theta f}{v}\right) = \frac{\Delta f}{v} - \frac{\Delta e}{u}.$$

From this we get that

$$\begin{aligned}\mathrm{Bias}_{Y=+1}(\widehat{Y}_{\mathrm{corr}}) &= \left|\left(\frac{\Delta}{u} - \frac{p_{-1,0}\theta}{u}\right)(\alpha_1 - e) - \left(\frac{\Delta}{v} - \frac{p_{-1,1}\theta}{v}\right)(\beta_1 - f)\right| \\ &\overset{(13)\&(14)}{=} |\alpha_1 - \beta_1| \cdot \left|\left(\frac{\Delta}{u} - \frac{p_{-1,0}\theta}{u}\right) \cdot \Pr[A=1 \,|\, Y=1, A_{\mathrm{c}}=0] + \left(\frac{\Delta}{v} - \frac{p_{-1,1}\theta}{v}\right) \cdot \Pr[A=0 \,|\, Y=1, A_{\mathrm{c}}=1]\right| \\ &\overset{(18)}{=} |\alpha_1 - \beta_1| \cdot |(p_{1,0} - p_{-1,0}) \cdot \Pr[A=1 \,|\, Y=1, A_{\mathrm{c}}=0] + (p_{1,1} - p_{-1,1}) \cdot \Pr[A=0 \,|\, Y=1, A_{\mathrm{c}}=1]| \\ &\leq |\alpha_1 - \beta_1| \cdot \{\Pr[A=1 \,|\, Y=1, A_{\mathrm{c}}=0] + \Pr[A=0 \,|\, Y=1, A_{\mathrm{c}}=1]\},\end{aligned} \tag{21}$$

where the last inequality follows from the triangle inequality and $|p_{1,0} - p_{-1,0}| \leq 1$ and $|p_{1,1} - p_{-1,1}| \leq 1$ because of $p_{-1,0}, p_{-1,1}, p_{1,0}, p_{1,1} \in [0,1]$.

Similarly, we obtain

$$\mathrm{Bias}_{Y=-1}(\widehat{Y}_{\mathrm{corr}}) \leq |\alpha_2 - \beta_2| \cdot \{\Pr[A=1 \,|\, Y=-1, A_{\mathrm{c}}=0] + \Pr[A=0 \,|\, Y=-1, A_{\mathrm{c}}=1]\}. \tag{22}$$

It is, for $y \in \{-1,+1\}$,

$$\begin{aligned}\Pr[A=1 \,|\, Y=y, A_{\mathrm{c}}=0] &= \frac{\Pr[A=1, A_{\mathrm{c}}=0 \,|\, Y=y]}{\Pr[A_{\mathrm{c}}=0 \,|\, Y=y]} \\ &= \frac{\Pr[A_{\mathrm{c}}=0 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y]}{\Pr[A_{\mathrm{c}}=0 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y] + \Pr[A_{\mathrm{c}}=0 \,|\, Y=y, A=0] \cdot \Pr[A=0 \,|\, Y=y]}\end{aligned} \tag{23}$$

and $\Pr[A=0 \,|\, Y=y, A_{\mathrm{c}}=1] = 1 - \Pr[A=1 \,|\, Y=y, A_{\mathrm{c}}=1]$ with

$$\begin{aligned}\Pr[A=1 \,|\, Y=y, A_{\mathrm{c}}=1] &= \frac{\Pr[A=1, A_{\mathrm{c}}=1 \,|\, Y=y]}{\Pr[A_{\mathrm{c}}=1 \,|\, Y=y]} \\ &= \frac{\Pr[A_{\mathrm{c}}=1 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y]}{\Pr[A_{\mathrm{c}}=1 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y] + \Pr[A_{\mathrm{c}}=1 \,|\, Y=y, A=0] \cdot \Pr[A=0 \,|\, Y=y]}.\end{aligned} \tag{24}$$

Combining (9), (21), (22), (23), (24) and Lemma 2 yields Theorem 1. □

We prove Lemma 1 by means of counterexamples.

Proof of Lemma 1:

- Assumptions I (a) violated & Assumptions I (b) satisfied:

Assume that

$$\begin{gathered}\Pr[Y=y, A=a] = \frac{1}{4}, \quad y \in \{-1,+1\}, a \in \{0,1\}, \\ \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=1, A=0\right] = 0.65, \qquad \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=1, A=1\right] = 0.6, \\ \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=-1, A=0\right] = 0, \qquad \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=-1, A=1\right] = 0\end{gathered} \tag{25}$$

and that

$$\Pr\left[A_c \neq A \,\middle|\, Y=1, A=0, \widetilde{Y}=-1\right] = 0.15, \qquad \Pr\left[A_c \neq A \,\middle|\, Y=y, A=a, \widetilde{Y}=\tilde{y}\right] = 0, \quad (y,a,\tilde{y}) \neq (1,0,-1).$$

Then $\Pr\left[A_c \neq A \,\middle|\, Y=1, A=0\right] = 0.15 \cdot 0.35 = 0.0525$ and $\Pr\left[A_c \neq A \,\middle|\, Y=y, A=a\right] = 0$, $(y,a) \neq (1,0)$, and Assumptions I (b) is satisfied. However, Assumptions I (a) is not satisfied since $\Pr\left[A_c = 1 \,\middle|\, Y=1, A=0, \widetilde{Y}=-1\right] \neq \Pr\left[A_c = 1 \,\middle|\, Y=1, A=0, \widetilde{Y}=1\right]$. It is $\mathrm{Bias}_{Y=+1}(\widetilde{Y}) = 0.05$ and $\mathrm{Bias}_{Y=-1}(\widetilde{Y}) = 0$.

It is straightforward to compute all probabilities $\Pr\left[Y=y, A_c=a, \widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1 \,\middle|\, Y=y, A_c=a\right]$ and solve the the linear program (2) with $\Pr\left[Y=y, A=a, \widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1 \,\middle|\, Y=y, A=a\right]$ replaced by $\Pr\left[Y=y, A_c=a, \widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1 \,\middle|\, Y=y, A_c=a\right]$, respectively. In doing so, one ends up with an optimal solution $(p^*_{-1,0}, p^*_{-1,1}, p^*_{1,0}, p^*_{1,1}) \approx (0, 0, 0.83, 1)$. The bias of the equalized odds predictor $\widehat{Y}_{\mathrm{corr}}$ for the class $Y=+1$ is

$$\begin{aligned}\mathrm{Bias}_{Y=+1}(\widehat{Y}_{\mathrm{corr}}) &= \left|\Pr\left[\widetilde{Y}=1 \,\middle|\, Y=1, A=0\right] \cdot (p^*_{1,0} - p^*_{-1,0}) - \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=1, A=1\right] \cdot (p^*_{1,1} - p^*_{-1,1}) + p^*_{-1,0} - p^*_{-1,1}\right| \\ &\approx |0.65 \cdot 0.83 - 0.6| \approx 0.06 > 0.05 = \mathrm{Bias}_{Y=+1}(\widetilde{Y}).\end{aligned}$$

- Assumptions I (a) satisfied & Assumptions I (b) violated:

The top left plot of Figure 1 in Section 5.1 provides an example where Assumptions I (a) is satisfied and for $\Pr\left[A_c \neq A \,\middle|\, Y=1, A=0\right] = \Pr\left[A_c \neq A \,\middle|\, Y=1, A=1\right] > 0.5$ (and hence Assumptions I (b) being violated) we have $\mathrm{Bias}_{Y=+1}(\widehat{Y}_{\mathrm{corr}}) > \mathrm{Bias}_{Y=+1}(\widetilde{Y})$. □

Next, we prove Theorem 2.

Proof of Theorem 2:

We use the same notation as in the proof of Theorem 1. In particular, let $\alpha_1, \alpha_2, \beta_1, \beta_2$ be the probabilities defined in (8). Since we assume Assumption II to hold, we have $\alpha_1 > \alpha_2$ and $\beta_1 > \beta_2$. Furthermore, without loss of generality, we may assume that $\alpha_2\beta_1 \geq \alpha_1\beta_2$ (otherwise, we can simply swap the role of the groups $A=0$ and $A=1$ so that this condition holds).

Let $\gamma := \Pr\left[A_c \neq A \,|\, A=a, Y=y\right]$, which does not depend on the values of a and y, be the perturbation probability. In the training phase for $\widehat{Y}_{\mathrm{corr}}$ we have $\gamma = \gamma_0$ for some $\gamma_0 \in (0, \frac{1}{2}]$, and in the training phase for $\widehat{Y}_{\mathrm{true}}$ we have $\gamma = 0$.

Note that we have $\Pr[Y=+1] = \Pr[Y=-1] = \frac{1}{2}$. It follows from (13) to (19), (23) and (24) that for any fixed value of the perturbation probability $\gamma \in [0,1]$ the equalized odds method solves the following linear program:

$$\begin{aligned} &\min_{p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1} \in [0,1]} \quad \Delta \\ &\text{s.t.} \quad (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_1 + \gamma\beta_1\} + p_{-1,0} = (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_1 + \gamma\alpha_1\} + p_{-1,1}, \\ &\qquad\ (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_2 + \gamma\beta_2\} + p_{-1,0} = (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_2 + \gamma\alpha_2\} + p_{-1,1}, \end{aligned} \tag{26}$$

where

$$\Delta = (p_{1,0} - p_{-1,0})u = (p_{1,1} - p_{-1,1})v \tag{27}$$

with

$$u = \frac{1}{2}\left[(1-\gamma)(\alpha_2 - \alpha_1) + \gamma(\beta_2 - \beta_1)\right], \qquad v = \frac{1}{2}\left[(1-\gamma)(\beta_2 - \beta_1) + \gamma(\alpha_2 - \alpha_1)\right]. \tag{28}$$

Note that $u < 0$ and $v < 0$ for any $\gamma \in [0, 1]$ because of $\alpha_1 > \alpha_2$ and $\beta_1 > \beta_2$. Since $p_{1,0} = p_{1,1} = p_{-1,0} = p_{-1,1} = 0$ satisfies the constraints in (26) and has objective value $\Delta = 0$, in an equalized odds solution (i.e., an optimal solution to (26)) we must have $\Delta \leq 0$, $p_{-1,0} \leq p_{1,0}$ and $p_{-1,1} \leq p_{1,1}$ for any $\gamma \in [0, 1]$. Furthermore, for $\gamma \in [0, \frac{1}{2}]$ we obtain from the first constraint in (26) that

$$\begin{aligned}
p_{-1,0} - p_{-1,1} &= (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_1 + \gamma\alpha_1\} - (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_1 + \gamma\beta_1\} \\
&\overset{(27)}{=} \frac{\Delta}{v}((1-\gamma)\beta_1 + \gamma\alpha_1) - \frac{\Delta}{u}((1-\gamma)\alpha_1 + \gamma\beta_1) \\
&= \frac{\Delta}{uv}\Big(\beta_1((1-\gamma)u - \gamma v) - \alpha_1((1-\gamma)v - \gamma u)\Big) \\
&\overset{(28)}{=} \frac{\Delta(1-2\gamma)}{2uv}(\alpha_2\beta_1 - \alpha_1\beta_2) \\
&\leq 0,
\end{aligned} \tag{29}$$

where the last inequality holds because of $\Delta \leq 0$, $1 - 2\gamma \geq 0$, $u < 0$, $v < 0$ and $\alpha_2\beta_1 \geq \alpha_1\beta_2$. Hence, in an equalized odds solution, for any $\gamma \in [0, 1/2]$, we must have $p_{-1,0} \leq p_{-1,1}$ and $p_{-1,0} = \min\{p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1}\}$. It is straightforward to check that the error $\text{Error}(\widehat{Y})$ of a derived equalized odds predictor $\widehat{Y}$ with probabilities $p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1}$ is given by

$$\text{Error}(\widehat{Y}) = \frac{1}{4} \cdot \{(p_{1,0} - p_{-1,0})(\alpha_2 - \alpha_1) + (p_{1,1} - p_{-1,1})(\beta_2 - \beta_1)\} + \frac{1}{2} \tag{30}$$

and hence is invariant under translations of the probabilities (compare with the end of Section 2). Hence, without loss of generality, we may assume that $p_{-1,0} = 0$. Substituting in the expressions computed above we get that

$$p_{1,0} \overset{(27)}{=} \frac{\Delta}{u}, \tag{31}$$

$$p_{-1,1} \overset{(29)}{=} \frac{\Delta(1-2\gamma)}{2uv}(\alpha_1\beta_2 - \alpha_2\beta_1), \tag{32}$$

$$p_{1,1} \overset{(27)}{=} \frac{\Delta}{v} + p_{-1,1} = \Delta\left[\frac{1}{v} + \frac{(1-2\gamma)(\alpha_1\beta_2 - \alpha_2\beta_1)}{2uv}\right]. \tag{33}$$

The value of Δ must be the smallest value such that all these three probabilities are in $[0, 1]$. It follows that in an equalized odds solution, for any $\gamma \in [0, \frac{1}{2}]$, either $p_{1,0}$ or $p_{1,1}$ (or both) equals 1 and this depends on the sign of the difference

$$\begin{aligned}
p_{1,0} - p_{1,1} &\overset{(31)\&(33)}{=} \Delta\Big(\frac{1}{u} - \frac{1}{v} - \frac{(1-2\gamma)(\alpha_1\beta_2 - \alpha_2\beta_1)}{2uv}\Big) \\
&\overset{(28)}{=} \frac{\Delta(1-2\gamma)}{2uv}\Big(\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2\Big).
\end{aligned} \tag{34}$$

Importantly, the difference (34) has the same sign for any $\gamma \in [0, \frac{1}{2}]$. We distinguish two cases depending on whether $\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2$ is smaller than zero or not:

Case 1: $\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2 < 0$. In this case, for $\gamma \in [0, \frac{1}{2}]$, the difference (34) is non-negative and we have $p_{1,0} = 1$.

Let $p^0_{1,0}, p^0_{-1,0}, p^0_{1,1}, p^0_{-1,1}$ be an equalized odds solution for $\gamma = 0$ (corresponding to $\widehat{Y}_{\text{true}}$) and $p^{\gamma_0}_{1,0}, p^{\gamma_0}_{-1,0}, p^{\gamma_0}_{1,1}, p^{\gamma_0}_{-1,1}$ be an equalized odds solution for $\gamma = \gamma_0 \in (0, \frac{1}{2}]$ (corresponding to $\widehat{Y}_{\text{corr}}$). It is $p^0_{1,0} = p^{\gamma_0}_{1,0} = 1$ and $p^0_{-1,0} = p^{\gamma_0}_{-1,0} = 0$. It follows from (30) that

$$\text{Error}(\widehat{Y}_{\text{true}}) - \text{Error}(\widehat{Y}_{\text{corr}}) = \frac{1}{4} \cdot \{(p^0_{1,1} - p^0_{-1,1})(\beta_2 - \beta_1) - (p^{\gamma_0}_{1,1} - p^{\gamma_0}_{-1,1})(\beta_2 - \beta_1)\}.$$

Using the fact that $(p^0_{1,0} - p^0_{-1,0})(\alpha_2 - \alpha_1) = (p^0_{1,1} - p^0_{-1,1})(\beta_2 - \beta_1)$, which follows from subtracting the first from the second constraint in (26) with $\gamma = 0$, we get that

$$\text{Error}(\widehat{Y}_{\text{true}}) - \text{Error}(\widehat{Y}_{\text{corr}}) = \frac{1}{4} \cdot \{(\alpha_2 - \alpha_1) - (p^{\gamma_0}_{1,1} - p^{\gamma_0}_{-1,1})(\beta_2 - \beta_1)\}.$$

We write $u(\gamma_0)$ and $v(\gamma_0)$ for u or v with $\gamma = \gamma_0$. Because of $p^{\gamma_0}_{1,0} - p^{\gamma_0}_{-1,0} = 1$, we have that

$$p^{\gamma_0}_{1,1} - p^{\gamma_0}_{-1,1} \overset{(27)}{=} \frac{u(\gamma_0)}{v(\gamma_0)}$$

and hence

$$\mathrm{Error}(\widehat{Y}_{\mathrm{true}}) - \mathrm{Error}(\widehat{Y}_{\mathrm{corr}}) = \frac{1}{4} \cdot \left\{(\alpha_2 - \alpha_1) - \frac{u(\gamma_0)}{v(\gamma_0)}(\beta_2 - \beta_1)\right\} \overset{(28)}{=} \frac{\gamma_0}{4} \frac{(\alpha_2 - \alpha_1)^2 - (\beta_2 - \beta_1)^2}{2v(\gamma_0)}.$$

Because of $\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2 < 0$ and $\alpha_2\beta_1 - \alpha_1\beta_2 \geq 0$, we have $\beta_1 - \beta_2 > \alpha_1 - \alpha_2 > 0$, and because of $v(\gamma_0) < 0$ it follows that

$$\mathrm{Error}(\widehat{Y}_{\mathrm{true}}) - \mathrm{Error}(\widehat{Y}_{\mathrm{corr}}) > 0$$

for all $\gamma_0 \in (0, \frac{1}{2}]$.

Case 2: $\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2 \geq 0$. In this case, for $\gamma \in [0, \frac{1}{2}]$, the difference (34) is non-positive and we have $p_{1,1} = 1$.

As before in Case 1, let $p^0_{1,0}, p^0_{-1,0}, p^0_{1,1}, p^0_{-1,1}$ be an equalized odds solution for $\gamma = 0$ (corresponding to $\widehat{Y}_{\mathrm{true}}$) and $p^{\gamma_0}_{1,0}, p^{\gamma_0}_{-1,0}, p^{\gamma_0}_{1,1}, p^{\gamma_0}_{-1,1}$ be an equalized odds solution for $\gamma = \gamma_0 \in (0, \frac{1}{2}]$ (corresponding to $\widehat{Y}_{\mathrm{corr}}$). It is $p^0_{1,1} = p^{\gamma_0}_{1,1} = 1$ and $p^0_{-1,0} = p^{\gamma_0}_{-1,0} = 0$. Similarly as in Case 1 we obtain that

$$\mathrm{Error}(\widehat{Y}_{\mathrm{true}}) - \mathrm{Error}(\widehat{Y}_{\mathrm{corr}}) = \frac{1}{4}\Big\{2(1 - p^0_{-1,1})(\beta_2 - \beta_1) - (1 - p^{\gamma_0}_{-1,1})(\beta_2 - \beta_1) - \frac{v(\gamma_0)}{u(\gamma_0)}(1 - p^{\gamma_0}_{-1,1})(\alpha_2 - \alpha_1)\Big\}. \tag{35}$$

When $p_{1,1} = 1$, we obtain from (33) that

$$\Delta = \frac{2uv}{2u + (1 - 2\gamma)(\alpha_1\beta_2 - \alpha_2\beta_1)}.$$

This implies that

$$1 - p^{\gamma_0}_{-1,1} \overset{(32)}{=} \frac{2u(\gamma_0)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)} \tag{36}$$

and

$$1 - p^0_{-1,1} \overset{(36)\&(28)}{=} \frac{\alpha_2 - \alpha_1}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1}.$$

Substituting these in (35) we get that

$$\begin{aligned}
\mathrm{Error}(\widehat{Y}_{\mathrm{true}}) - \mathrm{Error}(\widehat{Y}_{\mathrm{corr}}) &= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} - \frac{(\beta_2 - \beta_1)2u(\gamma_0) + (\alpha_2 - \alpha_1)2v(\gamma_0)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)}\right\} \\
&= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} - \frac{\gamma_0(\beta_2 - \beta_1 - \alpha_2 + \alpha_1)^2 + 2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)}\right\} \\
&= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} + \frac{\gamma_0(\beta_2 - \beta_1 - \alpha_2 + \alpha_1)^2 + 2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2)}\right\}.
\end{aligned} \tag{37}$$

Notice that in the second term the denominator is positive. Hence, we get that

$$\frac{\gamma_0(\beta_2 - \beta_1 - \alpha_2 + \alpha_1)^2 + 2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2)} \geq \frac{2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2)},$$

where for $\gamma_0 \in (0, \frac{1}{2}]$ equality holds if and only if $\alpha_1 - \alpha_2 = \beta_1 - \beta_2$. Next, we have that

$$\begin{aligned}
-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2) &= (1 - \gamma_0)(\alpha_1 - \alpha_2) + \gamma_0(\beta_1 - \beta_2) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2) \\
&= \alpha_1 - \alpha_2 + (1 - \gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2) - \gamma_0(\alpha_1 - \alpha_2 + \beta_2 - \beta_1 + \alpha_2\beta_1 - \alpha_1\beta_2).
\end{aligned}$$

Because of $\gamma_0 > 0$, $\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2 \geq 0$ and $\alpha_2\beta_1 - \alpha_1\beta_2 \geq 0$ we obtain that

$$\begin{aligned}
-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2) &\leq \alpha_1 - \alpha_2 + (1 - \gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2) \\
&\leq \alpha_1 - \alpha_2 + (\alpha_2\beta_1 - \alpha_1\beta_2),
\end{aligned}$$

where for $\gamma_0 > 0$ equality holds if and only if $\alpha_2\beta_1 = \alpha_1\beta_2$ and $\alpha_1 - \alpha_2 = \beta_1 - \beta_2$. We conclude that

$$\frac{\gamma_0(\beta_2-\beta_1-\alpha_2+\alpha_1)^2 + 2(\beta_2-\beta_1)(\alpha_2-\alpha_1)}{-2u(\gamma_0)+(1-2\gamma_0)(\alpha_2\beta_1-\alpha_1\beta_2)} \geq 2\frac{(\beta_2-\beta_1)(\alpha_2-\alpha_1)}{\alpha_1-\alpha_2+\alpha_2\beta_1-\alpha_1\beta_2},$$

where equality holds if and only if $\alpha_2\beta_1 = \alpha_1\beta_2$ and $\alpha_1 - \alpha_2 = \beta_1 - \beta_2$. It is not hard to see that $\alpha_1 - \alpha_2 = \beta_1 - \beta_2$ and $\alpha_2\beta_1 = \alpha_1\beta_2$ is equivalent to $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$. It follows from (37) that

$$\mathrm{Error}(\widehat{Y}_{\mathrm{true}}) - \mathrm{Error}(\widehat{Y}_{\mathrm{corr}}) \geq 0,$$

where equality holds if and only if $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$.

Note that in Case 1 we can never have $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$ and that $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$ is equivalent to $\mathrm{Bias}_{Y=+1}(\widetilde{Y}) = \mathrm{Bias}_{Y=-1}(\widetilde{Y}) = 0$ (compare with (9)). Hence, we have proved Theorem 2. □

A.3 Long Version of Section 4 on Related Work

By now, there is a huge body of work on fairness in ML, mainly in supervised learning (e.g., Kamiran and Calders, 2012; Kamishima et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Hardt et al., 2016; Kleinberg et al., 2017; Pleiss et al., 2017; Woodworth et al., 2017; Zafar et al., 2017a,b; Agarwal et al., 2018; Donini et al., 2018; Menon and Williamson, 2018; Xu et al., 2018; Kallus and Zhou, 2019), but more recently also in unsupervised learning (e.g., Chierichetti et al., 2017; Celis et al., 2018; Schmidt et al., 2018; Samadi et al., 2018; Kleindessner et al., 2019a,b; Tantipongpipat et al., 2019). All of these papers assume to know the true value of the protected attribute for every data point. We will discuss some papers not making this assumption below. First we discuss the pieces of work related to the fairness notion of equalized odds, which is central to our paper and one of the most prominent fairness notions in the ML literature (see Verma and Rubin, 2018, for a summary of the various notions and a citation count).

Equalized Odds Our paper builds upon the EO postprocessing method of Hardt et al. (2016) as described in Section 2. Hardt et al. also show how to derive an optimal predictor satisfying the EO criterion based on a biased score function rather than a binary classifier $\widetilde{Y}$. However, in this case the resulting optimization problem is no longer a linear program and it is unclear how to extend our analysis to it. Concurrently with the paper by Hardt et al., the fairness notion of EO has also been proposed by Zafar et al. (2017b) under the name of disparate mistreatment. Zafar et al. incorporate a proxy for the EO criterion into the training phase of a decision boundary-based classifier, which leads to a convex-concave optimization problem and does not come with any theoretical guarantees. The seminal paper of Kleinberg et al. (2017) proves that, except for trivial cases, a classifier cannot satisfy the EO criterion and the fairness notion of calibration within groups at the same time. Subsequently, Pleiss et al. (2017) show how to achieve calibration within groups and a relaxed form of the EO constraints simultaneously. Woodworth et al. (2017) show that postprocessing a Bayes optimal unfair classifier in order to obtain a fair classifier (fair / unfair with respect to the notion of EO) can be highly suboptimal and propose a two-step procedure as remedy. In the first step, some approximate fairness constraints are incorporated into the empirical risk minimization framework to get a classifier that is fair to a non-trivial degree, and in the second step, the EO postprocessing method of Hardt et al. (2016) is used to obtain the final classifier. This procedure is computationally intractable, however, and Woodworth et al. propose the notion of equalized correlations as a relaxation of the notion of EO, which leads to a computationally tractable learning problem. We also want to mention the critical work of Corbett-Davies and Goel (2018), which points out some limitations of prominent group fairness notions based on inframarginal statistics, including equalized odds.

Fairness with Only Limited Information about the Protected Attribute Dwork et al. (2012) phrased the notion of individual fairness mentioned in Section 1, according to which similar data points (as measured by a given metric) should be treated similarly by a randomized classifier. Only recently there have been works studying how to satisfy group fairness criteria when having only limited information about the protected attribute. Most important to mention are the works by Gupta et al. (2018) and Lamy et al. (2019). Gupta et al. (2018) empirically show that when the protected attribute is not known, improving a fairness metric for a proxy of the true attribute can improve the fairness metric for the true attribute. Our paper provides theoretical evidence for their observations. Lamy et al. (2019) study a scenario related to ours and consider training a fair classifier when the protected attribute is corrupted according to a mutually contaminated model (Scott et al., 2013). In their case, training is done by means of constrained empirical risk minimization and requires to solve a non-convex optimization problem. Similarly to our Theorem 1, they show that the bias of a classifier trained with the corrupted attribute grows in a certain way with the amount of corruption (where the bias is defined according to the fairness notions of EO or demographic parity). However, they do not investigate the error of such a classifier. Importantly, Lamy et al. only consider classifiers that do not use the protected attribute when making a prediction for a test point. Also important to mention is the paper by Hashimoto et al. (2018), which uses distributionally robust optimization in order to minimize the worst-case misclassification risk in a χ^2-ball around the data generating distribution. In doing so, under the assumption that the resulting non-convex optimization problem was solved exactly, one provably controls the risk of each protected group without knowing which group a data point belongs to. Hashimoto et al. also show that their approach helps to avoid disparity amplification in a sequential classification setting in which a group's fraction in the data decreases as its misclassification risk increases. As an application of our results, in Section 5.3 / Appendix A.9 we experimentally compare the approach of Hashimoto et al. to the EO method with perturbed attribute information in such a sequential setting. There are a couple of more works that we want to discuss. Botros and Tomczak (2018) propose a variational autoencoder for learning fair representations (Zemel et al., 2013; Louizos et al., 2016) that also works when the protected attribute is only partially observed. Kilbertus et al. (2018) provide an approach to fair classification when users to be classified are not willing to share their protected attribute but only an encrypted version of it. Their approach assumes the existence of a regulator with fairness aims and is based on secure multi-party computation. Chen et al. (2019) study the problem of assessing the demographic disparity of a classifier when the protected attribute is unknown and has to be estimated from data. Coston et al. (2019) study fair classification in a covariate shift setting where the attribute is only available in the source domain but not in the target domain (or the other way round). Finally, we want to mention the recent line of work on *rich subgroup fairness* (Hébert-Johnson et al., 2018; Kearns et al., 2018, 2019). This notion falls between the categories of individual and group fairness in that it requires some statistic to be similar for a *large* (or even infinite) number of subgroups, which are defined via a function class rather than a protected attribute.

A.4 Detailed Expressions Required for the Experiments of Section 5.1

We need to solve the linear program

$$\begin{aligned}
&\min_{\substack{p_{1,0},\, p_{1,1},\\ p_{-1,0},\, p_{-1,1}\in[0,1]}} \sum_{\substack{y\in\{-1,+1\}\\ a\in\{0,1\}}} \left\{\Pr\left[Y=-1, A_c=a, \widetilde{Y}=y\right] - \Pr\left[Y=1, A_c=a, \widetilde{Y}=y\right]\right\}\cdot p_{y,a}\\
&\text{s.t.}\quad \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=y, A_c=0\right]\cdot p_{1,0} + \Pr\left[\widetilde{Y}=-1 \,\middle|\, Y=y, A_c=0\right]\cdot p_{-1,0} =\\
&\qquad\qquad \Pr\left[\widetilde{Y}=1 \,\middle|\, Y=y, A_c=1\right]\cdot p_{1,1} + \Pr\left[\widetilde{Y}=-1 \,\middle|\, Y=y, A_c=1\right]\cdot p_{-1,1}, \quad y\in\{-1,1\},
\end{aligned} \tag{38}$$

where we have to express all coefficients in terms of the problem parameters $\Pr[Y=y, A=a]$ and $\Pr\left[\widetilde{Y}=1|Y=y, A=a\right]$ and the perturbation probabilities $\Pr\left[A_c\neq A|Y=y, A=a\right]$. As in Section 5.1, we let $\gamma_{y,a} := \Pr\left[A_c\neq A|Y=y, A=a\right]$, $y\in\{-1,+1\}, a\in\{0,1\}$. From (11) to (16) in the proof of Theorem 1 we obtain that the objective function equals

$$\begin{aligned}
&\Pr\left[Y=-1, A_c=0\right]\cdot\{p_{1,0}\cdot g + p_{-1,0}\cdot(1-g)\} + \Pr\left[Y=-1, A_c=1\right]\cdot\{p_{1,1}\cdot h + p_{-1,1}\cdot(1-h)\}\\
&- \Pr\left[Y=1, A_c=0\right]\cdot\{p_{1,0}\cdot e + p_{-1,0}\cdot(1-e)\} - \Pr\left[Y=1, A_c=1\right]\cdot\{p_{1,1}\cdot f + p_{-1,1}\cdot(1-f)\}
\end{aligned}$$

and that the constraints are equivalent to

$$\begin{aligned}
p_{1,0}\cdot e + p_{-1,0}\cdot(1-e) &= p_{1,1}\cdot f + p_{-1,1}\cdot(1-f),\\
p_{1,0}\cdot g + p_{-1,0}\cdot(1-g) &= p_{1,1}\cdot h + p_{-1,1}\cdot(1-h)
\end{aligned}$$

with

$$\begin{aligned}
e &:= \alpha_1 + (\beta_1-\alpha_1)\cdot\Pr\left[A=1 \,|\, Y=1, A_c=0\right], &\quad f &:= \alpha_1 + (\beta_1-\alpha_1)\cdot\Pr\left[A=1 \,|\, Y=1, A_c=1\right],\\
g &:= \alpha_2 + (\beta_2-\alpha_2)\cdot\Pr\left[A=1 \,|\, Y=-1, A_c=0\right], &\quad h &:= \alpha_2 + (\beta_2-\alpha_2)\cdot\Pr\left[A=1 \,|\, Y=-1, A_c=1\right]
\end{aligned}$$

and $\alpha_1, \beta_1, \alpha_2, \beta_2$ defined in (8). It is

$$\Pr\left[Y=y, A_c=a\right] = \sum_{a'\in\{0,1\}} \underbrace{\Pr\left[A_c=a \,|\, Y=y, A=a'\right]}_{\gamma_{y,a'} \text{ or } 1-\gamma_{y,a'}}\cdot\Pr\left[Y=y, A=a'\right]$$

and from (23) and (24) in the proof of Theorem 1 we obtain that

$$\begin{aligned}
\Pr[A=1\,|\,Y=y,A_c=0] &= \frac{\gamma_{y,1}\cdot\Pr[A=1,Y=y]}{\gamma_{y,1}\cdot\Pr[A=1,Y=y]+(1-\gamma_{y,0})\cdot\Pr[A=0,Y=y]},\\
\Pr[A=1\,|\,Y=y,A_c=1] &= \frac{(1-\gamma_{y,1})\cdot\Pr[A=1,Y=y]}{(1-\gamma_{y,1})\cdot\Pr[A=1,Y=y]+\gamma_{y,0}\cdot\Pr[A=0,Y=y]}.
\end{aligned}$$

Hence, we have written all coefficients of (38) in terms of the problem parameters and perturbation probabilities.

After solving (38) and obtaining a solution $p_{1,0}$, $p_{1,1}$, $p_{-1,0}$, $p_{-1,1}$, we need to compute the bias and the error of the equalized odds predictor $\widehat{Y}$ that is based on $p_{1,0}$, $p_{1,1}$, $p_{-1,0}$, $p_{-1,1}$. From (20) in the proof of Theorem 1 we obtain that

$$\begin{aligned}
\text{Bias}_{Y=+1} &= \left|\alpha_1\cdot(p_{1,0}-p_{-1,0})-\beta_1\cdot(p_{1,1}-p_{-1,1})+p_{-1,0}-p_{-1,1}\right|,\\
\text{Bias}_{Y=-1} &= \left|\alpha_2\cdot(p_{1,0}-p_{-1,0})-\beta_2\cdot(p_{1,1}-p_{-1,1})+p_{-1,0}-p_{-1,1}\right|.
\end{aligned}$$

It is easy to verify that the error of $\widehat{Y}$ is given by (recall that the error refers to the test error and that in the test phase $\widehat{Y}$ gets to see the true protected attribute)

$$\begin{aligned}
\text{Error}(\widehat{Y}) = \Pr[Y=1] &+ \big\{\alpha_2\Pr[Y=-1,A=0]-\alpha_1\Pr[Y=1,A=0]\big\}\cdot p_{1,0}\\
&+ \big\{\beta_2\Pr[Y=-1,A=1]-\beta_1\Pr[Y=1,A=1]\big\}\cdot p_{1,1}\\
&+ \big\{\Pr[Y=-1,A=0]-\Pr[Y=1,A=0]-\alpha_2\Pr[Y=-1,A=0]+\alpha_1\Pr[Y=1,A=0]\big\}\cdot p_{-1,0}\\
&+ \big\{\Pr[Y=-1,A=1]-\Pr[Y=1,A=1]-\beta_2\Pr[Y=-1,A=1]+\beta_1\Pr[Y=1,A=1]\big\}\cdot p_{-1,1}.
\end{aligned} \tag{39}$$

Finally, we have

$$\text{Bias}_{Y=+1}(\widetilde{Y})=|\alpha_1-\beta_1|,\quad \text{Bias}_{Y=-1}(\widetilde{Y})=|\alpha_2-\beta_2|$$

and (simply set $p_{1,0}=p_{1,1}=1$ and $p_{-1,0}=p_{-1,1}=0$ in (39))

$$\text{Error}(\widetilde{Y})=\Pr[Y=1]+\alpha_2\Pr[Y=-1,A=0]-\alpha_1\Pr[Y=1,A=0]+\beta_2\Pr[Y=-1,A=1]-\beta_1\Pr[Y=1,A=1].$$

A.5 Problem Parameters for the Experiments of Figure 1

Table 3 provides the problem parameters for the experiments shown in Figure 1.

Table 3: Problem parameters for the experiments of Figure 1.

Plot	$\Pr[\widetilde{Y}=1\,\vert\,Y=y,A=a]$				$(\gamma_{1,1},\gamma_{-1,0},\gamma_{-1,1})$
	$y=1$ $a=0$	$y=1$ $a=1$	$y=-1$ $a=0$	$y=-1$ $a=1$	
top left	0.9	0.8	0.4	0.1	$(\gamma_{1,0},\gamma_{1,0},\gamma_{1,0})$
top right	0.9	0.6	0.7	0.1	$(\gamma_{1,0},\frac{\gamma_{1,0}}{2},\frac{\gamma_{1,0}}{2})$
bottom left	0.9	0.6	0.3	0.8	$(\frac{\gamma_{1,0}}{2},\frac{\gamma_{1,0}}{4},\frac{\gamma_{1,0}}{8})$
bottom right	0.9	0.5	0.0	0.4	$(\gamma_{1,0},\gamma_{1,0},\gamma_{1,0})$

A.6 Further Experiments as in Section 5.1

In Figure 3 and Figure 4, we present a number of further experiments as described in Section 5.1. The problem parameters can be read from the titles of the plots and Tables 4 and 5, respectively. In these tables, we also report whether inequality (5) is true or not (with $\widehat{Y}_{\text{true}}$ corresponding to $\widehat{Y}$ for $\gamma_{1,0} = 0$, and $\widehat{Y}_{\text{corr}}$ corresponding to $\widehat{Y}$ for $\gamma_{1,0} = 0.05$). We chose the parameters to be presented here in a rather non-systematic way, but such that (i) the given classifier $\widetilde{Y}$ is biased (i.e., $\text{Bias}_{Y=+1}(\widetilde{Y}) > 0$), (ii) $\widetilde{Y}$ satisfies Assumption II, (iii) we do not only observe constant curves in a plot (i.e., the EO method does not yield the same classifier for all values of $\gamma_{1,0}$), and (iv) the parameters cover a wide range of settings. In these experiments, we make the same observations as in the experiments of Section 5.1, and we obtain further confirmation of the main claims of our paper.

Table 4: Problem parameters for the experiments of Figure 3. We use $r(\gamma_{1,0}) := \min\{2\gamma_{1,0}, 0.8\}$.

Plot	$\Pr[\widetilde{Y} = 1 \mid Y = y, A = a]$				$(\gamma_{1,1}, \gamma_{-1,0}, \gamma_{-1,1})$	(5) is true
	$y = 1$ $a = 0$	$y = 1$ $a = 1$	$y = -1$ $a = 0$	$y = -1$ $a = 1$		
1st row left	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
1st row right	0.8	0.9	0.1	0.0	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$	yes
2nd row left	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, \frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{2})$	yes
2nd row right	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, r(\gamma_{1,0}), r(\gamma_{1,0}))$	yes
3rd row left	0.9	0.6	0.7	0.1	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
3rd row right	0.9	0.4	0.1	0.1	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$	yes
4th row left	0.7	0.9	0.3	0.0	$(\gamma_{1,0}, \frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{2})$	yes
4th row right	0.7	0.9	0.3	0.0	$(\gamma_{1,0}, r(\gamma_{1,0}), r(\gamma_{1,0}))$	yes
5th row left	0.3	0.8	0.1	0.2	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
5th row right	0.3	0.8	0.1	0.2	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
6th row left	0.9	0.6	0.4	0.1	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$	yes
6th row right	0.9	0.6	0.4	0.4	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$	yes
7th row left	0.5	0.8	0.1	0.4	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	**no**
7th row right	0.6	0.8	0.1	0.4	$(\gamma_{1,0}, r(\gamma_{1,0}), r(\gamma_{1,0}))$	**no**

Table 5: Problem parameters for the experiments of Figure 4.

Plot	$\Pr[\widetilde{Y} = 1 \mid Y = y, A = a]$				$(\gamma_{1,1}, \gamma_{-1,0}, \gamma_{-1,1})$	(5) is true
	$y = 1$ $a = 0$	$y = 1$ $a = 1$	$y = -1$ $a = 0$	$y = -1$ $a = 1$		
1st row left	0.6	0.55	0.1	0.3	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
1st row right	0.9	0.6	0.4	0.1	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
2nd row left	1.0	0.8	0.0	0.1	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
2nd row right	0.4	0.95	0.1	0.15	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$	yes
3rd row left	0.3	0.7	0.1	0.5	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$	**no**
3rd row right	0.35	0.95	0.1	0.15	$(\gamma_{1,0}, \frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{2})$	**no**

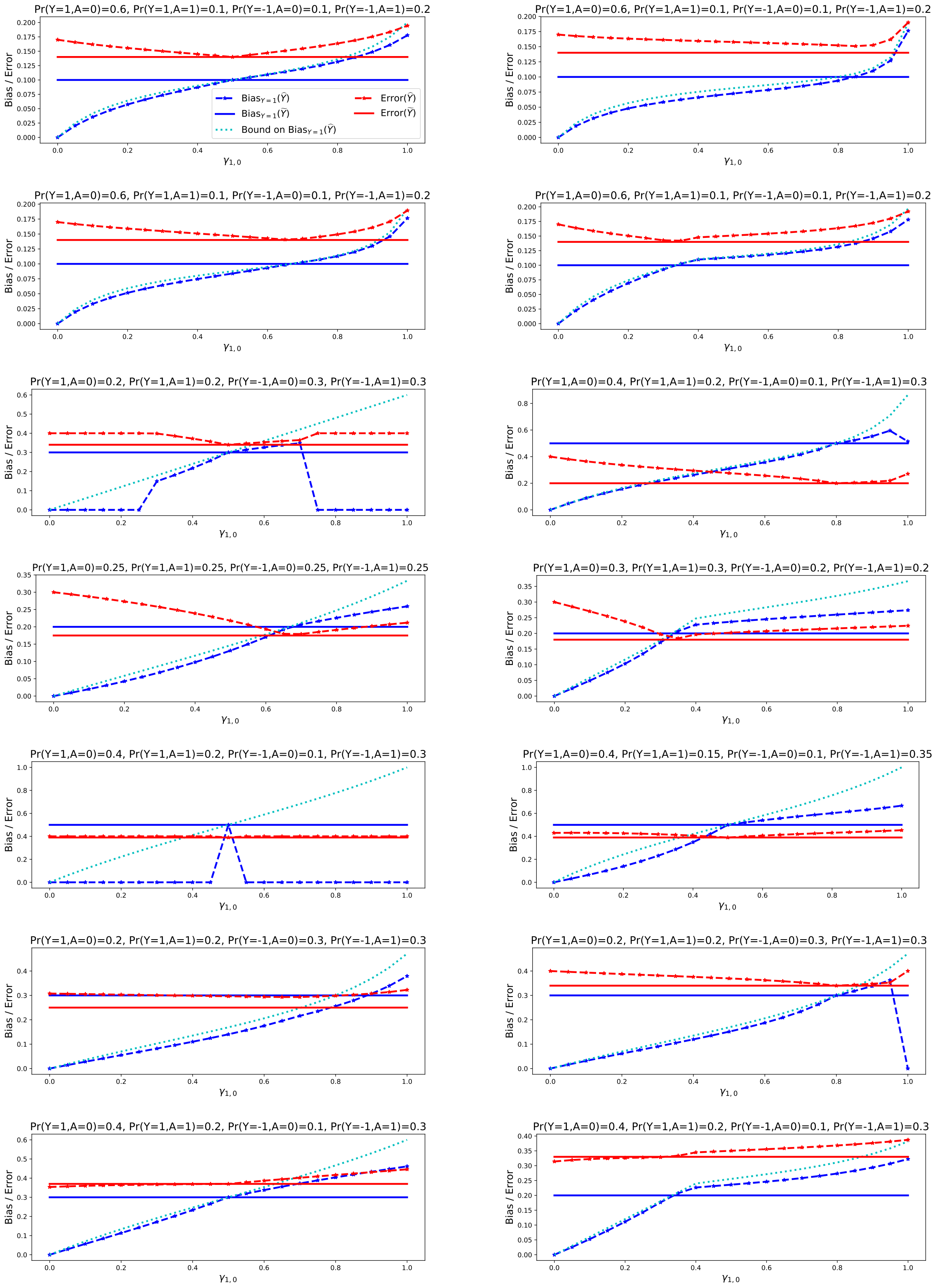

Figure 3: Similar experiments as shown in Figure 1. The dashed blue curve shows $\text{Bias}_{Y=1}(\widehat{Y})$ and the dashed red curve shows $\text{Error}(\widehat{Y})$ as a function of the perturbation level. The solid blue line shows $\text{Bias}_{Y=1}(\widetilde{Y})$ and the solid red line shows $\text{Error}(\widetilde{Y})$. The dotted cyan curve shows the upper bound on $\text{Bias}_{Y=1}(\widehat{Y})$ provided in (4) in Theorem 1. The problem parameters can be read from the titles of the plots and Table 4.

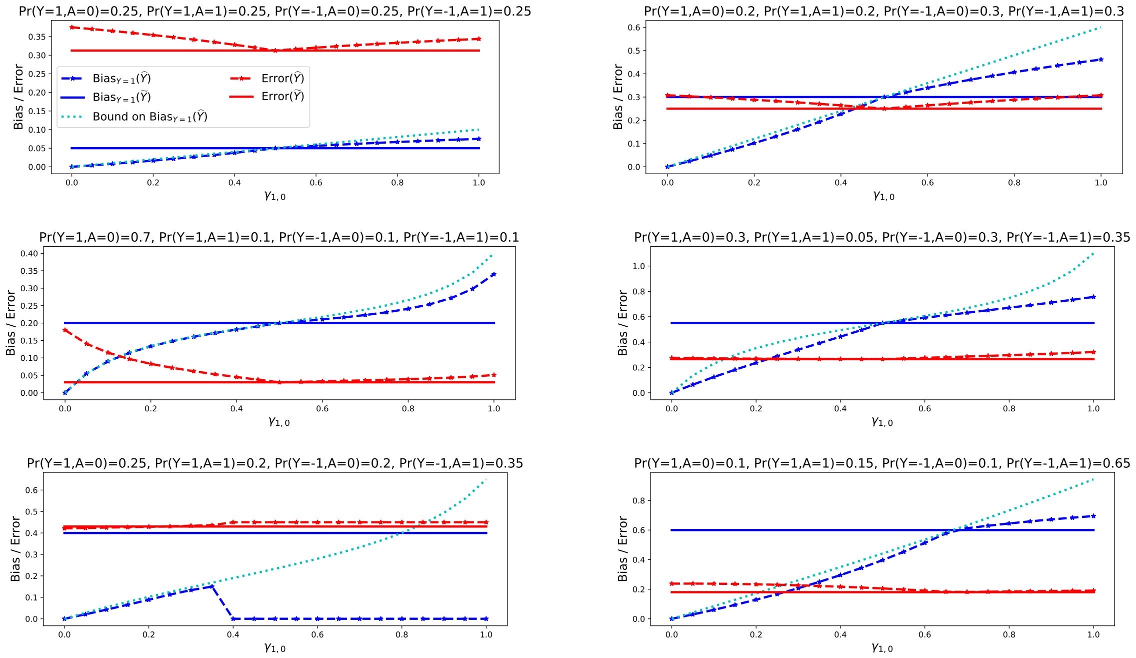

Figure 4: Similar experiments as shown in Figures 1 and 3. The dashed blue curve shows $\mathrm{Bias}_{Y=1}(\widehat{Y})$ and the dashed red curve shows $\mathrm{Error}(\widehat{Y})$ as a function of the perturbation level. The solid blue line shows $\mathrm{Bias}_{Y=1}(\widetilde{Y})$ and the solid red line shows $\mathrm{Error}(\widetilde{Y})$. The dotted cyan curve shows the upper bound on $\mathrm{Bias}_{Y=1}(\widehat{Y})$ provided in (4) in Theorem 1. The problem parameters can be read from the titles of the plots and Table 5.

A.7 Full Table and Additional Statistics of the Experiment on the Drug Consumption Data Set of Section 5.2

Table 7 provides the complete results for the experiment on the drug consumption data set of Section 5.2. Note that we do not consider the drugs Alcohol, Caff, Choc and the fictitious drug Semer since for these drugs it is $\Pr[Y=1] > 0.96$ or $\Pr[Y=1] < 0.01$ and there is a significant chance of observing $\Pr[Y=y, A=a] = 0$ for some $y \in \{-1,+1\}$ and $a \in \{0,1\}$ when working with only a random third of the data set. However, the equalized odds postprocessing method requires $\Pr[Y=y, A=a] > 0$ for $y \in \{-1,+1\}$ and $a \in \{0,1\}$.

Table 6 provides for each drug the number of runs (out of the 200 in total) in which Assumptions I (b) and Assumption II, respectively, is satisfied.

Table 6: Number of runs (out of 200) with Assumptions I (b) / Assumption II being satisfied.

	Amphet	Amyl	Benzos	Cannabis	Coke	Crack	Ecstasy	Heroin
Assumptions I (b)	200	196	198	200	199	191	200	166
Assumption II	200	199	200	200	200	164	200	194
	Ketamine	Legalh	LSD	Meth	Mushroom	Nicotine	VSA	
Assumptions I (b)	169	197	197	164	198	200	197	
Assumption II	191	200	200	200	200	200	197	

.

Table 7: Experiment on the Drug Consumption data set.

Y	$\Pr[Y=1]$	$\text{Bias}_{Y=1/-1}(\widetilde{Y})$	$\text{Bias}_{Y=1/-1}(\widehat{Y}_{\text{corr}})$	$\text{Bias}_{Y=1/-1}(\widehat{Y}_{\text{true}})$	$\text{Error}(\widetilde{Y})$	$\text{Error}(\widehat{Y}_{\text{corr}})$	$\text{Error}(\widehat{Y}_{\text{true}})$	C. I. (6)
Amphet	0.36	0.085 / 0.106	0.076 / 0.065	0.043 / 0.027	0.317	0.339	0.352	0.033
Amyl	0.19	0.08 / 0.032	0.002 / 0.001	0.0 / 0.0	0.226	0.195	0.195	0.032
Benzos	0.41	0.074 / 0.132	0.064 / 0.1	0.041 / 0.034	0.351	0.369	0.39	0.036
Cannabis	0.67	0.092 / 0.052	0.091 / 0.073	0.041 / 0.077	0.214	0.227	0.255	0.032
Coke	0.36	0.075 / 0.107	0.054 / 0.068	0.04 / 0.024	0.331	0.347	0.358	0.032
Crack	0.1	0.075 / 0.025	0.0 / 0.0	0.0 / 0.0	0.129	0.101	0.101	0.039
Ecstasy	0.4	0.095 / 0.117	0.109 / 0.084	0.064 / 0.049	0.294	0.313	0.331	0.032
Heroin	0.11	0.086 / 0.022	0.002 / 0.0	0.0 / 0.0	0.137	0.112	0.112	0.042
Ketamine	0.19	0.067 / 0.043	0.0 / 0.0	0.0 / 0.0	0.236	0.185	0.185	0.035
Legalh	0.4	0.098 / 0.062	0.119 / 0.047	0.071 / 0.044	0.261	0.281	0.289	0.031
LSD	0.29	0.076 / 0.082	0.095 / 0.059	0.061 / 0.032	0.246	0.264	0.279	0.032
Meth	0.22	0.07 / 0.063	0.015 / 0.009	0.003 / 0.002	0.229	0.223	0.223	0.038
Mushroom	0.37	0.084 / 0.106	0.094 / 0.075	0.071 / 0.041	0.279	0.297	0.316	0.031
Nicotine	0.67	0.081 / 0.077	0.041 / 0.047	0.014 / 0.026	0.317	0.329	0.332	0.03
VSA	0.12	0.074 / 0.037	0.0 / 0.0	0.0 / 0.0	0.148	0.12	0.12	0.043

A.8 Plots for the Experiment of Section 5.2 on the Adult Data Set and some Statistics of the COMPAS and Adult Data Sets

Figure 5 provides the plots for the experiment of Section 5.2 on the Adult data set. Table 8 provides several statistics of the COMPAS and Adult data sets (before splitting them into a training and a test set).

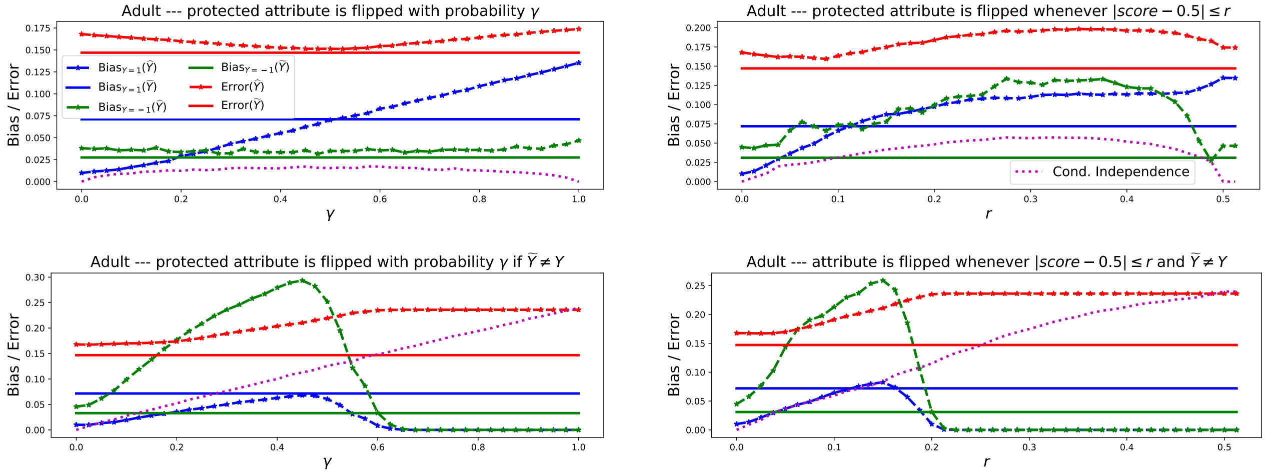

Figure 5: Adult data set. $\text{Bias}_{Y=+1/-1}(\widehat{Y})$ (dashed blue / dashed green) and $\text{Error}(\widehat{Y})$ (dashed red) as a function of the perturbation level in four perturbation scenarios. The solid lines show the bias (blue and green) and the error (red) of $\widetilde{Y}$. The magenta line shows an estimate of (6) and how heavily Assumptions I (a) is violated.

Table 8: Statistics of the real data sets used in Section 5.2.

	COMPAS	Adult
# records	6150	9768
$\frac{\#(Y=1 \wedge A=0)}{\#\text{ records}}$	0.157	0.470
$\frac{\#(Y=1 \wedge A=1)}{\#\text{ records}}$	0.309	0.294
$\frac{\#(Y=-1 \wedge A=0)}{\#\text{ records}}$	0.242	0.201
$\frac{\#(Y=-1 \wedge A=1)}{\#\text{ records}}$	0.292	0.036
$\frac{\#(\widetilde{Y}=1)}{\#\text{ records}}$	0.394	0.795
$\frac{\#(\widetilde{Y} \neq Y)}{\#\text{ records}}$	0.344	0.147
$\frac{\#(\widetilde{Y}=1 \wedge Y=1 \wedge A=0)}{\#(Y=1 \wedge A=0)}$	0.408	0.897
$\frac{\#(\widetilde{Y}=1 \wedge Y=1 \wedge A=1)}{\#(Y=1 \wedge A=1)}$	0.628	0.968
$\frac{\#(\widetilde{Y}=1 \wedge Y=-1 \wedge A=0)}{\#(Y=-1 \wedge A=0)}$	0.147	0.374
$\frac{\#(\widetilde{Y}=1 \wedge Y=-1 \wedge A=1)}{\#(Y=-1 \wedge A=1)}$	0.343	0.398

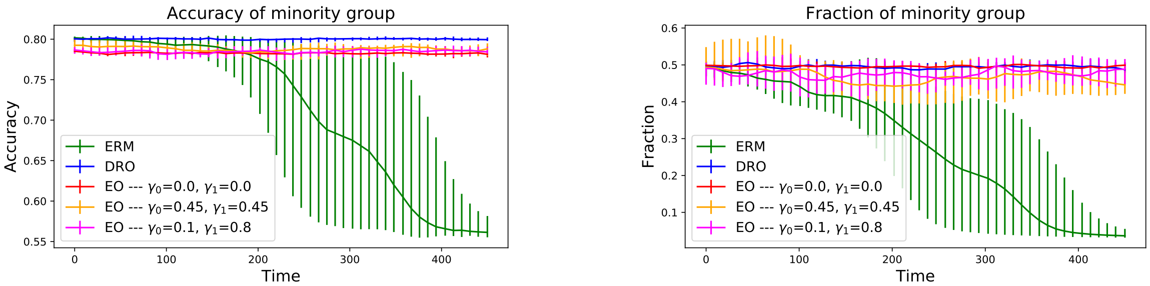

Figure 6: Repeated loss minimization experiment of Hashimoto et al. (2018) (Figure 5 in their paper). Not only the method proposed by Hashimoto et al. (DRO), but also equalized odds postprocessing guarantees high user retention, and hence high accuracy, for both groups over time, even when the protected attribute is highly perturbed. The curves and error bars show the accuracy **(left)** and fraction **(right)** of the minority group over time over 10 replicates of the experiment.

A.9 Repeated Loss Minimization Experiment Outlined in Section 5.3

As another application of our results, we compare the equalized odds postprocessing method to the method of Hashimoto et al. (2018), discussed in Section 4, in the sequential classification setting studied by Hashimoto et al.. In this setting, at each time step a classifier is trained on a data set that comprises several protected groups. The fraction of a group at time step t depends on the group's fraction and the classifier's accuracy for the group at time step $t-1$. Hashimoto et al. show that in such a sequential setting standard empirical risk minimization can lead to disparity amplification with a group having a very small fraction / classification accuracy after some time while their proposed method helps to avoid this situation.

In Figure 6 we present an experiment that reproduces and extends the experiment shown in Figure 5 in Hashimoto et al. (2018).[3] Figure 6 shows the classification accuracy (left plot) and the fraction (right plot) of the minority group over time for various classification strategies. In this experiment, there are only two groups that initially have the same size, and by minority group we mean the group that has a smaller fraction on average over time (hence, at some time steps the fraction of the minority group can be greater than one half). The classification strategies that we consider are all based on logistic regression. ERM refers to a "standard" logistic regression classifier trained with empirical risk minimization and DRO to a logistic regression classifier trained with distributionally robust optimization (the method proposed by Hashimoto et al.; see their paper for details). EO refers to the ERM strategy with equalized odds postprocessing. We consider EO using the true protected attribute and when the true attribute A is perturbed and replaced by A_c, which is obtained by flipping A to its complementary value with probabilities $\gamma_0 := \Pr[A_c \neq A | A=0]$ and $\gamma_1 := \Pr[A_c \neq A | A=1]$, respectively, independently for each data point. We can see from the plots that EO achieves the same goal as DRO, namely avoiding disparity amplification, even when the protected attribute is highly perturbed (orange and magenta curves with $\gamma_0 = \gamma_1 = 0.45$ and $\gamma_0 = 0.1$ / $\gamma_1 = 0.8$, respectively). DRO achieves a slightly higher accuracy, at least in this experiment, and other than EO, it does not require knowledge about the protected attribute at all. However, the underlying optimization problem for DRO is non-convex, and as a result DRO does not come with theoretical per-step guarantees. Hence, we believe that in situations where one has access to a perturbed version of the protected attribute and can assume Assumptions I and II to be satisfied, the equalized odds postprocessing method is a more trustworthy alternative.

[3] We used the code provided by Hashimoto et al. and extended it without changing any parameters.