Effectiveness of Equalized Odds for Fair Classification under Imperfect Group Information

Pranjal Awasthi
Department of Computer Science
Rutgers University
Piscataway, NJ 08854
pranjal.awasthi@rutgers.edu

Matthäus Kleindessner
Department of Computer Science
Rutgers University
Piscataway, NJ 08854
matthaeus.kleindessner@rutgers.edu

Jamie Morgenstern
College of Computing
Georgia Tech
Atlanta, GA 30332
jamiemmt@cs.gatech.edu

Abstract

Most approaches for ensuring or improving a model's fairness with respect to a protected attribute (such as race or gender) assume access to the true value of the protected attribute for every data point. In many scenarios, however, perfect knowledge of the protected attribute is unrealistic. In this paper, we ask to what extent fairness interventions can be effective even with imperfect information about the protected attribute. In particular, we study this question in the context of the prominent equalized odds method of Hardt et al. (2016). We claim that as long as the perturbation of the protected attribute is somewhat moderate, one should still run equalized odds if one would run it knowing the true protected attribute: the bias of the classifier that we obtain using the perturbed attribute is smaller than the bias of the original classifier, and its error is not larger than the error of the equalized odds classifier obtained when working with the true protected attribute.

1 Introduction

As machine learning (ML) algorithms become more mainstream and embedded into our society, evidence has surfaced questioning whether they produce high-quality predictions for most members of diverse populations. The work on *fairness* in machine learning aims to understand the extent to which existing ML methods produce equally high-quality predictions for different individuals, and what new methods can remove the discrepancies therein (Barocas et al., 2018). The appropriate formalization of fair or high-quality predictions necessarily varies based upon the domain, leading to a variety of definitions, largely falling into either the category of *individual fairness* (e.g., Dwork et al., 2012; Dwork and Ilvent, 2018) or *group fairness* (e.g., Kamishima et al., 2012; Hardt et al., 2016; Kleinberg et al., 2017; Pleiss et al., 2017; Zafar et al., 2017a,b). The former focuses on ensuring some property for every individual (and usually is agnostic to any group membership), with the latter asking that some statistic (e.g., accuracy or false positive rate) be similar for different groups. One key drawback of individual fairness is the need for the existence of a similarity metric over the space of individuals. Group fairness, analogously, usually requires knowledge of group membership (such as gender or race), encoded by a *protected attribute*. Arguably a more reasonable requirement than asking for a similarity metric, in many practical applications perfect knowledge of the protected attribute is still an invalid assumption. In this work, we ask to what extent one can guarantee group fairness criteria with only limited information about the protected attribute, generalizing the applicability of such methods.

Our work explores the question of when imperfect or perturbed protected attribute information can be substituted for the true protected attribute into an existing algorithmic framework for fairness with limited harm to the resulting model's fairness and accuracy. In particular, one would never want to end up in a situation where the "fair" classifier obtained from perturbed protected attribute information has worse fairness guarantees than a classifier that ignores fairness altogether, when tested on the true data distribution. In this work we explore the question posed above in the context of fair classification. In particular, due to its simplicity and widespread applicability, we study the prominent postprocessing method of Hardt et al. (2016) for ensuring equalized odds.

Another motivation for studying the robustness of an existing ML method for fairness comes from the fact that an adversary, with the knowledge that the method incorporates fairness, can easily corrupt the data. For example, the adversary could simply change the protected attribute labels of some fraction of the data points. Such corruptions might not be easily detectable via standard methods such as PCA. Hence, it is important to characterize the robustness of existing methods to such perturbations. We give surprisingly strong theoretical and empirical evidence that the equalized odds postprocessing method of Hardt et al. performs well even when based on data with perturbed attribute information.

Our main theoretical result is that as long as the perturbation of the protected attribute in the training data is somewhat moderate (in the balanced case, where all classes and groups have the same size, the attribute of almost half of the data points can be incorrect), the equalized odds postprocessing method of Hardt et al. based on the perturbed attribute produces a classifier $\widehat{Y}$ that is more fair than the original classifier $\widetilde{Y}$. At the same time, under some natural assumptions, the accuracy of $\widehat{Y}$ will never be worse than the classifier obtained from running equalized odds with the true protected attribute. While a similar phenomenon was empirically observed in the recent work of Gupta et al. (2018) (see Section 3 for related work), our work is the first to provide formal guarantees on the effectiveness of a prominent method for fairness in ML even under highly perturbed group information. We further validate our claims empirically under a wide range of settings on both synthetic and real data. We also compare to a group agnostic approach recently proposed by Hashimoto et al. (2018) in a setting of repeated loss minimization.

2 Equalized odds with a perturbation of the protected attribute

We first review the equalized odds postprocessing method of Hardt et al. (2016), assuming the true protected attribute for every data point is known. We then describe our noise model for perturbing the protected attribute and present our analysis of equalized odds under this noise model. Like Hardt et al. and as it is common in the literature on fair machine learning (e.g., Pleiss et al., 2017; Hashimoto et al., 2018), we deal with the distributional setting and ignore the effect of estimating probabilities from finite training samples.

2.1 Equalized odds

Let $X \in \mathcal{X}$, $Y \in \{-1, +1\}$ and $A \in \{0, 1\}$ be random variables with some joint probability distribution. The variable X represents a data point ($\mathcal{X}$ is some suitable set), Y is the data point's ground-truth label and A its protected attribute. Like Hardt et al. (2016), we only consider the case of binary classification and a binary protected attribute. The goal in fair classification is to predict Y from X, or from (X, A), such that the prediction is "fair" with respect to the two groups defined by $A = 0$ and $A = 1$. Think of the standard example of hiring: in this case, X would be a collection of features describing an applicant such as his / her GPA, work experience or language skills, Y would encode whether the applicant is a good fit for the job or not, and A could encode the applicant's gender or skin color. There are numerous formulations of what it means for a prediction to be fair in such an example (some of them contradicting each other; see Section 3), among which the notion of equalized odds as introduced by Hardt et al. is one of the most prominent ones. Denoting the (possibly randomized) prediction by $\widehat{Y} \in \{-1, +1\}$, the prediction satisfies the equalized odds criterion if[1]

$$\Pr\left[\widehat{Y} = 1 \,\middle|\, Y = y, A = 0\right] = \Pr\left[\widehat{Y} = 1 \,\middle|\, Y = y, A = 1\right], \quad y \in \{-1, +1\}. \tag{1}$$

Equation (1) for $y = +1$ requires that $\widehat{Y}$ has equal true positive rates for the two groups $A = 0$ and $A = 1$, and for $y = -1$ it requires $\widehat{Y}$ to have equal false positive rates. In their paper, Hardt et al.

[1]Throughout the paper we assume that $\Pr[Y = y, A = a] > 0$ for $y \in \{-1, +1\}$ and $a \in \{0, 1\}$.

propose a simple postprocessing method to derive a predictor $\widehat{Y}$ that satisfies the equalized odds criterion from a predictor $\widetilde{Y}$ that does not, which works as follows: given a data point with $\widetilde{Y} = y$ and $A = a$, the predictor $\widehat{Y}$ predicts $+1$ with probability $p_{y,a}$ (note that $\widehat{Y}$ depends on X and Y only via $\widetilde{Y}$ and A). The four probabilities $p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1}$ are computed in such a way that (i) $\widehat{Y}$ satisfies the equalized odds criterion, and (ii) the probability of $\widehat{Y}$ not equaling Y is minimized. The former requirement and the latter objective naturally give rise to the following linear program:

$$\begin{aligned}
&\min_{\substack{p_{1,0},\, p_{1,1},\\ p_{-1,0},\, p_{-1,1} \in [0,1]}} \sum_{\substack{y \in \{-1,+1\}\\ a \in \{0,1\}}} \left\{ \Pr\left[Y = -1, A = a, \widetilde{Y} = y\right] - \Pr\left[Y = 1, A = a, \widetilde{Y} = y\right] \right\} \cdot p_{y,a} \\
&\text{s.t. } \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = y, A = 0\right] \cdot p_{1,0} + \Pr\left[\widetilde{Y} = -1 \,\middle|\, Y = y, A = 0\right] \cdot p_{-1,0} = \\
&\qquad \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = y, A = 1\right] \cdot p_{1,1} + \Pr\left[\widetilde{Y} = -1 \,\middle|\, Y = y, A = 1\right] \cdot p_{-1,1}, \quad y \in \{-1, 1\}.
\end{aligned} \tag{2}$$

Note that the linear program (2) does not have a unique solution: rewriting the objective function by exploiting the constraints, it is easy to see that if $p^*_{1,0}, p^*_{1,1}, p^*_{-1,0}, p^*_{-1,1}$ is an optimal solution, then $p^*_{1,0}+c, p^*_{1,1}+c, p^*_{-1,0}+c, p^*_{-1,1}+c$, for any c such that $p^*_{1,0}+c, p^*_{1,1}+c, p^*_{-1,0}+c, p^*_{-1,1}+c \in [0, 1]$, is an optimal solution too, and there might be even more other optimal solutions. Hence, the derived predictor $\widehat{Y}$ is not uniquely defined. We will refer to any predictor that is derived via an optimal solution to (2) as *a* derived equalized odds predictor. Throughout the paper, we will use the terms predictor and classifier interchangeably.

2.2 Noise model for perturbing the protected attribute

When deriving an equalized odds predictor $\widehat{Y}$ from a given classifier $\widetilde{Y}$ one has to estimate the probabilities $\Pr\left[Y = y', A = a, \widetilde{Y} = y\right]$ and $\Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A = a\right]$ that appear in the linear program (2) from training data and then solve the resulting linear program (2) for some optimal probabilities $p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1}$. This is the *training phase* in the equalized odds procedure. In the *test phase*, when applying $\widehat{Y}$ in order to predict the ground-truth label of a test point, one just needs to toss a coin and output a label estimate of $+1$ with probability $p_{y,a}$, or -1 with probability $1 - p_{y,a}$, if $\widetilde{Y} = y$ and $A = a$ for the test point.

The noise model that we consider captures the scenario that the protected attribute in the training data has been corrupted. More specifically, we assume that in the training phase the probabilities $\Pr\left[Y = y', A = a, \widetilde{Y} = y\right]$ and $\Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A = a\right]$ are replaced by $\Pr\left[Y = y', A_{\text{sw}} = a, \widetilde{Y} = y\right]$ and $\Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A_{\text{sw}} = a\right]$, respectively. The random variable A_{sw} denotes the perturbed, or corrupted, protected attribute. We assume that given the ground-truth label Y and the true protected attribute A, the prediction $\widetilde{Y}$ and the corrupted attribute A_{sw} are independent, that is we have, for all $y', y \in \{-1, +1\}$ and $a', a \in \{0, 1\}$,

$$\begin{aligned}
\Pr\left[\widetilde{Y} = y', A_{\text{sw}} = a' \,\middle|\, Y = y, A = a\right] = \Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A = a\right] \cdot \\
\Pr\left[A_{\text{sw}} = a' \,\middle|\, Y = y, A = a\right].
\end{aligned} \tag{3}$$

Other than in the training phase, in the test phase we assume that we have access to the true protected attribute A without any corruption. Hence, the probabilities $p_{y,a}$ of a derived equalized odds predictor for predicting $+1$ depend upon the perturbed protected attribute, but the predictions themselves depend on the true protected attribute. Our noise model applies to scenarios in which a classifier is trained on unreliable data (e.g., crowdsourced data, data obtained from a third party, or when a classifier predicts the unavailable protected attribute) and then applied to test data for which the protected attribute can easily be verified (for example, when performing in-person hiring).

2.3 Bias and error of a derived equalized odds predictor under perturbation

We define the bias for the class $Y = y$ of a predictor $\widehat{Y}$ as the absolute error in the equalized odds condition (1) for this class, that is

$$\text{Bias}_{Y=y}(\widehat{Y}) = \left|\Pr\left[\widehat{Y} = 1 \,|\, Y = y, A = 0\right] - \Pr\left[\widehat{Y} = 1 \,|\, Y = y, A = 1\right]\right|, \quad y \in \{-1, +1\}. \tag{4}$$

Similarly, we define $\text{Bias}_{Y=y}(\widetilde{Y})$. The error of $\widehat{Y}$ or $\widetilde{Y}$ is simply $\text{Error}(\widehat{Y}) = \Pr\left[\widehat{Y} \neq Y\right]$ and $\text{Error}(\widetilde{Y}) = \Pr\left[\widetilde{Y} \neq Y\right]$, respectively. Note that $\text{Bias}_{Y=y}(\widehat{Y})$ and $\text{Error}(\widehat{Y})$ refer to the bias and error of $\widehat{Y}$ in the test phase, and recall from Section 2.2 that in the test phase, according to our noise model, a derived equalized odds predictor $\widehat{Y}$ always makes its prediction based on $\widetilde{Y}$ and the true protected attribute A, regardless of whether the attribute has been corrupted in the training phase.

Let now $\widehat{Y}_{\text{corr}}$ be a derived equalized odds predictor for which the protected attribute in the equalized odds training phase has been corrupted, that is it is based on the linear program (2) with A replaced by A_{sw}, and $\widehat{Y}_{\text{true}}$ be a derived equalized odds predictor without any corruption. The main claim of our paper is that, under some mild assumptions, we have

$$\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) < \text{Bias}_{Y=y}(\widetilde{Y}), \quad y \in \{-1, +1\}, \tag{5}$$

where $\widetilde{Y}$ is the given predictor from which $\widehat{Y}_{\text{corr}}$ and $\widehat{Y}_{\text{true}}$ are derived, and that

$$\text{Error}(\widehat{Y}_{\text{corr}}) \leq \text{Error}(\widehat{Y}_{\text{true}}). \tag{6}$$

Our claim states a beneficial property of the equalized odds method: if one is willing to pay the *price of fairness* (i.e., a loss in prediction accuracy) and would run equalized odds when being guaranteed to observe the true protected attribute, one should also run it when the protected attribute in the training phase might have been corrupted. By running the equalized odds method, one still reduces the bias of $\widetilde{Y}$ while increasing its error by not more than what one is willing to pay for fairness. Actually, our proofs and experiments show that the bias and the error of $\widehat{Y}_{\text{corr}}$ interpolate nicely between those of $\widetilde{Y}$ and $\widehat{Y}_{\text{true}}$.

We begin with relating the bias of $\widehat{Y}_{\text{corr}}$ to the bias of $\widetilde{Y}$ in the following theorem. Recall from Section 2.1 that a derived equalized odds predictor is not uniquely defined.

Theorem 1 (Bias of $\widehat{Y}_{\text{corr}}$ vs. bias of $\widetilde{Y}$)**.** *Assume that* $Pr\left[A_{sw} = A \,|\, A = a, Y = y\right] > 0$ *for* $y \in \{-1, +1\}$ *and* $a \in \{0, 1\}$*. We distinguish two cases depending on whether*

$$Pr\left[\widetilde{Y} = 1 \,|\, Y = -1, A_{sw} = a\right] \cdot Pr\left[Y = -1\right] \neq Pr\left[\widetilde{Y} = 1 \,|\, Y = +1, A_{sw} = a\right] \cdot Pr\left[Y = +1\right], \tag{7}$$

for $a \in \{0, 1\}$*, holds or not.*

1. *Assume that (7) holds for* $a \in \{0, 1\}$*. Then, for* $y \in \{-1, +1\}$*, any derived equalized odds predictor* $\widehat{Y}_{corr}$ *satisfies*

$$Bias_{Y=y}(\widehat{Y}_{corr}) \leq Bias_{Y=y}(\widetilde{Y}) \cdot F(Pr\left[A_{sw} = 0 \,|\, A = 1, Y = y\right], Pr\left[A_{sw} = 1 \,|\, A = 0, Y = y\right], Pr\left[A = 1 \,|\, Y = y\right]), \tag{8}$$

   where $F = F(\gamma_1, \gamma_2, p)$ *is some differentiable function that is strictly increasing both in* γ_1 *and in* γ_2 *with* $F(0, 0, p) = 0$ *and* $F(\gamma_1, \gamma_2, p) < 1$ *for all* (γ_1, γ_2, p) *with* $\gamma_1 + \gamma_2 < 1$*.*

2. *In the degenerate case with (7) not being true for* $a \in \{0, 1\}$*, one derived equalized odds predictor is the constant predictor* $\widehat{Y}_{corr} = +1$ *or* $\widehat{Y}_{corr} = -1$ *with* $Bias_{Y=y}(\widehat{Y}_{corr}) = 0$*,* $y \in \{-1, +1\}$*.*

The proof of Theorem 1 can be found in Section A.2 in the appendix. Note that in the non-degenerate case we have $\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) < \text{Bias}_{Y=y}(\widetilde{Y})$ whenever the corruption of the protected attribute in the class $Y = y$ is moderate in the sense that $\Pr\left[A_{\text{sw}} = 1 \,|\, A = 0, Y = y\right] +$

$\Pr[A_{\text{sw}} = 0 \,|\, A = 1, Y = y] < 1$. If $\Pr[A = 0 \,|\, Y = y] = \Pr[A = 1 \,|\, Y = y] = 1/2$, this condition is equivalent to $\Pr[A_{\text{sw}} \neq A] < 1/2$.

Next, we analyze the error of $\widehat{Y}_{\text{corr}}$ and relate it to the error of $\widehat{Y}_{\text{true}}$. We will assume that the given predictor $\widetilde{Y}$ is correlated with the ground-truth label Y in the sense that

$$\Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = 1, A = a\right] > \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = -1, A = a\right], \quad a \in \{0, 1\}. \tag{9}$$

In our experiments in Section 4.1 we will see that assumption (9) is necessary for our claim (6) to hold. For our theoretical analysis we also make two simplifying assumptions. First, we assume a balanced case in which $\Pr[Y = y, A = a] = \frac{1}{4}$, $y \in \{-1, +1\}$, $a \in \{0, 1\}$. Second, we assume that $\Pr[A_{\text{sw}} \neq A \,|\, A = a, Y = y]$ does not depend on the values of a and y (to give an example, this is the case if every protected attribute is flipped independently with the same probability γ). However, our experiments in Section 4.1 show that our claim (6) also holds in the unbalanced case and when $\Pr[A_{\text{sw}} = A \,|\, A = a, Y = y]$ does depend on the values of a and y. Note that in the balanced case the assumption (9) is equivalent to $\Pr\left[\widetilde{Y} \neq Y \,\middle|\, A = a\right] < \frac{1}{2}$, $a \in \{0, 1\}$, that is to $\widetilde{Y}$ being a weak learner for both of the groups $A = 0$ and $A = 1$. We have the following theorem:

Theorem 2 (Error of $\widehat{Y}_{\text{corr}}$ vs. error of $\widetilde{Y}$)**.** *Assume that* $Pr[Y = y, A = a] = \frac{1}{4}$, $y \in \{-1, +1\}$, $a \in \{0, 1\}$, *and that the given classifier* $\widetilde{Y}$ *is a weak learner for both of the groups* $A = 0$ *and* $A = 1$. *Furthermore, assume that* $Pr\,[A_{sw} \neq A \,|\, A = a, Y = y] \in (0, \frac{1}{2}]$ *and that this probability does not depend on* a *and* y. *Then we have for any derived equalized odds predictors* $\widehat{Y}_{corr}$ *and* $\widehat{Y}_{true}$ *that*

$$Error(\widehat{Y}_{corr}) \leq Error(\widehat{Y}_{true}),$$

where the equality holds if and only if the given classifier $\widetilde{Y}$ *is unbiased, that is* $Bias_{Y=+1}(\widetilde{Y}) = Bias_{Y=-1}(\widetilde{Y}) = 0$.

The proof of Theorem 2 can be found in Section A.2 in the appendix.

3 Related work

By now, there is a huge body of work on fairness in ML, mainly in supervised learning (e.g., Kamishima et al., 2012; Kamiran and Calders, 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), but more recently also in unsupervised learning (e.g., Chierichetti et al., 2017; Celis et al., 2018; Samadi et al., 2018; Schmidt et al., 2018; Kleindessner et al., 2019a,b). All of these papers assume to know the true value of the protected attribute for each 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 (see Verma and Rubin, 2018, for a summary of the various notions and a citation count).

Equalized odds Our paper builds upon the equalized odds postprocessing method of Hardt et al. (2016) as described in Section 2.1. Hardt et al. also show how to derive an optimal predictor satisfying the equalized odds criterion based on a biased score function S (with values in $[0, 1]$ expressing the likelihood of $Y = 1$) 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 equalized odds has also been proposed by Zafar et al. (2017b) under the name of disparate mistreatment. Zafar et al. incorporate a proxy for the equalized odds 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. Kleinberg et al. (2017) show that, except for trivial cases, a classifier cannot satisfy the equalized odds criterion and the fairness notion of calibration within groups (Kleinberg et al., 2017) at the same time. Subsequently, Pleiss et al. (2017) show how to achieve calibration within groups and a relaxed form of the equalized odds constraints at the same time. The work of Woodworth et al. (2017) shows that for certain loss functions postprocessing a Bayes optimal unfair classifier does not necessarily lead to a Bayes optimal fair classifier (fair / unfair with respect to the fairness notion of

equalized odds). They propose a two stage procedure where 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 then using the equalized odds postprocessing method to get the final classifier.

Fairness without protected attributes Dwork et al. (2012) phrased the notion of individual fairness already 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 is the work of Gupta et al. (2018). Their paper empirically shows that when the protected attribute is not known, improving a fairness metric for a proxy of the true protected attribute might be a valuable strategy to improve the fairness metric for the true attribute. Also important to mention are the works by Lamy et al. (2019) and Hashimoto et al. (2018). Lamy et al. (2019) study a scenario related to ours and consider training a fair classifier when the protected attribute is corrupted. 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 notion of equalized odds or demographic parity). However, they do not investigate the error / accuracy 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 and pose it as an open question to extend their results to classifiers that do use the protected attribute when making a prediction. In our paper, we study the bias *and* the error of a derived equalized odds predictor under a perturbation of the protected attribute in the training phase of the equalized odds method. A derived equalized odds predictor crucially depends on the protected attribute when making a prediction, and hence our paper addresses the question raised by Lamy et al.. The paper by Hashimoto et al. (2018) 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 (compare with Section 4.3), 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 4.3 we experimentally compare the approach of Hashimoto et al. to the equalized odds method with perturbed protected attribute information in such a sequential setting. The paper by Kilbertus et al. (2018) provides 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. Finally, Coston et al. (2019) study fair classification in a covariate shift setting where the protected attribute is only available in the source domain but not in the target domain (or the other way round).

4 Experiments

In this section, we present a number of experiments. First, we study the bias and the error of the equalized odds predictor $\widehat{Y}$ as a function of the perturbation level in extensive simulations. Next, we show some experiments on real data. Finally, we consider the repeated loss minimization setting of Hashimoto et al. (2018) and demonstrate that the equalized odds method achieves the same goal as the strategy proposed by Hashimoto et al., even when the protected attribute is highly perturbed.

4.1 Simulations of bias and error

For various choices of the problem parameters $\Pr[Y = y, A = a]$ and $\Pr\left[\widetilde{Y} = 1 \,|\, Y = y, A = a\right]$, we study how the bias and the error of a derived equalized odds predictor $\widehat{Y}$ change as the perturbation probabilities $\Pr\left[A_{\text{sw}} \neq A \,|\, A = a, Y = y\right]$, with which the protected attribute in the training phase is perturbed, increase. For doing so, we solve the linear program (2), where in all probabilities the random variable A is replaced by A_{sw} (see Section A.3 in the appendix for details). We compare the bias and the error of $\widehat{Y}$ to the bias and the error of $\widetilde{Y}$, and we also compare the bias of $\widehat{Y}$ to our theoretical bound provided in (8) in Theorem 1. Let $\gamma_{y,a} := \Pr\left[A_{\text{sw}} \neq A \,|\, A = a, Y = y\right]$,

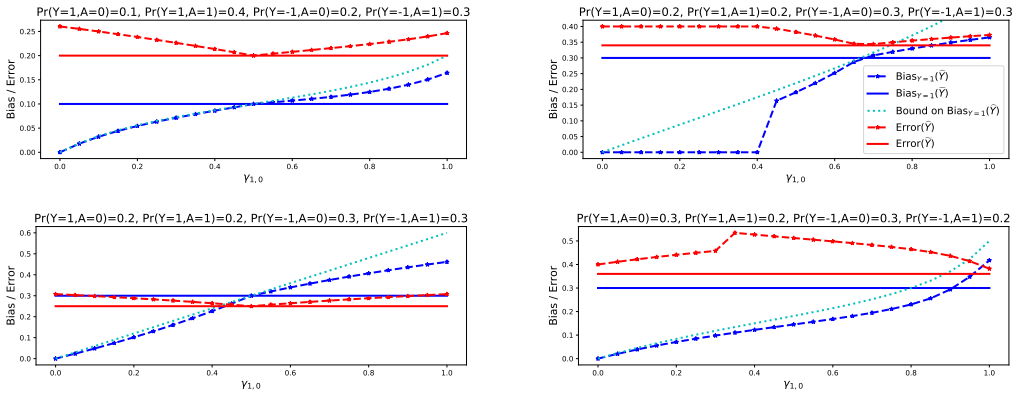

Figure 1: Bias$_{Y=1}(\widehat{Y})$ (dashed blue) and Error$(\widehat{Y})$ (dashed red) as a function of the perturbation level for various choices of the problem parameters (see the titles of the plots and Table 1 in Section A.1 in the appendix). The solid lines show the bias (blue) and the error (red) of the given predictor $\widetilde{Y}$. The dotted cyan curve shows the bound on Bias$_{Y=1}(\widehat{Y})$ provided in (8) in Theorem 1. Note that in the right bottom plot assumption (9) is not satisfied and here the error of $\widehat{Y}$ does not decrease initially.

$y \in \{-1, +1\}, a \in \{0, 1\}$. Figure 1 shows the quantities of interest as a function of $\gamma_{1,0}$, where $\gamma_{1,1}, \gamma_{-1,0}, \gamma_{-1,1}$ grow with $\gamma_{1,0}$ in a certain way, in various scenarios (the probabilities $\Pr[Y = y, A = a]$ can be read from the titles of the plots, and the other parameters are provided in Table 1 in Section A.1 in the appendix). For clarity, we only show the bias for the class $Y = 1$, but note that for a corresponding choice of the parameters the bias for the class $Y = -1$ behaves in the same way. As suggested by our upper bound (8) in Theorem 1, the bias of $\widehat{Y}$ is increasing as the perturbation level increases, and we can see that our upper bound is quite tight in most cases. For a moderate perturbation level with $\gamma_{1,0} + \gamma_{1,1} < 1$, the bias of $\widehat{Y}$ is smaller than the bias of $\widetilde{Y}$ as claimed by Theorem 1. Although all plots show a non-balanced case, which is not captured by Theorem 2, our claim (6) is still shown to be true: except for the bottom right plot, in which assumption (9) is not satisfied, the error of $\widehat{Y}$ decreases as the perturbation level increases up to the point that the error of $\widehat{Y}$ equals the error of $\widetilde{Y}$. The bottom right plot shows that assumption (9) is indeed necessary. We make similar observations in a number of further experiments of this type presented in the appendix in Section A.4. Our findings empirically validate the main claims of our paper.

4.2 Experiments on real data

We run the equalized odds method on two real data sets when we perturb the protected attribute in one of two ways: either we set each protected attribute to its complementary value with probability γ independently of each other, or we (deterministically) flip the protected attribute of every data point whose score lies in the intervall $[0.5 - r, 0.5 + r]$. The score of a data point is the likelihood predicted by a classifier for the data point to belong to the class $Y = 1$ and is related to the given predictor $\widetilde{Y}$ in that $\widetilde{Y}$ predicts $+1$ whenever the score is greater than 0.5 and -1 otherwise. We build upon the data provided by Pleiss et al. (2017). It contains the ground-truth labels, the true protected attributes and the predicted scores for the UCI Adult data set (Dua and Graff, 2019) and the COMPAS criminal recidivism risk assessment data set (Dieterich et al., 2016). The scores for the Adult data set are obtained from a multilayer perceptron, the scores for the COMPAS data set are the actual scores from the COMPAS risk assessment tool. We randomly split the data sets into a training and a test set of equal size (we report several statistics such as the sizes of the original data sets in the appendix in Section A.5). Figure 2 shows the bias and the error of the given predictor $\widetilde{Y}$ and a derived equalized odds predictor $\widehat{Y}$ for the two data sets and in the two perturbation scenarios as a function of the perturbation level γ and r, respectively. The shown curves are obtained from averaging the results of 100 runs of the experiment. They look quite similar to the ones that we obtained in the experiments of Section 4.1 and again validate the main claims of our paper.

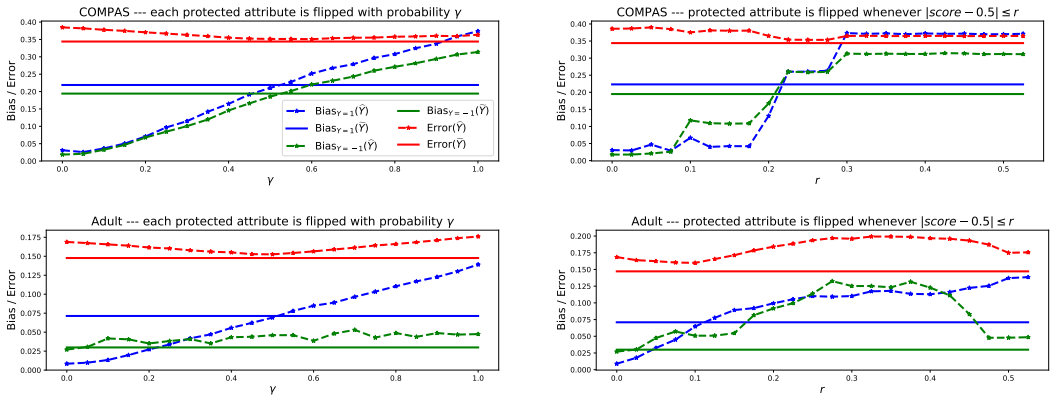

Figure 2: $\text{Bias}_{Y=1}(\widehat{Y})$ (dashed blue), $\text{Bias}_{Y=-1}(\widehat{Y})$ (dashed green) and $\text{Error}(\widehat{Y})$ (dashed red) as a function of the perturbation level for two real data sets and two perturbation scenarios. The solid lines show the bias (blue and green) and the error (red) of the given predictor $\widetilde{Y}$.

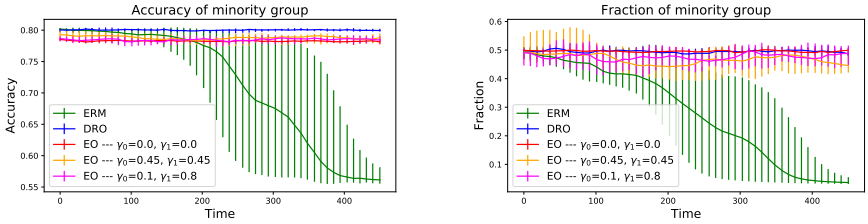

Figure 3: Repeated loss minimization experiment of Hashimoto et al. (2018) (Figure 5 in their paper). Not only the method proposed by Hashimoto et al. (DRO), but also equalized odds postprocessing guarantees high user retention, and hence high accuracy, for both groups over time, even when the protected 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.

4.3 Repeated loss minimization

We compare the equalized odds method to the method of Hashimoto et al. (2018), discussed in Section 3, 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 step t depends on the group's fraction and the classifier's accuracy on the group at time step $t-1$. Hashimoto et al. show that in such a setting standard empirical risk minimization can lead to disparity amplification with a group having a very small fraction, and thus very small classification accuracy, after some time while their proposed method helps to avoid this situation.

In Figure 3 we present an experiment that reproduces and extends the experiment shown in Figure 5 in Hashimoto et al. (2018).[2] Figure 3 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 intially 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 logistic regression classifier trained with empirical risk minimization and DRO to a logistic regression classifier trained with distributionally robust optimization (the method propopsed 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_{sw}, which is obtained by flipping A to its complementary value with probabilities $\gamma_0 := \Pr[A_{\text{sw}} \neq A \,|\, A = 0]$ and $\gamma_1 := \Pr[A_{\text{sw}} \neq A \,|\, A = 1]$, 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 higly perturbed (orange

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

and magenta curves). 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 the algorithm does not come with per step theoretical guarantees. Hence, we believe that in situations where one has access to a perturbed version of the protected attribute, the equalized odds method is a more sensible alternative.

5 Discussion

In this paper, we studied the equalized odds method of Hardt et al. (2016) for fair classification when the protected attribute is perturbed. We gave strong theoretical and empirical evidence that as long as the perturbation is somewhat moderate, one should still run the equalized odds method with the perturbed attribute. In doing so, one still reduces the bias of the original classifier while not suffering too much in terms of accuracy. We believe that without such a property the practical applicability of a "fair" machine learning method is only limited. While there is some empirical work demonstrating the usefulness of using a proxy for the protected attribute when the protected attribute is not available (Gupta et al., 2018; see Section 3), our paper is the first to provide a rigorous theoretical analysis for such a claim. This opens up a new line of research in fairness in ML, asking which methods are robust to a perturbation of the protected attribute and if so, to what extent.

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. http://www.fairmlbook.org.

L. E. Celis, V. Keswani, D. Straszak, A. Deshpande, T. Kathuria, and N. K. Vishnoi. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning (ICML)*, 2018.

J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on Fairness, Accountability, and Transparency (Fat*)*, 2019.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Neural Information Processing Systems (NIPS)*, 2017.

A. Coston, K. N. Ramamurthy, D. Wei, K. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*, 2019.

W. Dieterich, C. Mendoza, and T. Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe Inc., 2016. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems (NeurIPS)*, 2018.

D. Dua and C. Graff. UCI machine learning repository, 2019. https://archive.ics.uci.edu/ml/datasets/adult.

C. Dwork and C. Ilvent. Individual fairness under composition. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

M. Gupta, A. Cotter, M. M. Fard, and S. Wang. Proxy fairness. arXiv:1806.11212 [cs.LG], 2018.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NIPS)*, 2016.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018. Code available on https://bit.ly/2sFkDpE.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2012.

N. Kilbertus, A. Gascón, M. Kusner, M. Veale, K. P. Gummadi, and A. Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning (ICML)*, 2018.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

M. Kleindessner, P. Awasthi, and J. Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning (ICML)*, 2019a.

M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning (ICML)*, 2019b.

A. L. Lamy, Z. Zhong, A. K. Menon, and N. Verma. Noise-tolerant fair classification. *arXiv:1901.10837 [cs.LG]*, 2019.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability, and Transparency*, 2018.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Neural Information Processing Systems (NIPS)*, 2017. Code and data available on https://github.com/gpleiss/equalized_odds_and_calibration.

S. Samadi, U. Tantipongpipat, J. Morgenstern, M. Singh, and S. Vempala. The price of fair PCA: One extra dimension. In *Neural Information Processing Systems (NeurIPS)*, 2018.

M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means clustering. arXiv:1812.10854 [cs.DS], 2018.

S. Verma and J. Rubin. Fairness definitions explained. In *International Workshop on Software Fairness (FairWare)*, 2018.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, 2017.

D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data*, 2018.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017a.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web (WWW)*, 2017b.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, 2013.

A Appendix

A.1 Problem parameters for the experiments of Figure 1

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

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

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

A.2 Proofs

We require a simple technical lemma.

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

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

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) to be true. 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

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

We have

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

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

$$1 - (\gamma_1 + \gamma_2 - 1) \cdot \big\{p \cdot \big[(1-\gamma_1)p + \gamma_2(1-p)\big] - p \cdot \big[\gamma_1 p + (1-\gamma_2)(1-p)\big]\big\} \geq$$
$$1 - |\gamma_1 + \gamma_2 - 1| \cdot \big|p \cdot \big[(1-\gamma_1)p + \gamma_2(1-p)\big] - p \cdot \big[\gamma_1 p + (1-\gamma_2)(1-p)\big]\big| \geq 1 - p > 0.$$

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

$$\begin{aligned} \alpha_1 &:= \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = 1, A = 0\right], \qquad \beta_1 := \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = 1, A = 1\right], \\ \alpha_2 &:= \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = -1, A = 0\right], \qquad \beta_2 := \Pr\left[\widetilde{Y} = 1 \,\middle|\, Y = -1, A = 1\right]. \end{aligned} \tag{10}$$

Then

$$\text{Bias}_{Y=+1}(\widetilde{Y}) = |\alpha_1 - \beta_1|, \quad \text{Bias}_{Y=-1}(\widetilde{Y}) = |\alpha_2 - \beta_2|. \tag{11}$$

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 $\Pr\left[Y = y', A = a, \widetilde{Y} = y\right]$ and $\Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A = a\right]$ by $\Pr\left[Y = y', A_{\text{sw}} = a, \widetilde{Y} = y\right]$ and $\Pr\left[\widetilde{Y} = y' \,\middle|\, Y = y, A_{\text{sw}} = a\right]$, respectively, in the linear program (2). Note that the assumption $\Pr\left[A_{\text{sw}} = A \,|\, A = a, Y = y\right] > 0$ for $y \in \{-1, +1\}$ and $a \in \{0, 1\}$ implies that $\Pr[Y = y, A_{\text{sw}} = a] > 0$ for $y \in \{-1, +1\}$ and $a \in \{0, 1\}$. It is

$$\Pr\left[Y = y', A_{\text{sw}} = a, \widetilde{Y} = y\right] = \Pr\left[\widetilde{Y} = y \,\middle|\, Y = y', A_{\text{sw}} = a,\right] \cdot \Pr\left[Y = y', A_{\text{sw}} = a\right]$$

and, for $a \in \{0, 1\}$,

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

Hence, we end up with the new linear program

$$\min_{\substack{p_{1,0},\, p_{1,1},\\ p_{-1,0},\, p_{-1,1} \in [0,1]}} \sum_{\substack{y \in \{-1,+1\}\\ a \in \{0,1\}}} \left\{\Pr\left[Y = -1, A_{\text{sw}} = a, \widetilde{Y} = y\right] - \Pr\left[Y = 1, A_{\text{sw}} = a, \widetilde{Y} = y\right]\right\} \cdot p_{y,a}$$

$$\begin{aligned} \text{s.t. } & \{\beta_1 \cdot \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 0] + \alpha_1 \cdot (1 - \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 0])\} \cdot p_{1,0} \\ & \quad + \{1 - \beta_1 \cdot \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 0] - \alpha_1 \cdot (1 - \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 0])\} \cdot p_{-1,0} = \\ & \{\beta_1 \cdot \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 1] + \alpha_1 \cdot (1 - \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 1])\} \cdot p_{1,1} \\ & \quad + \{1 - \beta_1 \cdot \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 1] - \alpha_1 \cdot (1 - \Pr[A = 1 \,|\, Y = 1, A_{\text{sw}} = 1])\} \cdot p_{-1,1}, \\ & \{\beta_2 \cdot \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 0] + \alpha_2 \cdot (1 - \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 0])\} \cdot p_{1,0} \\ & \quad + \{1 - \beta_2 \cdot \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 0] - \alpha_2 \cdot (1 - \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 0])\} \cdot p_{-1,0} = \\ & \{\beta_2 \cdot \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 1] + \alpha_2 \cdot (1 - \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 1])\} \cdot p_{1,1} \\ & \quad + \{1 - \beta_2 \cdot \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 1] - \alpha_2 \cdot (1 - \Pr[A = 1 \,|\, Y = -1, A_{\text{sw}} = 1])\} \cdot p_{-1,1}. \end{aligned} \tag{12}$$

Some elementary calculations yield that the objective function $\Delta = \Delta(p_{1,0}, p_{1,1}, p_{-1,0}, p_{-1,1})$ in (12) equals

$$\begin{aligned}\Delta = {} & \Pr[Y=-1, A_{\text{sw}}=0]\left[(p_{1,0}-p_{-1,0})\cdot\{\alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=0]\}+p_{-1,0}\right]\\ & +\Pr[Y=-1, A_{\text{sw}}=1]\left[(p_{1,1}-p_{-1,1})\cdot\{\alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=1]\}+p_{-1,1}\right]\\ & -\Pr[Y=1, A_{\text{sw}}=0]\left[(p_{1,0}-p_{-1,0})\cdot\{\alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=0]\}+p_{-1,0}\right]\\ & -\Pr[Y=1, A_{\text{sw}}=1]\left[(p_{1,1}-p_{-1,1})\cdot\{\alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=1]\}+p_{-1,1}\right].\end{aligned} \tag{13}$$

and that the constraints are equivalent to

$$\begin{aligned}&(p_{1,0}-p_{-1,0})\cdot\{\alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=0]\}+p_{-1,0}\\ &\qquad=(p_{1,1}-p_{-1,1})\cdot\{\alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=1]\}+p_{-1,1},\\ &(p_{1,0}-p_{-1,0})\cdot\{\alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=0]\}+p_{-1,0}\\ &\qquad=(p_{1,1}-p_{-1,1})\cdot\{\alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=1]\}+p_{-1,1}.\end{aligned} \tag{14}$$

Let

$$e := \alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=0], \tag{15}$$
$$f := \alpha_1+(\beta_1-\alpha_1)\cdot\Pr[A=1\,|\,Y=1,A_{\text{sw}}=1], \tag{16}$$
$$g := \alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=0], \tag{17}$$
$$h := \alpha_2+(\beta_2-\alpha_2)\cdot\Pr[A=1\,|\,Y=-1,A_{\text{sw}}=1]. \tag{18}$$

Then the constraints are

$$\begin{aligned}(p_{1,0}-p_{-1,0})\cdot e+p_{-1,0}&=(p_{1,1}-p_{-1,1})\cdot f+p_{-1,1},\\ (p_{1,0}-p_{-1,0})\cdot g+p_{-1,0}&=(p_{1,1}-p_{-1,1})\cdot h+p_{-1,1}.\end{aligned} \tag{19}$$

Because of the constraints we have

$$\begin{aligned}\Delta &= p_{-1,0}\cdot\{\Pr[Y=-1]-\Pr[Y=1]\}+(p_{1,0}-p_{-1,0})\cdot u\\ &= p_{-1,1}\cdot\{\Pr[Y=-1]-\Pr[Y=1]\}+(p_{1,1}-p_{-1,1})\cdot v,\end{aligned} \tag{20}$$

where

$$\begin{aligned}u &:= g\cdot\Pr[Y=-1]-e\cdot\Pr[Y=1],\\ v &:= h\cdot\Pr[Y=-1]-f\cdot\Pr[Y=1].\end{aligned} \tag{21}$$

It is straightforward to verify that condition (7) for $a=0$ is equivalent to $u\neq 0$ and for $a=1$ it is equivalent to $v\neq 0$. If $u=0$ or $v=0$, one optimal solution to (12) 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 $\Pr[Y=-1]>\Pr[Y=1]$. These probabilities correspond to 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 the proof for the degenerate case is complete.

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

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

we have

$$\begin{aligned}\text{Bias}_{Y=+1}&=\left|\alpha_1\cdot(p_{1,0}-p_{-1,0})-\beta_1\cdot(p_{1,1}-p_{-1,1})+p_{-1,0}-p_{-1,1}\right|,\\ \text{Bias}_{Y=-1}&=\left|\alpha_2\cdot(p_{1,0}-p_{-1,0})-\beta_2\cdot(p_{1,1}-p_{-1,1})+p_{-1,0}-p_{-1,1}\right|.\end{aligned} \tag{22}$$

It is

$$\begin{aligned}\text{Bias}_{Y=+1} &\overset{(20)}{=} \left|\frac{\Delta\alpha_1}{u} - \frac{\Delta\beta_1}{v} + p_{-1,0}\left(1-\frac{\theta\alpha_1}{u}\right) - p_{-1,1}\left(1-\frac{\theta\beta_1}{v}\right)\right| \\ &= \left|\frac{\Delta\alpha_1}{u} - \frac{\Delta\beta_1}{v} + p_{-1,0}\left(1-\frac{\theta e}{u}\right) - p_{-1,1}\left(1-\frac{\theta f}{v}\right) + p_{-1,0}\frac{\theta(e-\alpha_1)}{u} - p_{-1,1}\frac{\theta(f-\beta_1)}{v}\right|.\end{aligned}$$

From (19) and (20) 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

$$\begin{aligned}\text{Bias}_{Y=+1} &= \left|\left(\frac{\Delta}{u} - \frac{p_{-1,0}\theta}{u}\right)(\alpha_1 - e) - \left(\frac{\Delta}{v} - \frac{p_{-1,1}\theta}{v}\right)(\beta_1 - f)\right| \\ &\overset{(15)\&(16)}{=} |\alpha_1 - \beta_1| \cdot \left|\left(\frac{\Delta}{u} - \frac{p_{-1,0}\theta}{u}\right) \cdot \Pr[A=1 \,|\, Y=1, A_{\text{sw}}=0] + \right. \\ &\qquad \left. \left(\frac{\Delta}{v} - \frac{p_{-1,1}\theta}{v}\right) \cdot \Pr[A=0 \,|\, Y=1, A_{\text{sw}}=1]\right| \qquad (23) \\ &\overset{(20)}{=} |\alpha_1 - \beta_1| \cdot |(p_{1,0} - p_{-1,0}) \cdot \Pr[A=1 \,|\, Y=1, A_{\text{sw}}=0] + \\ &\qquad (p_{1,1} - p_{-1,1}) \cdot \Pr[A=0 \,|\, Y=1, A_{\text{sw}}=1]| \\ &\leq |\alpha_1 - \beta_1| \cdot \{\Pr[A=1 \,|\, Y=1, A_{\text{sw}}=0] + \Pr[A=0 \,|\, Y=1, A_{\text{sw}}=1]\},\end{aligned}$$

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

$$\text{Bias}_{Y=-1} \leq |\alpha_2 - \beta_2| \cdot \{\Pr[A=1 \,|\, Y=-1, A_{\text{sw}}=0] + \Pr[A=0 \,|\, Y=-1, A_{\text{sw}}=1]\}. \qquad (24)$$

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

$$\begin{aligned}\Pr[A=1 \,|\, Y=y, A_{\text{sw}}=0] &= \frac{\Pr[A=1, A_{\text{sw}}=0 \,|\, Y=y]}{\Pr[A_{\text{sw}}=0 \,|\, Y=y]} \\ &= \frac{\Pr[A_{\text{sw}}=0 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y]}{\Pr[A_{\text{sw}}=0 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y] + \Pr[A_{\text{sw}}=0 \,|\, Y=y, A=0] \cdot \Pr[A=0 \,|\, Y=y]} \qquad (25)\end{aligned}$$

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

$$\begin{aligned}\Pr[A=1 \,|\, Y=y, A_{\text{sw}}=1] &= \frac{\Pr[A=1, A_{\text{sw}}=1 \,|\, Y=y]}{\Pr[A_{\text{sw}}=1 \,|\, Y=y]} \\ &= \frac{\Pr[A_{\text{sw}}=1 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y]}{\Pr[A_{\text{sw}}=1 \,|\, Y=y, A=1] \cdot \Pr[A=1 \,|\, Y=y] + \Pr[A_{\text{sw}}=1 \,|\, Y=y, A=0] \cdot \Pr[A=0 \,|\, Y=y]}. \qquad (26)\end{aligned}$$

The statement of Theorem 1 for the non-degenerate case follows from combining (11), (23), (24), (25), (26) and Lemma 1. □

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 (10). Since we assume that $\Pr[Y=y, A=a] = \frac{1}{4}$, $y \in \{-1,+1\}$, $a \in \{0,1\}$, and that $\widetilde{Y}$ is a weak learner for both of the groups $A=0$ and $A=1$, 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 $\gamma := \Pr[A_{\text{sw}} \neq A \,|\, A=a, Y=y]$, which does not depend on the values of a and y, be the perturbation probability. In the training phase for $\widehat{Y}_{\text{corr}}$ we have $\gamma = \gamma_0$ for some $\gamma_0 \in (0, \frac{1}{2}]$, and

in the training phase for $\widehat{Y}_{\text{true}}$ we have $\gamma = 0$. Since we are assuming a balanced case, we have $\Pr[Y = +1] = \Pr[Y = -1] = \frac{1}{2}$.

It follows from (15) to (21), (25) and (26) that for any value of the perturbation probability $\gamma \in [0, 1]$ the equalized odds method solves the following linear program:

$$\begin{aligned} \min_{p_{1,0},p_{1,1},p_{-1,0},p_{-1,1}\in[0,1]} \quad & \Delta \\ \text{s.t.} \quad & (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_1 + \gamma\beta_1\} + p_{-1,0} = (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_1 + \gamma\alpha_1\} + p_{-1,1}, \\ & (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_2 + \gamma\beta_2\} + p_{-1,0} = (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_2 + \gamma\alpha_2\} + p_{-1,1}, \end{aligned} \tag{27}$$

where

$$\Delta = (p_{1,0} - p_{-1,0})u = (p_{1,1} - p_{-1,1})v \tag{28}$$

with

$$\begin{aligned} u &= \frac{1}{2}\left[(1-\gamma)(\alpha_2 - \alpha_1) + \gamma(\beta_2 - \beta_1)\right], \\ v &= \frac{1}{2}\left[(1-\gamma)(\beta_2 - \beta_1) + \gamma(\alpha_2 - \alpha_1)\right]. \end{aligned} \tag{29}$$

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 (27) and has objective value $\Delta = 0$, in an equalized odds solution (i.e., an optimal solution to (27)) 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 (27) that

$$\begin{aligned} p_{-1,0} - p_{-1,1} &= (p_{1,1} - p_{-1,1}) \cdot \{(1-\gamma)\beta_1 + \gamma\alpha_1\} - (p_{1,0} - p_{-1,0}) \cdot \{(1-\gamma)\alpha_1 + \gamma\beta_1\} \\ &\overset{(28)}{=} \frac{\Delta}{v}((1-\gamma)\beta_1 + \gamma\alpha_1) - \frac{\Delta}{u}((1-\gamma)\alpha_1 + \gamma\beta_1) \\ &= \frac{\Delta}{uv}\Big(\beta_1((1-\gamma)u - \gamma v) - \alpha_1((1-\gamma)v - \gamma u)\Big) \\ &\overset{(29)}{=} \frac{\Delta(1-2\gamma)}{2uv}(\alpha_2\beta_1 - \alpha_1\beta_2) \\ &\leq 0, \end{aligned} \tag{30}$$

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

$$\text{Error}(\widehat{Y}) = \frac{1}{4} \cdot \{(p_{1,0} - p_{-1,0})(\alpha_2 - \alpha_1) + (p_{1,1} - p_{-1,1})(\beta_2 - \beta_1)\} + \frac{1}{2} \tag{31}$$

and hence is invariant under translations of the probabilities (compare with the end of Section 2.1). Hence, without loss of generality, we may assume that $p_{-1,0} = 0$. Substituting in the expressions computed above we get that

$$p_{1,0} \overset{(28)}{=} \frac{\Delta}{u}, \tag{32}$$

$$p_{-1,1} \overset{(30)}{=} \frac{\Delta(1-2\gamma)}{2uv}(\alpha_1\beta_2 - \alpha_2\beta_1), \tag{33}$$

$$p_{1,1} \overset{(28)}{=} \frac{\Delta}{v} + p_{-1,1} = \Delta\left[\frac{1}{v} + \frac{(1-2\gamma)(\alpha_1\beta_2 - \alpha_2\beta_1)}{2uv}\right]. \tag{34}$$

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

$$\begin{aligned} p_{1,0} - p_{1,1} &\overset{(32)\&(34)}{=} \Delta\Big(\frac{1}{u} - \frac{1}{v} - \frac{(1-2\gamma)(\alpha_1\beta_2 - \alpha_2\beta_1)}{2uv}\Big) \\ &\overset{(29)}{=} \frac{\Delta(1-2\gamma)}{2uv}\Big(\beta_2 - \beta_1 + \alpha_1 - \alpha_2 + \alpha_2\beta_1 - \alpha_1\beta_2\Big). \end{aligned} \tag{35}$$

Importantly, the difference (35) 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 (35) 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 (31) 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 (27) 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

$$p^{\gamma_0}_{1,1} - p^{\gamma_0}_{-1,1} \overset{(28)}{=} \frac{u(\gamma_0)}{v(\gamma_0)}$$

and hence

$$\text{Error}(\widehat{Y}_{\text{true}}) - \text{Error}(\widehat{Y}_{\text{corr}}) = \frac{1}{4} \cdot \{(\alpha_2 - \alpha_1) - \frac{u(\gamma_0)}{v(\gamma_0)}(\beta_2 - \beta_1)\} \overset{(29)}{=} \frac{\gamma_0}{4} \frac{(\alpha_2 - \alpha_1)^2 - (\beta_2 - \beta_1)^2}{2v(\gamma_0)}.$$

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

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

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 (35) 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 $\widehat{Y}_{\text{true}}$) 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 $p^0_{1,1} = p^{\gamma_0}_{1,1} = 1$ and $p^0_{-1,0} = p^{\gamma_0}_{-1,0} = 0$. Similarly as in Case 1 we obtain that

$$\begin{aligned}\text{Error}(\widehat{Y}_{\text{true}}) - \text{Error}(\widehat{Y}_{\text{corr}}) = \frac{1}{4}\Big\{&2(1 - p^0_{-1,1})(\beta_2 - \beta_1) - (1 - p^{\gamma_0}_{-1,1})(\beta_2 - \beta_1) \\ &- \frac{v(\gamma_0)}{u(\gamma_0)}(1 - p^{\gamma_0}_{-1,1})(\alpha_2 - \alpha_1)\Big\}.\end{aligned} \tag{36}$$

When $p_{1,1} = 1$, we obtain from (34) that

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

This implies that

$$1 - p^{\gamma_0}_{-1,1} \overset{(33)}{=} \frac{2u(\gamma_0)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)} \tag{37}$$

and

$$1 - p^0_{-1,1} \overset{(37)\&(29)}{=} \frac{\alpha_2 - \alpha_1}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1}.$$

Substituting these in (36) we get that

$$\begin{aligned}\text{Error}(\widehat{Y}_{\text{true}}) - \text{Error}(\widehat{Y}_{\text{corr}}) &= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} - \frac{(\beta_2 - \beta_1)2u(\gamma_0) + (\alpha_2 - \alpha_1)2v(\gamma_0)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)}\right\} \\ &= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} - \frac{\gamma_0(\beta_2 - \beta_1 - \alpha_2 + \alpha_1)^2 + 2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_1\beta_2 - \alpha_2\beta_1)}\right\} \\ &= \frac{1}{4}\left\{2\frac{(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{\alpha_2 - \alpha_1 + \alpha_1\beta_2 - \alpha_2\beta_1} + \frac{\gamma_0(\beta_2 - \beta_1 - \alpha_2 + \alpha_1)^2 + 2(\beta_2 - \beta_1)(\alpha_2 - \alpha_1)}{-2u(\gamma_0) + (1 - 2\gamma_0)(\alpha_2\beta_1 - \alpha_1\beta_2)}\right\}.\end{aligned} \tag{38}$$

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 (38) that

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

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 $\text{Bias}_{Y=+1}(\widetilde{Y})=\text{Bias}_{Y=-1}(\widetilde{Y})=0$ (compare with (11)). Hence, we have proved Theorem 2. □

A.3 Detailed expressions required for the experiments of Section 4.1

We need to solve the linear program

$$\begin{aligned}\min_{\substack{p_{1,0},\,p_{1,1},\\ p_{-1,0},\,p_{-1,1}\in[0,1]}} \quad & \sum_{\substack{y\in\{-1,+1\}\\ a\in\{0,1\}}} \left\{\Pr\left[Y=-1, A_{\text{sw}}=a, \widetilde{Y}=y\right]-\Pr\left[Y=1, A_{\text{sw}}=a, \widetilde{Y}=y\right]\right\}\cdot p_{y,a}\\ \text{s.t. } & \Pr\left[\widetilde{Y}=1\,\middle|\,Y=y, A_{\text{sw}}=0\right]\cdot p_{1,0}+\Pr\left[\widetilde{Y}=-1\,\middle|\,Y=y, A_{\text{sw}}=0\right]\cdot p_{-1,0} =\\ & \Pr\left[\widetilde{Y}=1\,\middle|\,Y=y, A_{\text{sw}}=1\right]\cdot p_{1,1}+\Pr\left[\widetilde{Y}=-1\,\middle|\,Y=y, A_{\text{sw}}=1\right]\cdot p_{-1,1}, \quad y\in\{-1,1\},\end{aligned} \tag{39}$$

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 $\Pr\left[A_{\text{sw}}\neq A\,|\,A=a, Y=y\right]$. As in Section 4.1, we let $\gamma_{y,a}:=\Pr\left[A_{\text{sw}}\neq A\,|\,A=a, Y=y\right]$, $y\in\{-1,+1\}, a\in\{0,1\}$. From (13) to (18) in the proof of Theorem 1 we obtain that the objective function equals

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

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

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

and $\alpha_1, \beta_1, \alpha_2, \beta_2$ defined in (10). It is

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

and from (25) and (26) in the proof of Theorem 1 we obtain that

$$\Pr[A=1 \,|\, Y=y, A_{\text{sw}}=0] = \frac{\gamma_{y,1} \cdot \Pr[A=1, Y=y]}{\gamma_{y,1} \cdot \Pr[A=1, Y=y] + (1-\gamma_{y,0}) \cdot \Pr[A=0, Y=y]},$$
$$\Pr[A=1 \,|\, Y=y, A_{\text{sw}}=1] = \frac{(1-\gamma_{y,1}) \cdot \Pr[A=1, Y=y]}{(1-\gamma_{y,1}) \cdot \Pr[A=1, Y=y] + \gamma_{y,0} \cdot \Pr[A=0, Y=y]}.$$

Hence, we have written all coefficients of (39) in terms of the problem parameters and perturbation probabilities.

After solving (39) 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 (22) in the proof of Theorem 1 we obtain that

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

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)

$$\begin{aligned}
\text{Error}(\widehat{Y}) = \Pr[Y=1] &+ \left\{\alpha_2 \Pr[Y=-1, A=0] - \alpha_1 \Pr[Y=1, A=0]\right\} \cdot p_{1,0} \\
&+ \left\{\beta_2 \Pr[Y=-1, A=1] - \beta_1 \Pr[Y=1, A=1]\right\} \cdot p_{1,1} \\
&+ \left\{\Pr[Y=-1, A=0] - \Pr[Y=1, A=0] - \alpha_2 \Pr[Y=-1, A=0] + \alpha_1 \Pr[Y=1, A=0]\right\} \cdot p_{-1,0} \\
&+ \left\{\Pr[Y=-1, A=1] - \Pr[Y=1, A=1] - \beta_2 \Pr[Y=-1, A=1] + \beta_1 \Pr[Y=1, A=1]\right\} \cdot p_{-1,1}.
\end{aligned} \tag{40}$$

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 (40))

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

A.4 Further experiments as in Section 4.1

In Figure 4, we present a number of further experiments as described in Section 4.1 of the main paper. The problem parameters can be read from the titles of the plots and Table 2. We make the same observations as for the experiments of Section 4.1 and hence obtain further validation of the main claims of our paper.

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

Plot	$\Pr[\widetilde{Y} = 1 \mid Y = y, A = a]$				$(\gamma_{1,1}, \gamma_{-1,0}, \gamma_{-1,1})$
	$y = 1$ $a = 0$	$y = 1$ $a = 1$	$y = -1$ $a = 0$	$y = -1$ $a = 1$	
1st row left	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
1st row right	0.8	0.9	0.1	0.0	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$
2nd row left	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, \frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{2})$
2nd row right	0.8	0.9	0.1	0.0	$(\gamma_{1,0}, r(\gamma_{1,0}), r(\gamma_{1,0}))$
3rd row left	0.9	0.6	0.7	0.1	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
3rd row right	0.9	0.4	0.1	0.1	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$
4th row left	0.7	0.9	0.3	0.0	$(\gamma_{1,0}, \frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{2})$
4th row right	0.7	0.9	0.3	0.0	$(\gamma_{1,0}, r(\gamma_{1,0}), r(\gamma_{1,0}))$
5th row left	0.5	0.8	0.1	0.4	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
5th row right	1.0	0.8	0.0	0.1	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
6th row left	0.3	0.8	0.1	0.2	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
6th row right	0.3	0.8	0.1	0.2	$(\gamma_{1,0}, \gamma_{1,0}, \gamma_{1,0})$
7th row left	0.9	0.6	0.4	0.1	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$
7th row right	0.9	0.6	0.4	0.4	$(\frac{\gamma_{1,0}}{2}, \frac{\gamma_{1,0}}{4}, \frac{\gamma_{1,0}}{8})$

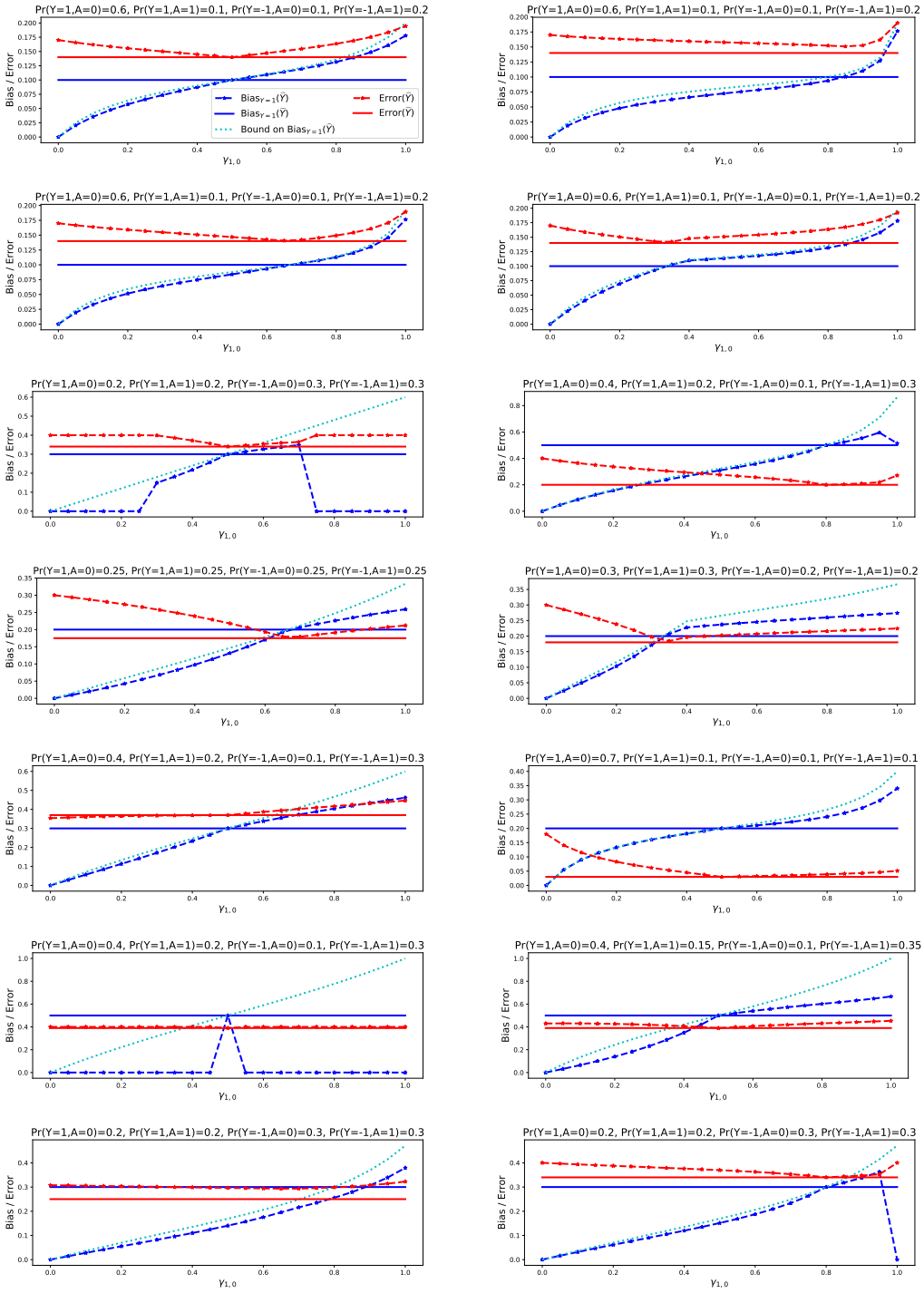

Figure 4: Similar experiments as shown in Figure 1 in the main paper. The dashed blue curve shows $\text{Bias}_{Y=1}(\widehat{Y})$ and the dashed red curve shows $\text{Error}(\widehat{Y})$ as a function of the perturbation level. The solid red line shows $\text{Bias}_{Y=1}(\widetilde{Y})$ and the solid blue line shows $\text{Error}(\widetilde{Y})$. The dotted cyan curve shows the bound on $\text{Bias}_{Y=1}(\widehat{Y})$ provided in (8) in Theorem 1. The problem parameters can be read from the titles of the plots and Table 2.

A.5 Some statistics of the real data sets used in Section 4.2

Table 3 provides several statistics of the real data sets used in Section 4.2, before splitting them into a training and a test set.

Table 3: Statistics of the real data sets used in Section 4.2.

	Adult	COMPAS
$\#$ records	9768	6150
$\frac{\#\,(Y=1 \wedge A=0)}{\#\text{ records}}$	0.470	0.157
$\frac{\#\,(Y=1 \wedge A=1)}{\#\text{ records}}$	0.294	0.309
$\frac{\#\,(Y=-1 \wedge A=0)}{\#\text{ records}}$	0.201	0.242
$\frac{\#\,(Y=-1 \wedge A=1)}{\#\text{ records}}$	0.036	0.292
$\frac{\#\,(\widetilde{Y}=1)}{\#\text{ records}}$	0.795	0.394
$\frac{\#\,(\widetilde{Y}=-1)}{\#\text{ records}}$	0.205	0.606
$\frac{\#\,(\widetilde{Y}\neq Y)}{\#\text{ records}}$	0.147	0.344
$\frac{\#\,(\widetilde{Y}=1 \wedge Y=1 \wedge A=0)}{\#\,(Y=1 \wedge A=0)}$	0.897	0.408
$\frac{\#\,(\widetilde{Y}=1 \wedge Y=1 \wedge A=1)}{\#\,(Y=1 \wedge A=1)}$	0.968	0.628
$\frac{\#\,(\widetilde{Y}=1 \wedge Y=-1 \wedge A=0)}{\#\,(Y=-1 \wedge A=0)}$	0.374	0.147
$\frac{\#\,(\widetilde{Y}=1 \wedge Y=-1 \wedge A=1)}{\#\,(Y=-1 \wedge A=1)}$	0.398	0.343