Equalized odds postprocessing under imperfect group information

Pranjal Awasthi
Rutgers University & Google

Matthäus Kleindessner
University of Washington

Jamie Morgenstern
University of Washington & Google

Abstract

Most approaches aiming to ensure a model's fairness with respect to a protected attribute (such as gender or race) assume to know the true value of the attribute for every data point. In this paper, we ask to what extent fairness interventions can be effective even when only imperfect information about the protected attribute is available. In particular, we study the prominent equalized odds postprocessing method of Hardt et al. (2016) under a perturbation of the attribute. We identify conditions on the perturbation that guarantee that the bias of a classifier is reduced even by running equalized odds with the perturbed attribute. We also study the error of the resulting classifier. We empirically observe that under our identified conditions most often the error does not suffer from a perturbation of the protected attribute. For a special case, we formally prove this observation to be true.

1 INTRODUCTION

As machine learning (ML) algorithms become more and more embedded into our society, evidence has surfaced questioning whether they produce equally high-quality predictions for most members of diverse populations. The work on fairness in ML aims to understand the extent to which existing ML methods produce *fair* predictions for different individuals, and what new methods can remove the discrepancies therein (Barocas et al., 2018). The appropriate formalization of "fair" 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 tries to ensure some property for every individual (and usually is agnostic to any group membership), while the latter asks some statistic (e.g., accuracy or false positive rate) to be similar for different groups. One key drawback of individual fairness is the need for a task-specific similarity metric over the space of individuals. Group fairness, on the other hand, usually requires knowledge of group membership (such as gender or race), encoded by a *protected attribute*. While arguably a more reasonable requirement than asking for a task-specific 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 imperfect information about the protected attribute, hence generalizing the applicability of such methods.

More specifically, we explore the question of when perturbed protected attribute information can be substituted for the true attribute in the training phase of an existing algorithmic framework for fair classification 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 study this question in the context of the prominent postprocessing method of Hardt et al. (2016) for ensuring equalized odds (EO).

Our main contribution is to identify (fairly natural) conditions on the perturbation of the protected attribute in the training data for the EO method that guarantee that the resulting classifier $\widehat{Y}$ is still more fair than the original classifier $\widetilde{Y}$ that it is based on. To illustrate the application of our general result, consider a balanced case, where the probability of a data point having label $y \in \{-1, +1\}$ and protected attribute $a \in \{0, 1\}$ equals $1/4$ independent of the values of y and a, and assume that in the training phase every attribute is independently flipped to its complementary value with probability γ. Our result implies that for

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

$\gamma < 0.5$, the bias (as defined in Section 3.2) of $\widehat{Y}$ will be strictly smaller than the bias of $\widetilde{Y}$. While a similar phenomenon was empirically observed in the recent work of Gupta et al. (2018) (see Section 4 for related work), our work is among the first to provide a formal guarantee on the effectiveness of a prominent method for fairness in ML under a perturbation of the attribute. To complement our result, we show that our identified conditions are necessary for providing such a guarantee.

We also study the error of the classifier $\widehat{Y}$. We observe that under our identified conditions, most often the error of $\widehat{Y}$ is not larger than the error of the classifier that we would obtain from running the EO method with the true protected attribute (if it is larger, the difference tends to be negligible—as long as the perturbation is moderate). In the balanced case outlined above, we formally prove this observation to be true.

2 EQUALIZED ODDS

We begin by reviewing the equalized odds (EO) post-processing method of Hardt et al. (2016), assuming the true protected attribute for every data point is known. Like Hardt et al. and as 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.

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., 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 her GPA, Y would encode whether the applicant is a good fit for the job, and A could encode the applicant's gender. 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 4), of 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 EO criterion if, for $y \in \{-1, +1\}$,

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

Throughout the paper we assume $\Pr[Y = y, A = a] > 0$ for $y \in \{-1, +1\}$ and $a \in \{0, 1\}$. For $y = +1$, Equation (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. propose a simple postprocessing method to derive a predictor $\widehat{Y}$ that satisfies the EO 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 $p_{y,a}$ (hence, $\widehat{Y}$ 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 EO criterion, and (ii) the error of $\widehat{Y}$, that is 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:

$$\min_{\substack{p_{-1,0},\ p_{-1,1},\\ p_{1,0},\ p_{1,1}\in[0,1]}} \sum_{\substack{\tilde{y}\in\{-1,+1\}\\ a\in\{0,1\}}} \{G(-1, a, \tilde{y}) - G(1, a, \tilde{y})\} \cdot p_{\tilde{y},a}$$
$$\text{s.t. } H(y,0)\cdot p_{1,0} + \{1 - H(y,0)\}\cdot p_{-1,0} = H(y,1)\cdot p_{1,1} + \{1 - H(y,1)\}\cdot p_{-1,1}, \quad y \in \{-1, 1\}, \tag{2}$$

where $G(y, a, \tilde{y}) = \Pr\left[Y = y, A = a, \widetilde{Y} = \tilde{y}\right]$ and $H(y, a) = \Pr\left[\widetilde{Y} = 1 \middle| Y = y, A = a\right]$. Note that this linear program is not guaranteed to have a unique solution: for example, in case of $\Pr[Y = 1] = \Pr[Y = -1] = 1/2$, it is not hard 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. Hence, the derived predictor $\widehat{Y}$ might not be uniquely defined. All our results apply to *any* derived EO predictor (derived via an arbitrary optimal solution to (2)), with one limitation: whenever the constant classifier $\widehat{Y} = +1$ or $\widehat{Y} = -1$ is an optimal EO predictor (corresponding to optimal probabilities $p_{-1,0} = p_{-1,1} = p_{1,0} = p_{1,1} = 1$ or $p_{-1,0} = p_{-1,1} = p_{1,0} = p_{1,1} = 0$), we assume the derived EO predictor to be this constant classifier. Throughout the paper, we use the terms predictor and classifier interchangeably.

3 ANALYSIS UNDER PERTURBATION OF THE ATTRIBUTE

We first describe our noise model for perturbing the protected attribute. We then study the bias and the error of the derived EO predictor under this noise model.

3.1 Noise Model

When deriving the equalized odds predictor $\widehat{Y}$ from a given classifier $\widetilde{Y}$, one needs to estimate the probabilities $\Pr\left[Y=y, A=a, \widetilde{Y}=\tilde{y}\right]$ and $\Pr\left[\widetilde{Y}=1 \middle| Y=y, A=a\right]$ that appear in (2) from training data and then solve the resulting linear program (2) for some optimal probabilities $p_{-1,0}, p_{-1,1}, p_{1,0}, p_{1,1}$. We refer to this as the *training phase* in the EO procedure. When applying the derived classifier $\widehat{Y}$ in order to predict the label of a test point, which we call the *test phase* of the EO procedure, one tosses a biased coin and outputs 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.

Our noise model captures the scenario that the protected attribute in the training data has been corrupted. Concretely, we assume that in the training phase the two probabilities mentioned above are 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. The random variable A_c denotes the perturbed, or corrupted, attribute. In the test phase we assume that we have access to the true attribute A without any corruption. Hence, the probabilities $p_{y,a}$ of the derived EO predictor for predicting +1 depend upon the perturbed attribute, but the predictions themselves depend on the true 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 attribute) and then applied to test data for which the attribute can be accessed directly or easily verified (as it usually is the case in hiring, for example). We discuss alternative settings and directions for future work in Section 6.

3.2 Bias of the Derived Equalized Odds Predictor under Perturbation

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

$$\text{Bias}_{Y=y}(\widehat{Y}) = \Big| \Pr\left[\widehat{Y}=1 \middle| Y=y, A=0\right] - \Pr\left[\widehat{Y}=1 \middle| Y=y, A=1\right] \Big|.$$

Similarly, we define $\text{Bias}_{Y=y}(\widetilde{Y})$. Note that $\text{Bias}_{Y=y}(\widehat{Y})$ refers to the bias of $\widehat{Y}$ in the test phase, and recall from Section 3.1 that in the test phase, according to our noise model, the derived EO 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 the derived EO predictor (derived from $\widetilde{Y}$) when the protected attribute in the EO training phase has been corrupted, that is $\widehat{Y}_{\text{corr}}$ is based on the linear program (2) with A replaced by A_c (the need for notation $\widehat{Y}_{\text{corr}}$ instead of $\widehat{Y}$ comes from Section 3.3, where we compare $\widehat{Y}_{\text{corr}}$ to the derived EO predictor $\widehat{Y}_{\text{true}}$ that is based on the true attribute; we provide a table collecting all random variables used in the paper in Table 2 in Appendix A.1). The main contribution of our paper is to establish that the following assumptions

Assumptions I

(a) given the ground-truth label Y and the true attribute A, the prediction $\widetilde{Y}$ and the corrupted attribute A_c are conditionally independent

(b) $\sum_{a\in\{0,1\}} \Pr\left[A_c \neq A \middle| Y=y, A=a\right] \leq 1$ and both summands strictly smaller than 1, $y \in \{-1,+1\}$

guarantee that

$$\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) \leq \text{Bias}_{Y=y}(\widetilde{Y}), \quad y \in \{-1,+1\}, \quad (3)$$

and that Assumptions I are necessary for guaranteeing (3). Furthermore, under Assumptions I (a), a strict inequality holds in (3) whenever $\text{Bias}_{Y=y}(\widetilde{Y}) > 0$ and a strict inequality holds in Assumptions I (b). Less surprising, $\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}})$ tends to zero as, for $a \in \{0,1\}$, $\Pr\left[A_c \neq A \middle| Y=y, A=a\right]$ tends to zero. These claims follow from Theorem 1 and Lemma 1 below. Note that the goal of our paper is to *analyze* the equalized odds method as it is and we do not try to *modify* the method.

Before stating Theorem 1 and Lemma 1, let us discuss their implications. To the practitioner who wants to run the EO method, but cannot rule out that the protected attribute might have been corrupted, it is higly relevant to know whether she can still expect to benefit from running EO or whether there is actually a risk of doing harm. According to Theorem 1, if she believes Assumptions I to be true, then she is guaranteed that (3) holds and that by running EO, at the very least, she does not increase the unfairness of the given classifier $\widetilde{Y}$. On the other hand, according to Lemma 1, if the practitioner expects Assumptions I to be violated, she should refrain from running EO as this might yield a predictor with higher bias than the given classifier $\widetilde{Y}$.

Assumptions I are fairly natural and might be satisfied in several practical situations. Assumptions I (a) asks for conditional independence (given Y and A) of the given classifier $\widetilde{Y}$ and the corrupted attribute A_c. For example, this is the case if A_c is the output of a classifier that only uses features that are conditionally independent of the features used by $\widetilde{Y}$ (e.g., body height is used for predicting gender and $\widetilde{Y}$ uses GPA for predicting aptitude for a job[1]). As another example,

[1] Another example, involving race as attribute, might be

Assumptions I (a) is also true if A_c is a crowdsourced estimate of A and one assumes that a crowdworker's probability of providing an incorrect estimate depends on the true label of the task (in our case A), but not on the task X itself, which is the standard assumption in most of the ML literature on crowdsourcing (cf. Kleindessner and Awasthi, 2018). Assumptions I (b) limits the level of perturbation of the protected attribute, but in a rather moderate way. For example, in case of $\Pr[A = a|Y = y] = 1/2$, $y \in \{-1, +1\}$, $a \in \{0, 1\}$, if $\Pr[A_c \neq A|Y = y] < 1/2$, $y \in \{-1, +1\}$, then Assumptions I (b) is satisfied.

Theorem 1 (Bias of $\widehat{Y}_{\text{corr}}$ vs. bias of $\widetilde{Y}$)**.** *Assume that Assumptions I (a) holds and that* $\Pr[A_c \neq A|Y = y, A = a] < 1$ *for* $y \in \{-1, +1\}$ *and* $a \in \{0, 1\}$*. Then, for* $y \in \{-1, +1\}$*, the derived equalized odds predictor* $\widehat{Y}_{\text{corr}}$ *satisfies*

$$\begin{aligned}\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) \leq \text{Bias}_{Y=y}(\widetilde{Y})\cdot \\ F\left(L(y,0), L(y,1), \Pr\left[A = 1|Y = y\right]\right),\end{aligned} \tag{4}$$

where $L(y, a) = \Pr[A_c \neq A|Y = y, A = a]$ *and* $F = F(\gamma_1, \gamma_2, p)$ *is some differentiable function (explicitly stated in (7) in Appendix A.2) that is strictly increasing both in* γ_1 *and in* γ_2 *with* $F(0, 0, p) = 0$ *and* $F(\gamma_1, \gamma_2, p) \underset{(=)}{\leq} 1$ *for all* (γ_1, γ_2, p) *with* $\gamma_1 + \gamma_2 \underset{(=)}{\leq} 1$*.*

Lemma 1 (Assumptions I are necessary for guaranteeing (3))**.** *If any of Assumptions I (a) or (b) is violated, inequality (3) might not be true.*

The proofs of Theorem 1 and Lemma 1 can be found in Appendix A.2, and we provide some intuition behind Theorem 1 in Section 3.4. The main difficulty in proving Theorem 1 comes from the fact that the equalized odds method uses the protected attribute in the test phase. This creates conditional dependencies that make it necessary to characterize how the optimal probabilities $p_{-1,0}, p_{-1,1}, p_{1,0}, p_{1,1}$ to the linear program (2) change under a perturbation of the attribute. Note that the counterexamples that we provide for proving Lemma 1 are not worst-case scenarios in which Assumptions I (a) or (b) would be heavily violated. Indeed, our counterexamples show that a moderate violation of Assumptions I (a) or a minimal violation of Assumptions I (b) can result in (3) not being true. Also note that (4) in Theorem 1 provides a quantitative bound on the bias of $\widehat{Y}_{\text{corr}}$ (in our experiments in Section 5.1 we will see that in most cases this bound is quite tight). In case the practitioner can estimate the various probabilities that it involves, this bound might provide additional benefit to her, but this idea goes beyond the scope of our paper (cf. Section 6).

the following: if one uses a person's surname to predict her race and her income to predict her creditworthiness, then it seems very plausible that conditional independence holds.

3.3 Error of the Derived Equalized Odds Predictor under Perturbation

The error of $\widehat{Y}$ is given by $\text{Error}(\widehat{Y}) = \Pr\left[\widehat{Y} \neq Y\right]$. Note that just as $\text{Bias}_{Y=y}(\widehat{Y})$, $\text{Error}(\widehat{Y})$ refers to the error of $\widehat{Y}$ in the test phase. As in Section 3.2, let $\widehat{Y}_{\text{corr}}$ be the derived EO predictor based on the corrupted protected attribute A_c, and let $\widehat{Y}_{\text{true}}$ be the EO predictor that is based on the true attribute A. In our experiments in Section 5 we observe that under Assumptions I from Section 3.2 and the following additional assumption

Assumption II

- the given predictor $\widetilde{Y}$ is correlated with the ground-truth label Y in the sense that for $a \in \{0, 1\}$

$$\begin{aligned}\Pr\left[\widetilde{Y} = 1 \middle| Y = 1, A = a\right] > \\ \Pr\left[\widetilde{Y} = 1 \middle| Y = -1, A = a\right]\end{aligned}$$

we most often have

$$\text{Error}(\widehat{Y}_{\text{corr}}) \leq \text{Error}(\widehat{Y}_{\text{true}}). \tag{5}$$

In our experiments, if inequality (5) is not true, it tends to be violated only to a negligible extent. In fact, our experiments and some intuition (outlined in Section 3.4) initially misled us to conjecture that (5) would *always* be true under Assumptions I and II (cf. the prior arXiv-version of this paper), but as we show in Section 5.1, such a conjecture is wrong. It only holds in a special balanced case as stated in Theorem 2 below. In general, it remains an open question which assumptions on A_c, $\widetilde{Y}$, and the base rates $\Pr[Y = y, A = a]$ would guarantee inequality (5) to hold (cf. Section 6).

Assumption II is mild and kind of a minimal requirement for $\widetilde{Y}$ to be considered useful. If $\Pr[Y = y|A = a] = 1/2$, $y \in \{-1, +1\}$, $a \in \{0, 1\}$, it is equivalent to requiring $\widetilde{Y}$ to be a weak learner for both groups $A = a$, that is to satisfy $\Pr\left[\widetilde{Y} \neq Y \middle| A = a\right] < 1/2$, $a \in \{0, 1\}$. However, in a special balanced case, together with Assumptions I, Assumption II is sufficient to guarantee that (5) holds as the following theorem states:

Theorem 2 (Error of $\widehat{Y}_{\text{corr}}$ vs. error of $\widehat{Y}_{\text{true}}$ in a special case)**.** *Assume that Assumptions I and II hold. Furthermore, assume that* $\Pr[Y = y, A = a] = 1/4$*,* $y \in \{-1, +1\}$*,* $a \in \{0, 1\}$*, and* $\Pr[A_c \neq A|Y = y, A = a] \in (0, 1/2]$ *does not depend on* y *and* a*. Then we have*

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

with equality holding if and only if the given classifier $\widetilde{Y}$ *is unbiased, that is* $\text{Bias}_{Y=+1}(\widetilde{Y}) = \text{Bias}_{Y=-1}(\widetilde{Y}) = 0$*.*

The proof of Theorem 2 can be found in Appendix A.2. Although several expressions in the analysis of the linear program (2) simplify in the special case considered in Theorem 2, the proof of Theorem 2 is still involved and requires a case analysis that distinguishes which of the probabilities in an optimal solution to (2) equal 1.

3.4 Intuition behind (3) and (5)

To provide some intuition behind our results and observations, consider the simple case of independently flipping each data point's protected attribute to its complementary value with probability γ. If $\gamma = 0$, the EO method gets to see the true attribute and we end up with the classifier $\widehat{Y}_{\text{true}}$, which has zero bias, but usually quite a larger error than the given classifier $\widetilde{Y}$. If $\gamma = 0.5$, the EO method gets to see random noise as the attribute and the given classifier $\widetilde{Y}$ appears to be totally fair. In this case, unless $\widetilde{Y}$ is rather bad and its accuracy can be improved simply by flipping its predictions from $+1$ to -1, or the other way round, for a group $A = a$ (in the balanced case studied in Theorem 2, Assumption II rules out such a situation), the EO method returns the given $\widetilde{Y}$, which has smaller error than $\widehat{Y}_{\text{true}}$, but higher test-phase bias. For $0 < \gamma < 0.5$, the EO method yields a classifier $\widehat{Y}_{\text{corr}}$ that interpolates between these two extremes: some amount of random noise in the attribute makes the given classifier $\widetilde{Y}$ appear more fair than it actually is and the EO method changes $\widetilde{Y}$ (i.e., decreases its bias / increases its error) in a less severe way than when it gets to see the true attribute. This interpolation behavior can be seen nicely in the simulations that we provide in Section 5.1.

While the conclusions in this simple case about the relationship between $\widetilde{Y}$, $\widehat{Y}_{\text{true}}$, and $\widehat{Y}_{\text{corr}}$ with respect to the bias carry over to the more general setting that we consider (we proved (3) to be always true), we do not know where our intuition breaks down with respect to the error in those rare situations in which (5) is not true (cf. Sections 5.1 and 6).

4 RELATED WORK

This section is a short version of a corresponding long version provided in Appendix A.3.

By now, there is a huge body of work on fairness in ML, mainly in supervised learning (e.g., 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; Xu et al., 2018; Kallus and Zhou, 2019), but more recently also in unsupervised learning (e.g., Chierichetti et al., 2017; Samadi et al., 2018; Kleindessner et al., 2019a,b). 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 method of Hardt et al. (2016) as described in Section 2. Concurrently with Hardt et al., the fairness notion of EO has also been proposed by Zafar et al. (2017b) under the name of disparate mistreatment. The seminal paper of Kleinberg et al. (2017) proves that, except for trivial cases, a classifier cannot satisfy the EO criterion and 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 can be suboptimal.

Fairness with Only Limited Information about the Attribute 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 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 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. 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. 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. In Section 5.3 / Appendix A.9, we experimentally compare their approach to the EO method with perturbed attribute information in such a sequential setting. Further works around group fairness with limited information about the attribute are the papers by Botros and Tomczak (2018), Kilbertus et al. (2018), Chen et al. (2019), and Coston et al. (2019).

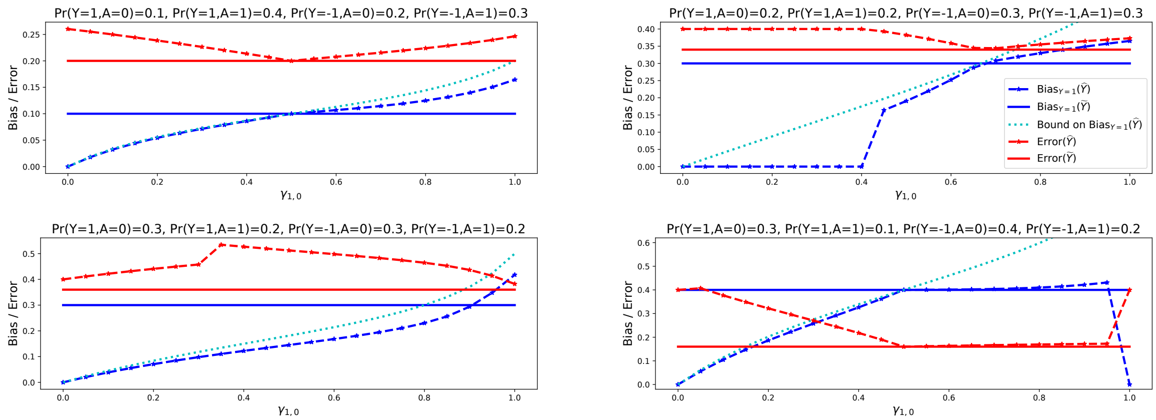

Figure 1: $\mathrm{Bias}_{Y=1}(\widehat{Y})$ (dashed blue) and $\mathrm{Error}(\widehat{Y})$ (dashed red) as a function of the perturbation level for various problem parameters (see the titles of the plots and Table 3 in Appendix A.5). For $\gamma_{1,0} = 0$ it is $\widehat{Y} = \widehat{Y}_{\mathrm{true}}$, and for $\gamma_{1,0} > 0$ it is $\widehat{Y} = \widehat{Y}_{\mathrm{corr}}$. The solid lines show the bias (blue) and the error (red) of $\widetilde{Y}$. The dotted cyan curve shows the bound on $\mathrm{Bias}_{Y=1}(\widehat{Y})$ provided in (4) in Theorem 1. In the left bottom plot, Assumption II is not satisfied and here the error of $\widehat{Y}$ clearly initially increases, that is (5) does not hold. In the right bottom plot, although Assumption II is satisfied, the error of $\widehat{Y}$ also initially increases, but in this case only to a negligible extent.

5 EXPERIMENTS

In this section, we present a number of experiments.[2] First, we study the bias and the error of the EO predictor $\widehat{Y}$ as a function of the level of perturbation of the protected attribute in extensive simulations. In doing so, we empirically validate Theorems 1 and 2 of Section 3 and provide evidence for our claim of Section 3.3 that most often we observe inequality (5) to be true. Next, we show some experiments on real data, providing some motivation for our paper and further support for its main results and claims. Finally, we consider the repeated loss minimization setting of Hashimoto et al. (2018) and demonstrate that the EO method achieves the same goal as their strategy, even when the protected attribute is highly perturbed.

5.1 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 the derived EO predictor $\widehat{Y}$ change as the perturbation probabilities $\Pr[A_{\mathrm{c}} \neq A | Y = y, A = a]$, with which the protected attribute in the EO training phase is perturbed, increase. For doing so, we solve the linear program (2) where in all probabilities A is replaced by A_{c}. We always assume that Assumptions I (a) is satisfied. The resulting linear program is provided in Appendix A.4. 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 (4) in Theorem 1. Let $\gamma_{y,a} := \Pr[A_{\mathrm{c}} \neq A | Y = y, A = a]$, $y \in \{-1, +1\}, a \in \{0, 1\}$. 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 3 in Appendix A.5). In the notation of Section 3, for $\gamma_{1,0} = 0$ it is $\widehat{Y} = \widehat{Y}_{\mathrm{true}}$ and for $\gamma_{1,0} > 0$ it is $\widehat{Y} = \widehat{Y}_{\mathrm{corr}}$. For clarity, we only show the bias for the class $Y = +1$. As suggested by our upper bound (4), in all four plots 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. Note that all four plots show a non-balanced case, which is not captured by Theorem 2. Still, in the two plots in the top row, 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 $\widetilde{Y}$, that is inequality (5) is true. In the bottom left plot, Assumption II is not satisfied and we do not expect inequality (5) to hold here. In the bottom right plot, Assumption II is satisfied, but (5) does not hold either since the error of $\widehat{Y}$ initially increases; however, the violation of (5) is negligible (for $\gamma_{1,0} = 0$, it is $\mathrm{Error}(\widehat{Y}) = 0.4$, and for $\gamma_{1,0} = 0.05$, it is $\mathrm{Error}(\widehat{Y}) = 0.407$). Note that we do not fully understand this behavior (cf. Sections 3.3, 3.4 and 6). We make similar observations in a number of further experiments of this type presented in Appendix A.6, and our findings confirm the main claims of our paper.

[2] Python code available on https://github.com/matthklein/equalized_odds_under_perturbation.

Table 1: Experiment on the Drug Consumption data set. The full table is provided in Appendix A.7.

Y	$\Pr[Y=1]$	$\text{Bias}_{Y=1/-1}(\widetilde{Y})$	$\text{Bias}_{Y=1/-1}(\widehat{Y}_{\text{corr}})$	$\text{Error}(\widetilde{Y})$	$\text{Error}(\widehat{Y}_{\text{corr}})$	$\text{Error}(\widehat{Y}_{\text{true}})$	Co. Ind. (6)
Amphet	0.36	0.085 / 0.106	0.076 / 0.065	0.317	0.339	0.352	0.033
Benzos	0.41	0.074 / 0.132	0.064 / 0.1	0.351	0.369	0.39	0.036
Cannabis	0.67	0.092 / 0.052	0.091 / 0.073	0.214	0.227	0.255	0.032

5.2 Experiments on Real Data

We first present an experiment in which we train a classifier to predict the protected attribute and replace the true attribute by the prediction in the EO training phase. Such a scenario is one of our motivations for studying the EO method under a perturbation of the protected attribute. We perform the experiment on the Drug Consumption data set (Fehrman et al., 2015). It comprises 1885 records of human subjects, and for each subject, it provides five demographic features (e.g., Age, Gender, or Education), seven features measuring personality traits (e.g., Nscore is a measure of neuroticism and Ascore of agreeableness), and 18 features each of which describes the subject's last use of a certain drug (e.g., Cannabis). We set the protected attribute A to be Gender, and, fixing a drug, we set the ground-truth label Y to indicate whether a subject has used the drug within the last decade ($Y=1$) or not ($Y=-1$). Randomly splitting the data set into three batches of equal size, we use the first batch to train a logistic regression classifier that predicts A using the features Nscore and Ascore (these two turned out to work best), and a one-hidden-layer perceptron that predicts Y using the demographic features except Gender and the five features for personality traits other than Nscore and Ascore. We consider the first classifier to provide a perturbed version A_c of the true attribute A and the second classifier to be the given classifier $\widetilde{Y}$. We use the second batch to derive EO predictors $\widehat{Y}_{\text{corr}}$ and $\widehat{Y}_{\text{true}}$ from $\widetilde{Y}$, where $\widehat{Y}_{\text{corr}}$ is based on A_c and $\widehat{Y}_{\text{true}}$ is based on A. The third batch is our test batch, on which we evaluate the bias and the error of $\widetilde{Y}$, $\widehat{Y}_{\text{corr}}$, and $\widehat{Y}_{\text{true}}$, the probability of A_c not equaling A, and also whether Assumptions I and II are satisfied. We measure the extent to which Assumptions I (a) is violated by the estimated l_∞-distance between the conditional (given Y and A) joint distribution of $\widetilde{Y}$ and A_c and the product of their conditional marginal distributions, that is

$$\max_{\substack{y,\tilde{y}\in\{-1,+1\}\\ a,\tilde{a}\in\{0,1\}}} \Big|\Pr\left[\widetilde{Y}=\tilde{y}, A_c=\tilde{a}\middle|Y=y,A=a\right] - \Pr\left[\widetilde{Y}=\tilde{y}\middle|Y=y,A=a\right]\cdot\Pr\left[A_c=\tilde{a}\middle|Y=y,A=a\right]\Big|. \tag{6}$$

Note that, in the distributional setting, Assumptions I (a) is satisfied if and only if this quantity is zero.

Table 1 shows the results for three of the drugs, where we report average results obtained from running the experiment for 200 times. For the sake of readability, we do not report $\text{Bias}_{Y=y}(\widehat{Y}_{\text{true}})$ (which equals zero in the distributional setting) in Table 1. A full table that shows $\text{Bias}_{Y=y}(\widehat{Y}_{\text{true}})$ as well as the results for the other drugs is provided as Table 7 in Appendix A.7. It is $\Pr[A=a]=1/2$ and $\Pr[A_c\neq A|A=a]=0.4$, $a\in\{0,1\}$. Almost always, Assumptions I (b) and Assumption II are satisfied (see Table 6 in Appendix A.7 for details). As we can see from the last column of Table 1 or Table 7, for all the drugs, the measure (6) is rather small, indicating that also Assumptions I (a) might be (almost) satisfied. In this light, the results for the bias and the error of the various classifiers are in accordance with the claims of our paper: we most often have $\text{Bias}_{Y=y}(\widehat{Y}_{\text{corr}}) < \text{Bias}_{Y=y}(\widetilde{Y})$ and we always have $\text{Error}(\widehat{Y}_{\text{corr}}) \leq \text{Error}(\widehat{Y}_{\text{true}})$. We consider finite-sample effects to be responsible for the first inequality not always being true since for Cannabis it even happens that $\text{Bias}_{Y=-1}(\widehat{Y}_{\text{true}}) > \text{Bias}_{Y=-1}(\widetilde{Y})$ (cf. Table 7).

In our second experiment, we run the EO method on two real data sets when we artificially perturb the protected attribute in one of four ways: either we set the attribute of each data point to its complementary value independently with probability γ, or we deterministically flip the attribute of every data point whose score lies in the interval $[0.5-r, 0.5+r]$, or we perturb the attribute in one of these two ways only for those data points for which $\widetilde{Y}\neq Y$. 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. 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 COMPAS criminal recidivism risk assessment data set (Dieterich et al., 2016) and the Adult data set (Dua and Graff, 2019). The scores for the COMPAS data set are the actual scores from the COMPAS risk assessment tool, the scores for the Adult data set are obtained from a multilayer perceptron. We randomly

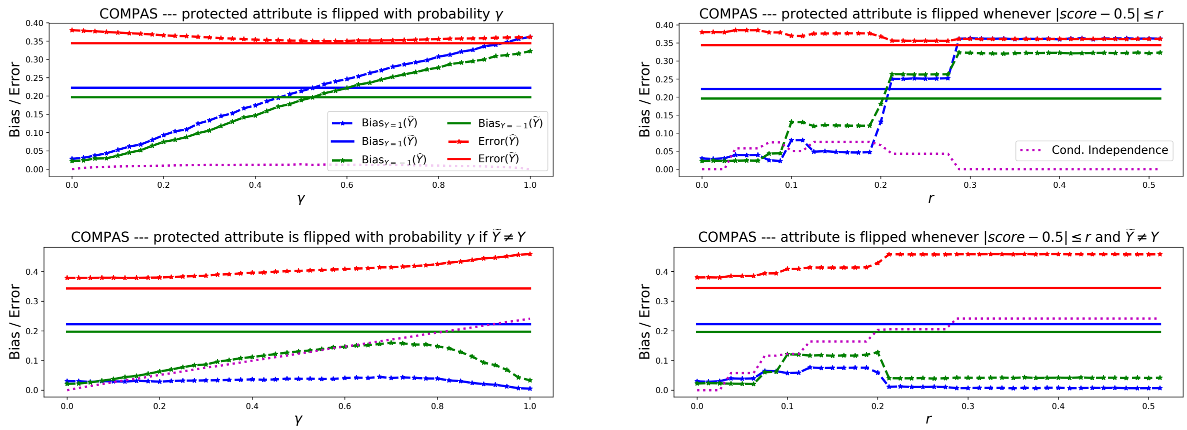

Figure 2: COMPAS data set. $\text{Bias}_{Y=+1/-1}(\widehat{Y})$ (dashed blue / dashed green) and $\text{Error}(\widehat{Y})$ (dashed red) as a function of the perturbation level in four perturbation scenarios. The solid lines show the bias (blue and green) and the error (red) of $\widetilde{Y}$. The magenta line shows an estimate of (6) and how heavily Assumptions I (a) is violated.

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 Appendix A.8). Figure 2 shows the bias and the error of $\widetilde{Y}$ and the derived EO predictor $\widehat{Y}$ as well as an estimate of (6) in the four perturbation scenarios as a function of the perturbation level γ and r, respectively, for the COMPAS data set. Figure 5 in Appendix A.8 shows analogous plots for the Adult data set. The shown curves are obtained from averaging the results of 200 runs of the experiment. In the first two perturbation scenarios, where (6) is small and Assumptions I (a) (almost) satisfied, the curves look quite similar to the ones that we obtained in the experiments of Section 5.1. In the third and the fourth perturbation scenario, Assumptions I (a) is clearly violated, and here the error of $\widehat{Y}$ does not initially decrease. Also, for the Adult data set, the bias of $\widehat{Y}$ explodes even for a moderate perturbation level, which once again shows that our identified Assumptions I are necessary for guaranteeing (3). Overall, also the findings of this experiment confirm the main claims of our paper.

5.3 Repeated Loss Minimization

As another application of our results, we compare the EO method to the method of Hashimoto et al. (2018), discussed in Section 4, in a sequential classification setting. This experiment is presented in Appendix A.9. It shows that just as the method of Hashimoto et al., the EO method can help avoid disparity amplification, even when the protected attribute is highly perturbed.

6 DISCUSSION

We studied the EO postprocessing method of Hardt et al. (2016) for fair classification when the protected attribute is perturbed. We identified conditions on the perturbation that guarantee that the bias of a classifier is reduced even by running the EO method with the perturbed attribute. We showed that our conditions are necessary for providing such a guarantee. For the error of the resulting classifier, we empirically observed that under our conditions and a mild additional assumption, most often the error is not larger than the error of the EO classifier based on the true attribute. In a special case, we formally proved this observation. Importantly, we *analyzed* the EO method as it is and did not try to *modify* it in order to make it more robust. We believe that often the practitioner with domain knowledge can assess whether our conditions hold and hence will benefit from our analysis. In contrast, modifying the method would require additional knowledge about the perturbation probabilities $\Pr[A_c \neq A | Y = y, A = a]$ (e.g., access to some estimates or knowledge about their order) that the practitioner often does not have.

There are several directions for future work: generally, one could analyze any of the many existing methods for fair ML (cf. Section 4) with respect to a perturbation of the protected attribute. Specifically related to our paper, a key question is to fully understand when inequality (5) holds and to provide upper bounds on its violation in case it does not hold. It would also be interesting to study alternative noise models in which the attribute is also corrupted in the test phase (where the corruption can be caused by either the same or a different mechanism as in the training phase). Finally, it would be interesting to extend our results to multiple groups (i.e., a non-binary attribute) or when one only requires $\widehat{Y}$ to have equal true positive rates (aka equality of opportunity). Based on our intuition as outlined in Section 3.4, we believe that such extensions are possible, but they still need to be formally established.

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.

P. Botros and J. M. Tomczak. Hierarchical VampPrior variational fair auto-encoder. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

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 (ACM FAT*)*, 2019.

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

S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023 [cs.CY], 2018.

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 (AIES)*, 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.

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk. arXiv:1506.06297 [stat.AP], 2015. Data available on https://archive.ics.uci.edu/ml/datasets/Drug+consumption+(quantified).

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.

Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, 2018.

N. Kallus and A. Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the XAUC metric. In *Neural Information Processing Systems (NeurIPS)*, 2019.

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.

M. Kearns, S. Neel, and Z. S. Roth, A. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*, 2018.

M. Kearns, S. Neel, and Z. S. Roth, A. Wu. An empirical study of rich subgroup fairness for machine learning. In *Conference on Fairness, Accountability, and Transparency (ACM FAT*)*, 2019.

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 and P. Awasthi. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning (ICML)*, 2018.

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. In *Neural Information Processing Systems (NeurIPS)*, 2019.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016.

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

C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on Learning Theory (COLT)*, 2013.

U. Tantipongpipat, S. Samadi, M. Singh, J. Morgenstern, and S. Vempala. Multi-criteria dimensionality reduction with applications to fairness. In *Neural Information Processing Systems (NeurIPS)*, 2019.

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.