Social Bias Meets Data Bias: The Impacts of Labeling and Measurement Errors on Fairness Criteria

Yiqiao Liao[1], **Parinaz Naghizadeh**[2]

[1] Department of Computer Science and Engineering, The Ohio State University
[2] Department of Electrical and Computer Engineering, The Ohio State University
{liao.489, naghizadeh.1}@osu.edu

Abstract

Although many fairness criteria have been proposed to ensure that machine learning algorithms do not exhibit or amplify our existing social biases, these algorithms are trained on datasets that can themselves be statistically biased. In this paper, we investigate the robustness of existing (demographic) fairness criteria when the algorithm is trained on biased data. We consider two forms of dataset bias: errors by prior decision makers in the labeling process, and errors in the measurement of the features of disadvantaged individuals. We analytically show that some constraints (such as Demographic Parity) can remain robust when facing certain statistical biases, while others (such as Equalized Odds) are significantly violated if trained on biased data. We provide numerical experiments based on three real-world datasets (the FICO, Adult, and German credit score datasets) supporting our analytical findings. While fairness criteria are primarily chosen under normative considerations in practice, our results show that naively applying a fairness constraint can lead to not only a loss in utility for the decision maker, but more severe unfairness when data bias exists. Thus, understanding how fairness criteria react to different forms of data bias presents a critical guideline for choosing among existing fairness criteria, or for proposing new criteria, when available datasets may be biased.

1 Introduction

Machine learning algorithms are being adopted widely in areas ranging from recommendation systems and ad-display, to hiring, loan approvals, and determining recidivism in courts. Despite their potential benefits, these algorithms can still exhibit or amplify existing societal biases (Angwin et al. 2016; Obermeyer et al. 2019; Lambrecht and Tucker 2019). This is referred to as algorithmic *(social) bias* or *unfairness*, as the algorithm makes decisions in favor or against individuals in a way that is inconsistent or discriminatory across groups with different social identities (e.g., race, gender). A commonly proposed method for assessing and preventing these forms of unfairness is through *fairness criteria* (e.g, equality of opportunity, equalized odds, or demographic parity) (Mehrabi et al. 2021; Barocas, Hardt, and Narayanan 2017). These criteria typically require the algorithm to make decisions in a way that (approximately) equalizes a statistical measure (e.g., selection rate, true positive rate) between different groups.

Despite the rising interest in this approach to developing fair algorithms, existing fairness criteria have been largely proposed and evaluated assuming access to unbiased training data. However, existing datasets are often themselves *statistically biased* due to biases or errors made during data collection, labeling, feature measurement, etc (Blum and Stangl 2020; Fogliato, Chouldechova, and G'Sell 2020; Jiang and Nachum 2020; Kallus and Zhou 2018; Wick, Tristan et al. 2019). Any machine learning algorithm is inevitably only as good as the data it is trained on, and so attempts to attain a desired notion of fairness can be thwarted by biases in the training dataset.

Our work identifies the impacts of statistical data biases on the efficacy of existing fairness criteria in addressing social biases. In particular, as we show both analytically and numerically, existing fairness criteria differ considerably in their robustness against different forms of statistical biases in the training data. Although fairness criteria are generally chosen under normative considerations in practice, our results show that naively applying fairness constraints can lead to more severe unfairness when data bias exists. Thus, understanding how fairness criteria react to different forms of data bias can serve as a critical guideline when choosing among existing fairness criteria, or when proposing new criteria.

Overview of our findings and contributions. Formally, we consider a setting in which a firm makes binary decisions (accept/reject) on agents from two demographic groups a and b, with b denoting the disadvantaged group. We assume the training data is statistically biased as a prior decision maker has made errors when either assessing the true qualification state (label) of individuals or when measuring their features. The firm selects its decision rule based on this biased data, potentially subject to one of four fairness criteria: Demographic Parity (DP), True/False Positive Rate Parity (TPR/FPR), or Equalized Odds (EO). Our main findings and contributions are summarized below.

(1) Some fairness criteria are more robust than others. We first analytically show (Proposition 1) that some existing fairness criteria (namely, DP and TPR) are more robust against labeling errors in the disadvantaged group compared to others (FPR and EO). That is, perhaps surprisingly, despite being trained on biased data, the resulting DP/TPR-constrained decision rules continue to satisfy the desired DP/TPR fairness

criteria when implemented on unbiased data. This can be interpreted as a positive byproduct of these fairness criteria, in that (social) fairness desiderata are not violated despite statistical data biases.

(2) Analysis for different forms of statistical data biases. We present similar analyses when the statistical biases are due to feature measurement errors on the disadvantaged group (Proposition 2), and labeling biases on the advantaged group (Proposition 4). We find that different sets of fairness criteria are robust against different forms of data bias.

(3) Guidelines for the selection of fairness criteria and data debiasing. We detail how these observations can be explained based on the effects of each type of data bias on the specific data statistics that a fairness criterion relies on in assessing and imposing its normative desiderata. Our findings can therefore be used to guide targeted data collection and debiasing efforts based on the selected fairness criterion. Alternatively, they could help the decision maker select the most robust fairness criteria among their options, so that it would either continue to be met, or be less drastically impacted, in spite of the suspected types of data bias.

(4) Supporting numerical experiments. We provide support for our analytical findings through numerical experiments based on three real-world datasets: FICO, Adult, and German credit score (Section 4).

(5) Fair algorithms may even increase firm utility if the data is biased. Notably, in contrast to the typically discussed "fairness-accuracy tradeoff", we show that at times using a fair algorithm can *increase* a firm's expected performance compared to an accuracy-maximizing (unfair) algorithm when training datasets are biased. We highlight this observation in Section 4 and provide an intuitive explanation for it by interpreting fairness constraints as having a regularization effect; we also provide an analytical explanation in Appendix E.5.

Related work. The interplay between data biases and fair machine learning has been a subject of growing interest (Ensign et al. 2018; Neel and Roth 2018; Bechavod et al. 2019; Kilbertus et al. 2020; Wei 2021; Blum and Stangl 2020; Jiang and Nachum 2020; Kallus and Zhou 2018; Rezaei et al. 2021; Fogliato, Chouldechova, and G'Sell 2020; Wick, Tristan et al. 2019), and our paper falls within this general category. Most of these works differ from ours in that they focus on the sources of these data biases such as feedback loops, censored feedback, and/or adaptive data collection, and on how these exacerbate algorithmic unfairness, how to debias data, and how to build fair algorithms robust to data bias. In contrast, we investigate how existing fair algorithms fare in the face of statistical data biases (without making adjustments to the algorithm or the data collection procedure), and provide potential guidelines for targeting data debiasing efforts accordingly.

Most closely related to our work are (Blum and Stangl 2020; Jiang and Nachum 2020; Fogliato, Chouldechova, and G'Sell 2020; Wick, Tristan et al. 2019), which also study the interplay between labeling biases and algorithmic fairness. Jiang and Nachum (2020) propose to address label biases in the data directly, by assigning appropriately selected weights to different samples in the training dataset. Blum and Stangl (2020) study labeling biases in the *qualified disadvantaged group*, as well as re-weighing techniques for debiasing data. Further, they show that fairness intervention in the form of imposing Equality of Opportunity can in fact improve the accuracy achievable on biased training data. Fogliato, Chouldechova, and G'Sell (2020) propose a sensitivity analysis framework to examine the fairness of a model obtained from biased data and consider errors in identifying the *unqualified advantaged group*. Wick, Tristan et al. (2019) consider errors in identifying both the *unqualified advantaged group and qualified disadvantaged group* together and focus on the fairness accuracy trade-off when applying different approaches to achieve Demographic Parity. In contrast to these works, we contribute through the study of a more comprehensive set of group fairness criteria, as well as a larger set of statistical biases (two types of labeling bias, and feature measurement errors). Our different analysis approach further allows us to provide new insights into which fairness criteria may remain robust (or even help increase a firm's utility), and why, against each form of statistical data bias.

We review additional related work in Appendix B.

2 Problem Setting

We analyze an environment consisting of a firm (the decision maker) and a population of agents, as detailed below. Table 1 in Appendix C summarizes the notation.

The agents. Consider a population of agents composed of two demographic groups, distinguished by a sensitive attribute $g \in \{a, b\}$. Let $n_g := \mathbb{P}(G = g)$ denote the fraction of the population who are in group g. Each agent has an observable feature $x \in \mathbb{R}$, representing information that is used by the firm in making its decisions; these could be e.g., exam scores or credit scores.[1] The agent further has a (hidden) binary qualification state $y \in \{0, 1\}$, with $y = 1$ and $y = 0$ denoting those qualified and unqualified to receive favorable decisions, respectively. Let $\alpha_g := \mathbb{P}(Y = 1|G = g)$ denote the qualification rate in group g. In addition, let $f_g^y(x) := \mathbb{P}(X = x|Y = y, G = g)$ denote the probability density function (pdf) of the distribution of features for individuals with qualification state y from group g. We make the following assumption on these feature distributions.

Assumption 1. *The pdfs $f_g^y(x)$ and their CDFs $F_g^y(x)$ are continuously differentiable, and the pdfs satisfy the strict monotone likelihood ratio property, i.e., $\frac{f_g^1(x)}{f_g^0(x)}$ is strictly increasing in $x \in \mathbb{R}$.*

This assumption implies that an individual is more likely to be qualified as their feature (score) increases.

We further define the qualification profile of group g as $\gamma_g(x) := \mathbb{P}(Y = 1|X = x, G = g)$, which captures the likelihood that an agent with feature x from group g is qualified. For instance, this could capture estimated repay probabilities given the observed credit scores (which may differ across

[1] We consider one-dimensional features (numerical scores) for ease of exposition in our analysis. Our experiments consider both one-dimensional and n-dimensional features.

groups). We let group b be the group with a lower likelihood of being qualified at the same feature ($\gamma_b(x) \leq \gamma_a(x), \forall x$), and refer to it as the disadvantaged group.

As we show in Section 3, the firm's optimal decision rule can be determined based on the qualification rates α_g and either one of the other problem primitives: the feature distributions $f_g^y(x)$ or the qualification profiles $\gamma_g(x)$. These quantities are related to each other as follows:

$$\gamma_g(x) = \frac{f_g^1(x)\alpha_g}{f_g^1(x)\alpha_g + f_g^0(x)(1-\alpha_g)} = \frac{1}{1+\frac{f_g^0(x)}{f_g^1(x)}(\frac{1}{\alpha_g}-1)} \ . \tag{1}$$

Existing real-world datasets also often provide information on the qualification rates α_g together with either the feature (distributions) $f_g^y(x)$ (e.g., the UCI Adult dataset) or the qualification profiles $\gamma_g(x)$ (e.g., the FICO credit score dataset); see Section 4. We will later detail how data biases (in the form of labeling or feature measurement errors) can be viewed as inaccuracies in these measures.

The firm. A firm makes binary decisions $d \in \{0, 1\}$ on agents from each group based on their observable features, with $d = 0$ and $d = 1$ denoting reject and accept decisions, respectively. The firm gains a benefit of u_+ from accepting qualified individuals, and incurs a loss of u_- from accepting unqualified individuals. The goal of the firm is to select a (potentially group-dependent) decision rule or policy $\pi_g(x) = \mathbb{P}(D = 1|X = x, G = g)$ to maximize its expected payoff. In this paper, we restrict attention to threshold policies $\pi_g(x) = 1(x \geq \theta_g)$, where $1(\cdot)$ denotes the indicator function and θ_g is the decision threshold for group g.[2] Let $U(\theta_a, \theta_b) = n_a U_a(\theta_a) + n_b U_b(\theta_b)$ denote the firm's expected payoff under policies $\{\theta_a, \theta_b\}$, with $U_g(\theta_g)$ denoting the payoff from group g agents.

The firm may further impose a (group) fairness constraint on the choice of its decision rule. While our framework is more generally applicable, we focus our analysis on Demographic Parity (DP) and True/False Positive Rate Parity (TPR/FPR).[3]

Let $\mathcal{C}_a^{\mathtt{f}}(\theta_a) = \mathcal{C}_b^{\mathtt{f}}(\theta_b)$ denote the fairness constraint,[4] where $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$. These constraints can be expressed as follows:

• DP: This constraint equalizes selection rate across groups, and is given by $C_g^{\mathtt{DP}}(\theta) = \int_\theta^\infty (\alpha_g f_g^1(x) + (1-\alpha_g) f_g^0(x))\mathrm{d}x$;

• TPR: Also known as Equality of Opportunity (Hardt, Price, and Srebro 2016), this constraint equalizes the true positive rate across groups, and can be expressed as $\mathcal{C}_g^{\mathtt{TPR}}(\theta) = \int_\theta^\infty f_g^1(x)\mathrm{d}x$;

• FPR: False positive rate parity is defined similarly, with $\mathcal{C}_g^{\mathtt{FPR}}(\theta) = \int_\theta^\infty f_g^0(x)\mathrm{d}x$.

Accordingly, the firm's optimal choice of decision thresholds can be determined by:

$$\max_{\theta_a, \theta_b} \sum_g n_g \int (\alpha_g u_+ f_g^1(x) - (1-\alpha_g) u_- f_g^0(x))\pi_g(x)\mathrm{d}x,$$
$$\text{s.t.} \quad \mathcal{C}_a^{\mathtt{f}}(\theta_a) = \mathcal{C}_b^{\mathtt{f}}(\theta_b) \ . \tag{2}$$

Let $\theta_g^{\mathtt{f}}$ denote the solution of (2) under fairness constraint $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$, and $\theta_g^{\mathtt{MU}}$ denote the Maximum Utility (MU) thresholds (i.e., maximizers of the firm's expected payoff in the absence of a fairness constraint).

Dataset biases. In order to solve (2), the firm relies on historical information and training datasets to obtain estimates of the underlying population characteristics: the qualification rates α_g, the feature distributions $f_g^y(x)$, and/or the qualification profiles $\gamma_g(x)$. However, the estimated quantities $\hat{\alpha}_g$, $\hat{f}_g^y(x)$, and/or $\hat{\gamma}_g(x)$ may differ from the true population characteristics. We refer to the inaccuracies in these estimates as data bias. Specifically, we focus our analysis on the following instantiations of our general model:

1. Qualification assessment (labeling) biases, reflected in the form of errors in profiles $\gamma_g(x)$. We note that such biases can also affect the estimate of α_g and $f_g^y(x)$. This case is most similar to labeling biases considered in prior work (Blum and Stangl 2020; Jiang and Nachum 2020; Fogliato, Chouldechova, and G'Sell 2020; Wick, Tristan et al. 2019). Here, we consider two specific forms of this type of bias:

- *Flipping labels on qualified disadvantaged agents.* We first consider biases that result in $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$, where $\beta \in (0, 1)$ is the underestimation rate. This can be viewed as label biases due to a prior decision maker/policy that only had a probability $\beta < 1$ of correctly identifying qualified agents from the disadvantaged group b. We start with this type of bias as it is one of the most difficult to rectify. Specifically, these biases will not be corrected post-decision due to censored feedback: once a qualified agent is labeled as 0 and rejected, the firm does not get the opportunity to observe the agent and assess whether this was indeed the correct label.
- *Flipping labels on unqualified advantaged agents.* We also consider biases of the form of $\hat{\gamma}_a(x) = (1-\beta)\gamma_a(x) + \beta$, with $\beta \in (0, 1)$, interpreted as prior errors by a decision maker who mistakenly labeled unqualified agents from the advantaged group as qualified with probability β.

2. Feature measurement errors, in the form of drops in the feature distribution likelihood ratios in the disadvantaged group. Formally, we consider biases that result in $\frac{\hat{f}_b^1(x)}{\hat{f}_b^0(x)} = \beta(x)\frac{f_b^1(x)}{f_b^0(x)}, \forall x$, where $\beta(x) : \mathbb{R} \to (0, 1)$ is the underestimation rate and is a non-decreasing function in x (including constant). In words, this results in a firm assessing that an agent with a given feature x is less likely to be qualified than it truly is. This type of bias can occur, for instance, if scores are normally distributed and systematically underestimated such that $\hat{x} = x - \epsilon$, where $\epsilon \geq 0$. This case generalizes measurement biases studied in (Liu et al. 2018).

[2] Prior work (Liu et al. 2018; Zhang et al. 2020) show that threshold policies are optimal under Assumption 1 when selecting fairness-unconstrained policies, and optimal in the fairness-constrained case given additional mild assumptions.

[3] We also study Equalized Odds (EO) (Hardt, Price, and Srebro 2016) in our experiments, which requires both TPR and FPR.

[4] The choice of hard constraints is for theoretical convenience. In Section 4, we allow for soft constraints $|\mathcal{C}_a^{\mathtt{f}}(\theta_a) - \mathcal{C}_b^{\mathtt{f}}(\theta_b)| < \epsilon$.

Note that both the firm's expected payoff (objective function in (2)) and the fairness criteria are impacted by such data biases. In the next sections, we analyze, both theoretically and numerically, the impacts of these types of data biases on the firm's ability to satisfy the desired fairness metric $\texttt{f}$, as well as on the firm's expected payoff.

3 Analytical Results

We begin by characterizing the decision thresholds $\theta_g^{\texttt{MU}}$ that maximize the firm's expected utility, in the absence of any fairness constraints, and investigate the impacts of data biases on these thresholds. All proofs are given in Appendix E.

Lemma 1 (Optimal $\texttt{MU}$ thresholds)**.** *The thresholds $\{\theta_a^{MU}, \theta_b^{MU}\}$ maximizing the firm's utility satisfy $\frac{f_g^1(\theta_g^{MU})}{f_g^0(\theta_g^{MU})} = \frac{(1-\alpha_g)u_-}{\alpha_g u+}$. Equivalently, $\gamma_g(\theta_g^{MU}) = \frac{u_-}{u_+ + u_-}$.*

Lemma 2 (Impact of data biases on $\texttt{MU}$ thresholds and firm's utility)**.** *Let θ_g^{MU} and $\hat{\theta}_g^{MU}$ denote the optimal $\texttt{MU}$ decision thresholds for group g, obtained given unbiased data and data with biases on group b, respectively. If (i) $\hat{\gamma}_b(\theta_b^{MU}) < \gamma_b(\theta_b^{MU})$, or, (ii) $\frac{\hat{f}_b^1(\theta_b^{MU})}{\hat{f}_b^0(\theta_b^{MU})} < \frac{f_b^1(\theta_b^{MU})}{f_b^0(\theta_b^{MU})}$, then the decision threshold on group b increases, i.e., $\hat{\theta}_b^{MU} > \theta_b^{MU}$. The reverse holds if the inequalities above are reversed. In all these cases, the decisions on group a are unaffected, i.e., $\hat{\theta}_a^{MU} = \theta_a^{MU}$. Further, the firm's utility decreases in all cases, i.e., $U(\theta_a^{MU}, \hat{\theta}_b^{MU}) < U(\theta_a^{MU}, \theta_b^{MU})$.*

As intuitively expected, biases against the disadvantaged group (underestimation of their qualification profiles, or scores) lead to an increase in their disadvantage; the reverse is true if the group is perceived more favorably. We also note that the decisions on group a remain unaffected by any biases in group b's data. This implies that if the representation of group b is small, the firm has less incentive for investing resources in removing data biases on group b. In the remainder of this section, we will show that the coupling introduced between the group's decisions due to fairness criteria breaks this independence. A takeaway from this observation is that once a fairness-constrained algorithm couples the decision rules between two groups, it also makes statistical data debiasing efforts advantageous to *both* groups, and therefore increases a (fair) firm's incentives for data debiasing.

The next lemma characterizes the optimal fairness-constrained decision thresholds.

Lemma 3 (Optimal fair thresholds)**.** *The thresholds $\{\theta_a^{\texttt{f}}, \theta_b^{\texttt{f}}\}$ maximizing the firm's expected utility subject to a fairness constraint $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$ satisfy $\sum_g n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{f}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{f}})}{\partial C_g^{\texttt{f}}(\theta_g^{\texttt{f}})/\partial\theta} = 0$.*

This characterization is similar to those obtained in prior works (Zhang et al. 2019, 2020) (our derivation technique is different). Using Lemma 3, we can further characterize the thresholds for fairness criteria $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$, as derived in Tables 3 and 4 in Appendix E.4. These form the basis of the next set of results, which shed light on the sensitivity of different fairness criteria to biased training data.

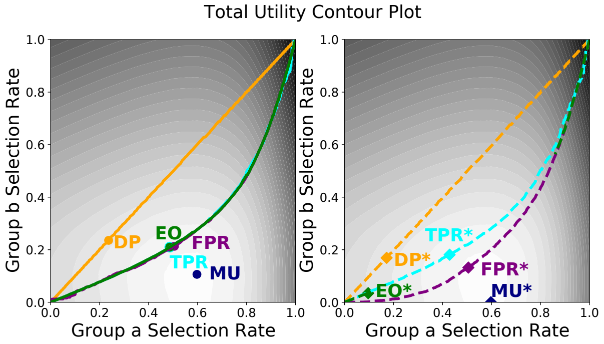


Figure 1: Firm's utility as a function of selection rates. Brighter regions represent higher utility. The curves highlight solutions satisfying the fairness constraints. Left: unbiased data. Right: biased data with 20% of the qualification states of the qualified agents from group b flipped from 1 to 0.

3.1 Impacts of Labeling Biases

We now assess the sensitivity of fairness-constrained policies to biases in qualification assessment (labeling). We first consider biases that result in $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$, where $\beta \in (0, 1]$ is the underestimation rate (this could be due to, e.g., labeling errors on qualified individuals from the disadvantaged group). We first analyze the impacts of such biases on the decision thresholds and on the firm's utility.

Proposition 1. *Assume the qualification profile of group b is underestimated so that $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$, where $\beta \in (0, 1]$. Let $\theta_g^{\texttt{f}}$ and $\hat{\theta}_g^{\texttt{f}}(\beta)$ denote the optimal decision thresholds satisfying fairness constraint $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$, obtained from unbiased data and from data with biases on group b given β, respectively. Then,*

(i) $\hat{\theta}_g^{\texttt{f}}(\beta) \geq \theta_g^{\texttt{f}}$ for $g \in \{a, b\}$, $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$, $\beta \in (0, 1]$. Further, $\hat{\theta}_g^{\texttt{f}}(\beta)$ is decreasing in β.

(ii) The $\texttt{DP}$ and $\texttt{TPR}$ criteria continue to be met, while $\texttt{FPR}$ is violated, at their $\{\hat{\theta}_a^{\texttt{f}}(\beta), \hat{\theta}_b^{\texttt{f}}(\beta)\}$.

(iii) The firm's utility decreases under $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}\}$, i.e., $U(\hat{\theta}_a^{\texttt{f}}(\beta), \hat{\theta}_b^{\texttt{f}}(\beta)) \leq U(\theta_a^{\texttt{f}}, \theta_b^{\texttt{f}})$.

(iv) The firm's utility may increase or decrease under $\texttt{FPR}$.

Figure 1 illustrates Proposition 1 on a synthetic dataset inspired by the FICO dataset (Hardt, Price, and Srebro 2016).

The proof of this proposition relies on the characterizations of the optimal fairness-constrained thresholds in Lemma 3, together with identifying the changes in the problem primitives when $\hat{\gamma}_b(x) = \beta\gamma_b(x)$. In particular, we show that $\hat{\alpha}_b(x) = \beta\alpha_b(x)$, $\hat{f}_b^1(x) = f_b^1(x)$, and $\hat{f}_b^0(x) = \frac{1-\alpha_b}{1-\beta\alpha_b}\frac{1-\beta\gamma_b(x)}{1-\gamma_b(x)}f_b^0(x)$. Intuitively, these changes can be explained as follows: $\hat{\gamma}_b(x) = \beta\gamma_b(x)$ can be viewed as flipping label 1 to label 0 in the training data on group b with probability β. This leaves the feature distribution of qualified agents unchanged, whereas it adds (incorrect) data on unqualified agents, hence biasing $f_b^0(x)$. Such label flipping also decreases estimated qualification rates α_b by a factor β. Using these, we show that $\texttt{DP}/\texttt{TPR}$ continue to hold at the biased thresholds given the changes in the statistics they rely on, while FPR is violated. As we show later, the impacts of other types of data bias, and the robustness of any fairness criteria against them, can be similarly tracked to the impacts of those statistical data biases on different data statistics.

We next note two main differences of this lemma with Lemma 2 in the unconstrained setting: (1) the biases in group b's data now lead to under-selection of *both* groups compared to the unbiased case. That is, the introduction of fairness constraints couples the groups in the impact of data biases as well. (2) Perhaps more interestingly, there exist scenarios in which the adoption of a fairness constraint *benefits* a firm facing biased qualification assessments. (We provide additional intuition for this in Appendix E.5). Note however that the fairness criterion is no longer satisfied in such scenarios.

In addition, Proposition 1 shows that the DP and TPR fairness criteria are *robust* to underestimation of qualification profiles of the disadvantaged group, in that the obtained thresholds continue to satisfy the desired notion of fairness. That said, the proposition also states that the pair of decision thresholds $\{\theta_a^{\texttt{f}}, \theta_b^{\texttt{f}}\}$ are different from (and higher than) those that would be obtained if data was unbiased, and hence lead to the loss of utility for the firm. To better assess the impacts of these changes on the firm's expected payoff, we investigate the sensitivity of DP and TPR thresholds to the error rate β.

Formally, we can use the results of Lemma 3 together with the implicit function theorem to characterize $\frac{\partial\hat{\theta}_g^{\texttt{f}}(\beta)}{\partial\beta}$, the rates of change in the thresholds as a function of the underestimation rate β; see Proposition 3 in Appendix E.7. The following corollary of that proposition shows that, under mild conditions, DP is more sensitive to qualification assessment biases than TPR when facing the same bias rates.

Corollary 1. *Consider $\frac{\partial\hat{\theta}_b^{\texttt{f}}(1)}{\partial\beta}$, the rate of change of group b's thresholds at $\beta = 1$. There exists a $\bar{\alpha}_b$ such that for all $\alpha_b \leq \bar{\alpha}_b$, we have $|\frac{\partial\hat{\theta}_b^{\texttt{TPR}}(1)}{\partial\beta}| < |\frac{\partial\hat{\theta}_b^{\texttt{DP}}(1)}{\partial\beta}|$; that is, DP is more sensitive to qualification assessment biases than TPR.*

This higher sensitivity of DP compared to TPR leads to a higher drop in the utility of the firm when DP-constrained classifiers are used on such biased datasets; we further illustrate this in our experiments in Section 4.

Finally, our analysis can be similarly applied to study the impacts of labeling biases on the *advantaged group*; we detail this analysis in Appendix E.8. In particular, we consider biases of the form of $\hat{\gamma}_a(x) = (1-\beta)\gamma_a(x)+\beta$, with $\beta \in [0, 1)$, interpreted as prior errors by a decision maker who mistakenly labeled unqualified individuals from the advantaged group as qualified with probability β. In Proposition 4, we show that this time, DP and FPR are robust against these biases, while TPR is in general violated. Notably, DP remains robust against *both* types of qualification assessment bias. Our experiments in Section 4 further support this observation by showing that DP-constraint thresholds are more robust to label flipping biases induced in different real-world datasets.

3.2 Impacts of Feature Measurement Errors

We now analyze the sensitivity of fairness-constrained decisions to an alternative form of statistical biases: errors in feature measurements of the disadvantaged group.

Proposition 2. *Assume the features of group b are incorrectly measured, so that $\frac{\hat{f}_b^1(x)}{\hat{f}_b^0(x)} = \beta(x)\frac{f_b^1(x)}{f_b^0(x)}, \forall x$, where $\beta(x) : \mathbb{R} \to (0, 1)$ is a non-decreasing function. Let $\theta_g^{\texttt{f}}$ and $\hat{\theta}_g^{\texttt{f}}(\beta)$ denote the optimal decision thresholds satisfying fairness constraint $\texttt{f} \in \{$DP, TPR, FPR$\}$, obtained from unbiased data and data with biases on group b with error function β, respectively. Then,*

(i) If $\hat{f}_b^1(x) = f_b^1(x), \forall x$ (resp. $\hat{f}_b^0(x) = f_b^0(x), \forall x$), TPR (resp. FPR) will be met at the new thresholds.

(ii) If $\hat{F}_b^1(x) < F_b^1(x), \forall x \geq \theta_b^{\texttt{TPR}}$, then $\hat{\theta}_g^{\texttt{TPR}}(\beta) > \theta_g^{\texttt{TPR}}$ for both groups and any function $\beta(x)$. Further, the TPR constraint is violated at the new thresholds.

(iii) If $\hat{F}_b^0(x) < F_b^0(x), \forall x \geq \theta_b^{\texttt{FPR}}$, then $\hat{\theta}_g^{\texttt{FPR}}(\beta) > \theta_g^{\texttt{FPR}}$ for both groups and any function $\beta(x)$. Further, the FPR constraint is violated at the new thresholds.

(iv) If $\hat{F}_b(x) < F_b(x), \forall x \geq \theta_b^{\texttt{DP}}$, then $\hat{\theta}_g^{\texttt{DP}}(\beta) > \theta_g^{\texttt{DP}}$ for both groups and any function $\beta(x)$. Further, the DP constraint is violated at the new thresholds.

(v) There exist problem instances in which $\hat{\theta}_b^{\texttt{f}}(\beta) < \theta_b^{\texttt{f}}$ for any of the three constraints.

We provide a visualization of this proposition in Figure 10 in Appendix D. This proposition shows that unless the feature measurement errors only affect one label (as in part (i)), the considered fairness constraints will in general not remain robust against feature measurement errors. The conditions in parts (ii)-(iv) require that a CDF in the biased distribution first-order stochastically dominates that of the unbiased distribution. This holds if, e.g., the corresponding features (qualified agents, unqualified agents, or all agents, respectively) are underestimated, $\hat{x} = x-\epsilon$ for some $\epsilon \geq 0$. We also note that in contrast to Proposition 1, the decision threshold can in fact *decrease* when biases are introduced; we illustrate this in our experiments in Section 4.

Similar to Proposition 3, we can also characterize the sensitivity of each constraint to bias rates, and investigate the impacts of other problem parameters on these sensitivities. We present this in detail in Proposition 5 in Appendix E.10.

4 Numerical Experiments

We now provide numerical support for our analytical results, and additional insights into the robustness of different fairness measures, through experiments on both real-world and synthetic datasets. Details about the datasets, experimental setup, and additional experiments, are given in Appendix D.

4.1 FICO Credit Score Dataset

We begin with numerical experiments on the FICO dataset preprocessed by (Hardt, Price, and Srebro 2016). The FICO credit scores ranging from 300 to 850 correspond to the one-dimensional feature x in our model, and race is the sensitive feature g (we focus on the white and black groups). The data provides repay probabilities for each score and group, which corresponds to our qualification profile $\gamma_g(x)$. We take this data to be the unbiased ground truth. (We discuss some implications of this assumption in Appendix A.) To induce labeling biases in the data, we drop the repay probabilities

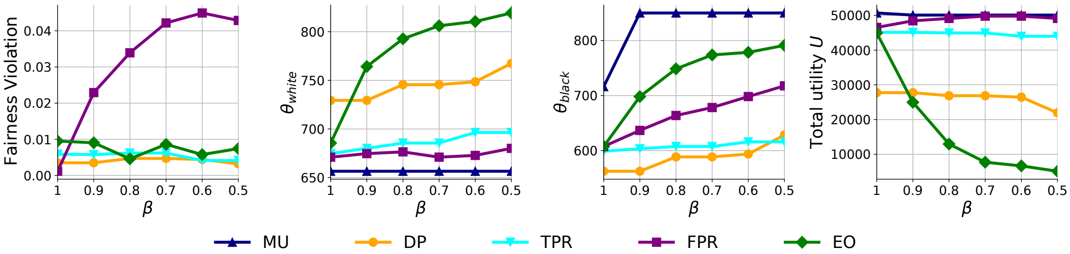


Figure 2: Experiments on qualification assessment (labeling) errors on the disadvantaged group in the FICO credit score dataset.

of the black group to model the underestimation of their qualification profiles, and generate training data on this group based on this biased profile. We use β to parametrize the bias (with $\hat{\gamma}_b(x) = \beta\gamma_b(x)$). Decision rules will be found on the biased data and applied to the unbiased data.

Violation of fairness constraints. The left-most panel in Figure 2 illustrates the fairness violation under each fairness constraint (measured as $|\mathcal{C}_a^{\mathsf{f}}(\hat{\theta}_a^{\mathsf{f}}) - \mathcal{C}_b^{\mathsf{f}}(\hat{\theta}_b^{\mathsf{f}})|$) as qualification assessment biases on the disadvantaged group increase. These observations are consistent with Proposition 1. In particular, DP and TPR are both robust to these biases in terms of achieving their notions of fairness, while FPR has an increasing trend in fairness violation. This means that the set of possible decision rules of FPR changes when bias is imposed. Note that though a violation of 0.04 may not be severe, it is more than 300% higher than a violation below 0.01 that FPR can achieve when the data is unbiased. We will also observe more significant violations of FPR on other datasets. Finally, from Figure 2, it may seem that EO also remains relatively robust to bias. This observation is not in general true (as shown in our experiments on other datasets in Section 4.2); however, it can be explained for the FICO dataset by noting how EO's feasible pairs of decision rules change due to data bias (similar to Figure 1) and how the problem setup can influence the results. We provide additional discussion in Appendix D.2.

Changes in decision thresholds and firm's utility. In Figure 2, we also display the threshold change of each group. In line with Proposition 1, thresholds for both groups increase as the bias level increases. Notably, the maximum utility (fairness unconstrained) decision rule would have led the firm to fully exclude the black group even at relatively low bias rates; all fairness-constrained thresholds prevent this from happening. In addition, the threshold's increase under TPR is less drastic than DP and EO (and consistent with Corollary 1).

Finally, we note the changes in the firm's utility. As the decision thresholds increase due to biases, the net utility from the white/black groups (U_{white}, U_{black}) increases/decreases. Overall, due to the fact that the white group is the majority in this data, an increase in the threshold θ_{white} will lead to a greater loss in total utility of the firm (as is the case in DP/TPR/EO seen in Figure 2). That said, the total utility may increase under FPR, as pointed out in Proposition 1 and observed in Figure 2, since there is a gain from U_{black} is larger than the loss from U_{white}. This increase may even make the FPR-constrained classifier attain higher utility than MU when training data is highly biased.

4.2 Adult Dataset and German Credit Dataset

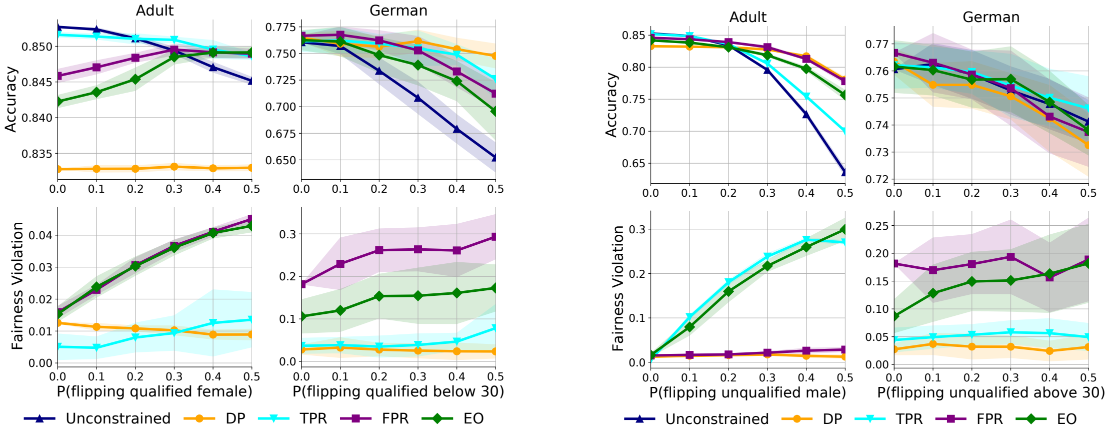
We next conduct experiments on two additional benchmark datasets: the Adult dataset and the German credit dataset (Dua and Graff 2017). In both these datasets, instead of maximizing utility, the objective is classification accuracy. We first train a logistic regression classifier on the training set using scikit-learn (Pedregosa et al. 2011) with default parameter settings as the base model. The logistic regression output in the range 0 to 1 can be interpreted as the score of being qualified as in the FICO dataset. Then, we obtain the fair classifier by applying the exponentiated gradient reduction (Agarwal et al. 2018) using Fairlearn (Bird et al. 2020). Although the exponentiated gradient reduction produces a randomized classifier, it can be viewed as an abstract threshold: given a randomized classifier, the expected rates (e.g., selection rates, true positive rates) for both groups are deterministic. We thus find that our claims on how labeling biases impact DP/TPR/FPR still hold.

We introduce qualification assessment biases by flipping the qualification states y of the (1) qualified agents from the disadvantaged group (i.e., female in Adult and age below 30 in German), (2) unqualified agents from the advantaged group (i.e., male in Adult and age above 30 in German).[5]

The results are presented in Figure 3. To quantify the trend in fairness violation, we fit a linear regression model to the fairness violation and present all model weights in Table 2 in Appendix D.3. In all three cases, the robustness of each constraint in terms of achieving fairness matches our findings in Propositions 1 and 4. One exception, however, is that TPR remains robust when we flip the labels of the unqualified advantaged group in the German dataset. This is primarily because, while flipping the unqualified individuals in the training set will in general make the classifier accept more of these individuals, the flip will have a minor effect on the TPR violation because 1) there is a limited number of unqualified individuals with age above 30 (15.2% of the dataset) compared to qualified individuals with age above 30 (43.7% of the dataset), and 2) there is little room for the true positive rate values on the test set to increase since the values for both groups start close to 1 (0.923 and 0.845) with a small difference (0.078) (see Figure 8 in Appendix D.3).

Interestingly, we also observe that fairness-constrained ac-

[5] Results from concurrently flipping the labels of qualified agents from the disadvantaged group and unqualified agents from the advantaged group are shown in Appendix D.3.



(a) Flip the qualified disadvantaged group.

(b) Flip the unqualified advantaged group.

Figure 3: Accuracy and fairness violation on Adult and German datasets. The results are averaged across 10 runs for Adult and 30 runs for German; shaded areas indicate plus/minus standard deviation.

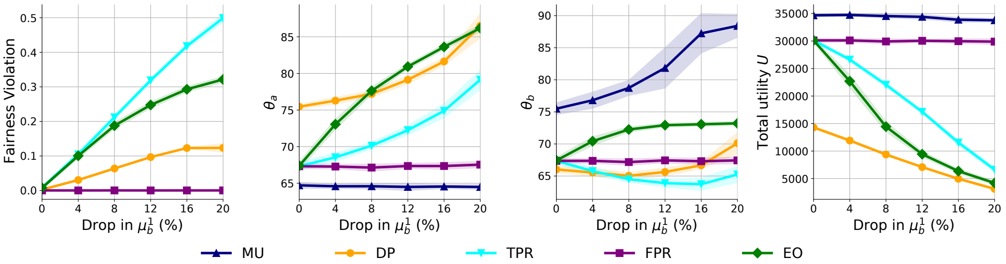


Figure 4: Fairness violation, thresholds, and utility under different constraints from synthetic simulation on measurement errors. The results are from 20 runs.

curacy can be *higher* than the utility-maximizing choices when training data is biased, as seen in the top row accuracy plots in Figure 3. This can be interpreted as fairness constraints having a regularization effect: since the constraints prevent the classifier from overfitting the biased training set to some extent, the test accuracy with fairness constraints imposed can be higher than that of the unconstrained case.

4.3 Impacts of Feature Measurement Errors

Lastly, we conduct experiments on a synthetic dataset inspired by the FICO credit score data (details in Appendix D.1). To bias the feature measurements, we drop the estimate $\hat{\mu}_b^1$ of the mean of the qualified agents from group b relative to the true value μ_b^1. As a result, $\hat{f}_b^1$ will be biased relative to its true value, while $\hat{f}_b^0$ will remain unchanged. As shown in Figure 4, and consistent with Proposition 2, under these choices, FPR will remain unaffected, while DP/TPR will no longer be satisfied.

Finally, Figure 4 also highlights the changes in the decision thresholds and firm's utility under this type of bias. As noted in Proposition 2, now the thresholds on the disadvantaged group can *decrease* compared to the unbiased case at low bias rates. Notably, we also observe that the firm's overall utility is lower under DP (similar to the labeling bias case), but that TPR is more sensitive to bias levels than DP (unlike the labeling bias case). This points to the fact that the choice of a robust fairness constraint has to be made subject to the type of data bias that the decision maker foresees.

5 Conclusion

We investigated the robustness of different fairness criteria when an algorithm is trained on statistically biased data. We provided both analytical results and numerical experiments based on three real-world datasets (FICO, Adult, and German credit score). We find that different constraints exhibit different sensitivity to labeling biases and feature measurement errors. In particular, we identified fairness constraints that can remain robust against certain forms of statistical biases (e.g., Demographic Parity and Equality of Opportunity given labeling biases on the disadvantaged group), as well as instances in which the adoption of a fair algorithm can increase the firm's expected utility when training data is biased, providing additional motivation for adopting fair machine learning algorithms. Our findings present an additional guideline to go along with normative considerations, for choosing among existing fairness criteria when available datasets are biased. We provide additional discussion about other implications of our findings, limitations, and future directions, in Appendix A.

Acknowledgments

The authors are grateful for support from the NSF program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040800, and Cisco Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, Amazon, or Cisco.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2022-03-01.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2.

Bechavod, Y.; Ligett, K.; Roth, A.; Waggoner, B.; and Wu, S. Z. 2019. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, 8974–8984.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Blum, A.; and Stangl, K. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Accessed: 2022-03-01.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, 160–171. PMLR.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, 2325–2336. PMLR.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Jiang, H.; and Nachum, O. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, 702–712.

Kallus, N.; and Zhou, A. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, 2439–2448. PMLR.

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

Kilbertus, N.; Rodriguez, M. G.; Schölkopf, B.; Muandet, K.; and Valera, I. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, 277–287. PMLR.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 65(7): 2966–2981.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Neel, S.; and Roth, A. 2018. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, 3720–3729. PMLR.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9419–9427.

Wang, J.; Liu, Y.; and Levy, C. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 526–536.

Wei, D. 2021. Decision-Making Under Selective Labels: Optimal Finite-Domain Policies and Beyond. In *International Conference on Machine Learning*, 11035–11046. PMLR.

Wick, M.; Tristan, J.-B.; et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Zhang, X.; Khaliligarekani, M.; Tekin, C.; et al. 2019. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.