Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS)
{mzafar, ivalera, manuelgr, gummadi}@mpi-sws.org

ABSTRACT

Automated data-driven decision making systems are increasingly being used to assist, or even replace humans in many settings. These systems function by learning from historical decisions, often taken by humans. In order to maximize the utility of these systems (or, classifiers), their training involves minimizing the errors (or, misclassifications) over the given historical data. However, it is quite possible that the optimally trained classifier makes decisions for people belonging to different social groups with different *misclassification* rates (e.g., misclassification rates for females are higher than for males), thereby placing these groups at an unfair disadvantage. To account for and avoid such unfairness, in this paper, we introduce a new notion of unfairness, *disparate mistreatment*, which is defined in terms of misclassification rates. We then propose intuitive measures of disparate mistreatment for decision boundary-based classifiers, which can be easily incorporated into their formulation as convex-concave constraints. Experiments on synthetic as well as real world datasets show that our methodology is effective at avoiding disparate mistreatment, often at a small cost in terms of accuracy.

1. INTRODUCTION

The emergence and widespread usage of automated data-driven decision making systems in a wide variety of applications, ranging from content recommendations to pretrial risk assessment, has raised concerns about their potential unfairness towards people with certain traits [8, 22, 24, 27]. Anti-discrimination laws in various countries prohibit unfair treatment of individuals based on specific traits, also called *sensitive* attributes (*e.g.*, gender, race). These laws typically distinguish between two different notions of unfairness [5] namely, ***disparate treatment*** and ***disparate impact***. More specifically, there is disparate treatment when the decisions an individual user receives change with changes to her sensitive attribute information, and there is disparate impact when the decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups. A number of recent studies [10, 21, 29], including our own prior work [28], have focused on designing decision making systems that avoid one or both of these types of unfairness.

These prior designs have attempted to tackle unfairness in decision making scenarios where the historical decisions in the training data are *biased* (*i.e.*, groups of people with certain sensitive attributes may have historically received unfair treatment) and there is no ground truth about the *correctness* of the historical decisions (*i.e.*, one cannot tell whether a historical decision used during the training phase was right or wrong). However, when the ground truth for historical decisions is available, disproportionately beneficial outcomes for certain sensitive attribute value groups can be justified and explained by means of the ground truth. Therefore, disparate impact would not be a suitable notion of unfairness in such scenarios.

In this paper, we propose an alternative notion of unfairness, ***disparate mistreatment***, especially well-suited for scenarios where ground truth is available for historical decisions used during the training phase. We call a decision making process to be suffering from disparate mistreatment with respect to a given sensitive attribute (*e.g.*, race) if the *misclassification rates differ* for groups of people having different values of that sensitive attribute (*e.g.*, blacks and whites). For example, in the case of the NYPD Stop-question-and-frisk program (SQF) [1], where pedestrians are stopped on the suspicion of possessing an illegal weapon [12], having different weapon discovery rates for different races would constitute a case of disparate mistreatment.

In addition to *all* misclassifications in general, depending on the application scenario, one might want to measure disparate mistreatment with respect to different kinds of misclassifications. For example, in pretrial risk assessments, the decision making process might only be required to ensure that the false positive rates are equal for all groups, since it may be more acceptable to let a guilty person go, rather than incarcerate an innocent person.[1] On the other hand, in loan approval systems, one might instead favor a decision making process in which the false negative rates are equal, to ensure that deserving (positive class) people with a certain sensitive attribute value are not denied (negative class) loans disproportionately. Similarly, depending on the application

An open-source code implementation of our scheme is available at: **http://fate-computing.mpi-sws.org/**

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052660

[1] *"It is better that ten guilty persons escape than that one innocent suffer"*—William Blackstone

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

Figure 1: Decisions of three fictitious classifiers (C_1, C_2 and C_3) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon. Gender is a sensitive attribute, whereas the other two attributes (suspicious bulge in clothing and proximity to a crime scene) are non-sensitive. Ground truth on whether the person is actually in possession of an illegal weapon is also shown.

scenario at hand, and the cost of the type of misclassification, one may choose to measure disparate mistreatment using false discovery and false omission rates, instead of false positive and false negative rates (see Table 1).

In the remainder of the paper, we first formalize disparate treatment, disparate impact and disparate mistreatment in the context of (binary) classification. Then, we introduce intuitive measures of disparate mistreatment for decision boundary-based classifiers and show that, for a wide variety of linear and nonlinear classifiers, these measures can be incorporated into their formulation as convex-concave constraints. The resulting formulation can be solved efficiently using recent advances in convex-concave programming [26]. Finally, we experiment with synthetic as well as real world datasets and show that our methodology can be effectively used to avoid disparate mistreatment.

2. BACKGROUND AND RELATED WORK

In this section, we first elaborate on the three different notions of unfairness in automated decision making systems using an illustrative example and then provide an overview of the related literature.

Disparate mistreatment. Intuitively, disparate mistreatment can arise in any automated decision making system whose outputs (or decisions) are not perfectly (*i.e.*, 100%) accurate. For example, consider a decision making system that uses a logistic regression classifier to provide binary outputs (say, positive and negative) on a set of people. If the items in the training data with positive and negative class labels are not linearly separable, as is often the case in many real-world application scenarios, the system will misclassify (*i.e.*, produce false positives, false negatives, or both, on) some people. In this context, the misclassification rates may be different for groups of people having different values of sensitive attributes (*e.g.*, males and females; blacks and whites) and thus disparate mistreatment may arise.

Figure 1 provides an example of decision making systems (classifiers) with and without disparate mistreatment. In all cases, the classifiers need to decide whether to stop a pedestrian—on the suspicion of possessing an illegal weapon—using a set of features such as bulge in clothing and proximity to a crime scene. The "ground truth" on whether a pedestrian actually possesses an illegal weapon is also shown. We show decisions made by three different classifiers $\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$. We deem $\mathbf{C_1}$ and $\mathbf{C_2}$ as unfair due to disparate mistreatment because their rate of erroneous decisions for males and females are different: $\mathbf{C_1}$ has different false negative rates for males and females (0.0 and 0.5, respectively), whereas $\mathbf{C_2}$ has different false positive rates (0.0 and 1.0) as well as different false negative rates (0.0 and 0.5) for males and females.

Disparate treatment. In contrast to disparate mistreatment, disparate treatment arises when a decision making system provides different outputs for groups of people with the same (or similar) values of non-sensitive attributes (or features) but different values of sensitive attributes.

In Figure 1, we deem $\mathbf{C_2}$ and $\mathbf{C_3}$ to be unfair due to disparate treatment since $\mathbf{C2}$'s ($\mathbf{C3}$'s) decisions for $Male$ 1 and $Female$ 1 ($Male$ 2 and $Female$ 2) are different even though they have the same values of non-sensitive attributes. Here, disparate treatment corresponds to the very intuitive notion of fairness: two otherwise similar persons should not be treated differently solely because of a difference in gender.

Disparate impact. Finally, disparate impact arises when a decision making system provides outputs that benefit (hurt) a group of people sharing a value of sensitive attribute more frequently than other groups of people.

In Figure 1, assuming that a pedestrian benefits from a decision of not being stopped, we deem $\mathbf{C_1}$ as unfair due to disparate impact because the fraction of males and females that were stopped are different (1.0 and 0.66, respectively).

Application scenarios for disparate impact vs. disparate mistreatment. Note that unlike in the case of disparate mistreatment, the notion of disparate impact is independent of the "ground truth" information about the decisions, *i.e.*, whether or not the decisions are correct or valid. Thus, the notion of disparate impact is particularly appealing in application scenarios where ground truth information for decisions does not exist and the historical decisions used during training are not reliable and thus cannot be trusted. Unreliability of historical decisions for automated decision making systems is particularly concerning in scenarios like recruiting or loan approvals, where biased judgments by humans in the past may be used when training classifiers for the future. In such application scenarios, it is hard to distinguish correct and incorrect decisions, making it hard to assess or use disparate mistreatment as a notion of fairness.

However, in scenarios where ground truth information for decisions can be obtained, disparate impact can be quite misleading as a notion of fairness. That is, in scenarios where the validity of decisions can be reliably ascertained, it would be possible to distinguish disproportionality in decision outcomes for sensitive groups that arises from justifiable reasons (*e.g.*, qualification of the candidates) and disproportionality that arises for non-justifiable reasons (*i.e.*,

discrimination against certain groups). By requiring decision outcomes to be proportional, disparate impact risks introducing reverse-discrimination against qualified candidates. Such practices have previously been deemed unlawful by courts (*Ricci vs. DeStefano, 2009*). In contrast, when the correctness of decisions can be determined, disparate mistreatment can not only be accurately assessed, but also avoids reverse-discrimination, making it a more appealing notion of fairness.

Related Work. There have been a number of studies, including our own prior work [28], proposing methods for detecting [10, 21, 23, 25] and removing [9, 10, 13, 16, 17, 23, 28, 29] unfairness when it is defined in terms of disparate treatment, disparate impact or both. However, as pointed out earlier, the disparate impact notion might be less meaningful in scenarios where ground truth decisions are available.

A number of previous studies have pointed out racial disparities in both automated [4] as well as human [12, 14] decision making systems related to criminal justice. For example, a recent work by Goel et al. [12] detects racial disparities in NYPD SQF program, inspired by a notion of unfairness similar to our notion of disparate mistreatment. More specifically, it uses ground truth (stops leading to successful discovery of an illegal weapon on the suspect) to show that blacks were treated unfairly since false positive rates in stops were higher for them than for whites. The study's findings provide further justification for the need for data-driven decision making systems without disparate mistreatment.

A recent work by Hardt et al. [15] (concurrently conducted with our work) proposes a method to achieve a fairness notion equivalent to our notion of disparate mistreatment. This method works by post-processing the probability estimates of an unfair classifier to learn different decision thresholds for different sensitive attribute value groups, and applying these group-specific thresholds at decision making time. Since this method requires the sensitive attribute information at decision time, it cannot be used in cases where sensitive attribute information is unavailable (*e.g.*, due to privacy reasons) or prohibited from being used due to disparate treatment laws [5].

3. FORMALIZING NOTIONS OF FAIRNESS

In a binary classification task, the goal is to learn a mapping $f(\mathbf{x})$ between user feature vectors $\mathbf{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. Learning this mapping is often achieved by finding a decision boundary $\boldsymbol{\theta}^*$ in the feature space that minimizes a certain loss $L(\boldsymbol{\theta})$, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, computed on a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Then, for a given *unseen* feature vector $\mathbf{x}$, the classifier predicts the class label $\hat{y} = f_{\boldsymbol{\theta}^*}(\mathbf{x}) = 1$ if $d_{\boldsymbol{\theta}^*}(\mathbf{x}) \geq 0$ and $\hat{y} = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\mathbf{x})$ denotes the signed distance from $\mathbf{x}$ to the decision boundary. Assume that each user has an associated sensitive feature z. For ease of exposition, we assume z to be binary, *i.e.*, $z \in \{0, 1\}$. However, our setup can be easily generalized to categorical as well as multiple sensitive features.

Given the above terminology, we can formally express the absence of disparate treatment, disparate impact and disparate mistreatment as follows:

Existing notion 1: Avoiding disparate treatment. A binary classifier does not suffer from disparate treatment if:

$$P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x}), \tag{1}$$

		Predicted Label $\hat{y}=1$	Predicted Label $\hat{y}=-1$	
True Label	$y=1$	True positive	False negative	$P(\hat{y} \neq y \mid y=1)$ False Negative Rate
True Label	$y=-1$	False positive	True negative	$P(\hat{y} \neq y \mid y=-1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y}=1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y}=-1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Table 1: In addition to the overall misclassification rate, error rates can be measured in two different ways: false negative rate and false positive rate are defined as fractions over the *class distribution in the ground truth labels*, or true labels. On the other hand, false discovery rate and false omission rate are defined as fractions over the *class distribution in the predicted labels*.

i.e., if the probability that the classifier outputs a specific value of $\hat{y}$ given a feature vector $\mathbf{x}$ does not change after observing the sensitive feature z, there is no disparate treatment.

Existing notion 2: Avoiding disparate impact. A binary classifier does not suffer from disparate impact if:

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1), \tag{2}$$

i.e., if the probability that a classifier assigns a user to the positive class, $\hat{y} = 1$, is the same for both values of the sensitive feature z, then there is no disparate impact.

New notion 3: Avoiding disparate mistreatment. A binary classifier does not suffer from disparate mistreatment if the misclassification rates for different groups of people having different values of the sensitive feature z are the same. Table 1 describes various ways of measuring misclassification rates. Specifically, misclassification rates can be measured as fractions over the *class distribution in the ground truth labels*, *i.e.*, as false positive and false negative rates, or over the *class distribution in the predicted labels*, *i.e.*, as false omission and false discovery rates.[2] Consequently, the absence of disparate mistreatment in a binary classification task can be specified with respect to the different misclassification measures as follows:

overall misclassification rate (OMR):

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1), \tag{3}$$

false positive rate (FPR):

$$P(\hat{y} \neq y|z = 0, y = -1) = P(\hat{y} \neq y|z = 1, y = -1), \tag{4}$$

false negative rate (FNR):

$$P(\hat{y} \neq y|z = 0, y = 1) = P(\hat{y} \neq y|z = 1, y = 1), \tag{5}$$

false omission rate (FOR):

$$P(\hat{y} \neq y|z = 0, \hat{y} = -1) = P(\hat{y} \neq y|z = 1, \hat{y} = -1), \tag{6}$$

false discovery rates (FDR):

$$P(\hat{y} \neq y|z = 0, \hat{y} = 1) = P(\hat{y} \neq y|z = 1, \hat{y} = 1). \tag{7}$$

[2] In prediction tasks where a positive prediction entails a large cost (*e.g.*, cost involved in the treatment of a disease) one might be more interested in measuring error rates as fractions over the class distribution in the *predicted labels*, rather than over the class distribution in the *ground truth labels*, *e.g.*, to ensure that the false discovery rates, instead of false positive rates, for all groups are the same.

In the following section, we introduce a method to eliminate disparate mistreatment from decision boundary-based classifiers when disparate mistreatment is defined in terms of overall misclassification rate, false positive rate and false negative rate. Eliminating disparate mistreatment when it is defined in terms of false discovery rate and false omission rate presents significant additional challenges due to computational complexities involved and we leave it as a direction to be thoroughly explored in a future work.

Satisfying multiple fairness notions simultaneously. In certain application scenarios, it might be desirable to satisfy more than one notion of fairness defined above in Eqs. (1-7). In this paper, we consider scenarios where we attempt to avoid disparate treatment as well as disparate mistreatment measured as overall misclassification rate, false positive rate and false negative rate *simultaneously*, *i.e.*, satisfy Eqs. (1, 3-5).

Some recent works [7, 18] have investigated the impossibility of simultaneously satisfying multiple notions of fairness. Chouldechova [7] and Kleinberg et al. [18], show that, when the fraction of users with positive class labels differ between members of different sensitive attribute value groups, it is impossible to construct classifiers that are equally *well-calibrated* (where well-calibration essentially measures the false discovery and false omission rates of a classifier) and also satisfy the equal false positive and false negative rate criterion (except for a "dumb" classifier that assign all examples to a single class). These results suggest that satisfying all five criterion of disparate mistreatment (Table 1) simultaneously is impossible when the underlying distribution of data is different for different groups. However, in practice, it may still be interesting to explore the best, even if imperfect, extent of fairness a classifier can achieve. In the next section, we allow for bounded imperfections in our new fairness notions by allowing the left- and right-sides of Eqs. (3-5) to differ by no more than a threshold ϵ.

4. CLASSIFIERS WITHOUT DISPARATE MISTREATMENT

In this section, we describe how to train decision boundary-based classifiers (*e.g.*, logistic regression, SVMs) that do not suffer from disparate mistreatment. These classifiers generally learn the optimal decision boundary by minimizing a convex loss $L(\boldsymbol{\theta})$. The convexity of $L(\boldsymbol{\theta})$ ensures that a global optimum can be found *efficiently*. In order to ensure that the learned boundary is fair—it does not suffer from disparate mistreatment—one could incorporate the appropriate condition from Eqs. (3-5) (based on which kind of misclassifications disparate mistreatment is being defined for) into the classifier formulation. For example:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\ & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon, \end{array} \tag{8}$$

where $\epsilon \in \mathbb{R}^+$ and the smaller ϵ is, the more fair the decision boundary would be. The above formulation ensures that the classifier chooses the optimal decision boundary *within* the space of fair boundaries specified by the constraints. However, since the conditions in Eqs. (3-5) are, in general, non convex, solving the constrained optimization problem defined by (8) seems difficult.

To overcome the above difficulty, we propose a tractable proxy, inspired by the disparate impact proxy proposed by Zafar et al. [28]. In particular, we propose to measure disparate mistreatment using the covariance between the users' sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary, *i.e.*:

$$\begin{aligned} \text{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))] \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}), \end{aligned} \tag{9}$$

where the term $\mathbb{E}[(z - \bar{z})]\bar{g}_{\boldsymbol{\theta}}(\mathbf{x})$ cancels out since $\mathbb{E}[(z - \bar{z})] = 0$ and the function $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ is defined as:

$$g_{\boldsymbol{\theta}}(y, \mathbf{x}) = min(0, y d_{\boldsymbol{\theta}}(\mathbf{x})), \tag{10}$$

$$g_{\boldsymbol{\theta}}(y, \mathbf{x}) = min\left(0, \frac{1 - y}{2} y d_{\boldsymbol{\theta}}(\mathbf{x})\right), \text{ or} \tag{11}$$

$$g_{\boldsymbol{\theta}}(y, \mathbf{x}) = min\left(0, \frac{1 + y}{2} y d_{\boldsymbol{\theta}}(\mathbf{x})\right), \tag{12}$$

which approximates, respectively, the conditions in Eqs. (3-5). Note that, if a decision boundary satisfies Eqs. (3-5), the covariance defined above for that boundary will be close to zero, *i.e.*, $\text{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) \approx 0$. Moreover, in linear models for classification, such as logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\boldsymbol{\theta}^T \mathbf{x} = 0$, therefore, $d_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$.

Given the above proxy, one can rewrite (8) as:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \leq c, \\ & \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \geq -c, \end{array} \tag{13}$$

where the covariance threshold $c \in \mathbb{R}^+$ controls how *adherent* to disparate mistreatment the boundary should be.[3]

Solving the problem efficiently. While the constraints proposed in (13) can be an effective proxy for fairness, they are still non-convex, making it challenging to efficiently solve the optimization problem in (13). Next, we will convert these constraints into a Disciplined Convex-Concave Program (DCCP), which can be solved efficiently by leveraging recent advances in convex-concave programming [26].

First, consider the constraint described in (13), *i.e.*,

$$\sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \sim c, \tag{14}$$

where $\sim$ may denote '$\geq$' or '$\leq$'. Also, we drop the constant number $\frac{1}{N}$ for the sake of simplicity. Since the sensitive feature z is binary, *i.e.*, $z \in \{0, 1\}$, we can split the sum in the above expression into two terms:

$$\sum_{(\mathbf{x}, y) \in \mathcal{D}_0} (0 - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) + \sum_{(\mathbf{x}, y) \in \mathcal{D}_1} (1 - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \sim c, \tag{15}$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ are the subsets of the training dataset $\mathcal{D}$ taking values $z = 0$ and $z = 1$, respectively. Define $N_0 = |\mathcal{D}_0|$ and $N_1 = |\mathcal{D}_1|$, then one can write $\bar{z} = \frac{(0 \times N_0) + (1 \times N_1)}{N} = \frac{N_1}{N}$ and rewrite (15) as:

$$\frac{-N_1}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_0} g_{\boldsymbol{\theta}}(y, \mathbf{x}) + \frac{N_0}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_1} g_{\boldsymbol{\theta}}(y, \mathbf{x}) \sim c, \tag{16}$$

[3] Note that if one wants to have *both* equal false positive and equal false negative rates, one can apply separate constraints with $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ defined in both (11) and (12).

which, given that $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ is convex in $\boldsymbol{\theta}$, results into a convex-concave (or, difference of convex) function.

Finally, we can rewrite the problem defined by (13) as:

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \\ & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c, \end{aligned} \tag{17}$$

which is a Disciplined Convex-Concave Program (DCCP) for any convex loss $L(\boldsymbol{\theta})$, and can be efficiently solved using well-known heuristics such as the one proposed by Shen et al. [26]. Next, we particularize the formulation given by (17) for a logistic regression classifier [6].[4]

Logistic regression without disparate mistreatment. In logistic regression, the optimal decision boundary $\boldsymbol{\theta}^*$ can be found by solving a maximum likelihood problem of the form $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ in the training phase. Hence, a fair logistic regressor can be trained by solving the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & - \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \\ & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c. \end{aligned} \tag{18}$$

Simultaneously removing disparate treatment. Note that the above formulation for removing disparate mistreatment provides the flexibility to remove disparate treatment as well. That is, since our formulation does not require the sensitive attribute information at decision time, by keeping the features $\mathbf{x}$ disjoint from sensitive attribute z, one can remove disparate mistreatment and disparate treatment simultaneously.

5. EVALUATION

In this section, we conduct experiments on synthetic as well as real world datasets to evaluate the effectiveness of our scheme in controlling disparate mistreatment. To this end, we first generate several *synthetic* datasets that illustrate different variations of disparate mistreatment and show that our method can effectively remove disparate mistreatment in each of the variations, often at a small cost on accuracy. Then, we conduct experiments on the ProPublica COMPAS dataset [19] to show the effectiveness of our method on a *real world* dataset. In both the synthetic and real-world datasets, we compare the performance of our scheme with a baseline algorithm and a recently proposed method [15].

All of our experiments are conducted using logistic regression classifiers. To ensure the robustness of the experimental findings, for all of the datasets, we repeatedly (five times) split the data uniformly at random into train (50%) and test (50%) sets and report the average statistics for accuracy and fairness.

Evaluation metrics. In this evaluation, we consider that one wants to remove disparate mistreatment when it is measured in terms of false positive rate and false negative rate (Eqs. (4) and (5)). Specifically, We quantify the disparate mistreatment incurred by a classifier as:

$$D_{FPR} = P(\hat{y} \neq y | z = 0, y = -1) - P(\hat{y} \neq y | z = 1, y = -1),$$
$$D_{FNR} = P(\hat{y} \neq y | z = 0, y = 1) - P(\hat{y} \neq y | z = 1, y = 1),$$

where the closer the values of D_{FPR} and D_{FNR} to 0, the lower the degree of disparate mistreatment.

5.1 Experiments on synthetic data

In this section, we empirically study the trade-off between fairness and accuracy in a classifier that suffers from disparate mistreatment. To this end, we first start with a simple scenario in which the classifier is unfair in terms of *only* false positive rate *or* false negative rate. Then, we focus on a more complex scenario in which the classifier is unfair in terms of *both*.

5.1.1 *Disparate mistreatment on* only *false positive rate* or *false negative rate*

The first scenario considers a case where a classifier trained on the ground truth data leads to disparate mistreatment in terms of only the false positive rate (false negative rate), while being fair with respect to false negative rate (false positive rate), *i.e.*, $D_{FPR} \neq 0$ and $D_{FNR} = 0$ (or, alternatively, $D_{FPR} = 0$ and $D_{FNR} \neq 0$).

Experimental setup. We first generate 10,000 binary class labels ($y \in \{-1, 1\}$) and corresponding sensitive attribute values ($z \in \{0, 1\}$), both uniformly at random, and assign a two-dimensional user feature vector ($\mathbf{x}$) to each of the points. To ensure different distributions for negative classes of the two sensitive attribute value groups (so that the two groups have different false positive rates), the user feature vectors are sampled from the following distributions (we sample 2500 points from each distribution):

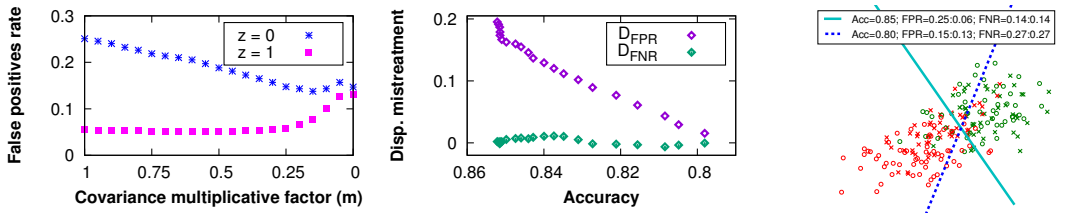
$$\begin{aligned} p(\mathbf{x}|z=0, y=1) &= \mathcal{N}([2,2],[3,1;1,3]) \\ p(\mathbf{x}|z=1, y=1) &= \mathcal{N}([2,2],[3,1;1,3]) \\ p(\mathbf{x}|z=0, y=-1) &= \mathcal{N}([1,1],[3,3;1,3]) \\ p(\mathbf{x}|z=1, y=-1) &= \mathcal{N}([-2,-2],[3,1;1,3]). \end{aligned}$$

Next, we train a (unconstrained) logistic regression classifier on this data. The classifier is able to achieve an accuracy of 0.85. However, due to difference in feature distributions for the two sensitive attribute value groups, it achieves $D_{FNR} = 0.14 - 0.14 = 0$ and $D_{FPR} = 0.25 - 0.06 = 0.19$, which constitutes a clear case of disparate mistreatment in terms of false positive rate.

We then train several logistic regression classifiers on the same training data subject to fairness constraints on false positive rate, *i.e.*, we train a logistic regressor by solving problem (18), where $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ is given by Eq. (11). Each classifier constrains the false positive rate covariance (c) with a multiplicative factor ($m \in [0, 1]$) of the covariance of the unconstrained classifier (c^*), that is, $c = mc^*$. Ideally, a smaller m, and hence a smaller c, would result in more fair outcomes.

Results. Figure 2 summarizes the results for this scenario by showing (a) the relation between decision-boundary covariance and the false positive rates for both sensitive attribute values; (b) the trade-off between accuracy and fairness; and (c) the decision boundaries for both the unconstrained classifier (solid) and the fair constrained classifier

[4] Our fairness constraints can be easily incorporated to other boundary-based classifiers such as (non)linear SVMs.



(a) Cov. vs FPR (b) Fairness vs. Acc. (c) Boundaries

Figure 2: [Synthetic data] Panel (a) shows that decreasing the covariance threshold causes the false positive rates for both groups to become similar. Panel (b) shows that an increasing degree of fairness corresponds to a steady decrease in accuracy. Panel (c) shows the original decision boundary (solid line) and fair decision boundary (dashed line), along with corresponding accuracy and false positive rates for groups $z = 0$ (crosses) and $z = 1$ (circles). Fairness constraints cause the original decision boundary to rotate such that previously misclassified examples with $z = 0$ are moved into the negative class (decreasing false positives), while well-classified examples with $z = 1$ are moved into the positive class (increasing false positives), leading to equal false positive rates for both groups.

(dashed). In this figure, we observe that: i) as the fairness constraint value $c = mc^*$ goes to zero, the false positive rates for both groups ($z = 0$ and $z = 1$) converge, and hence, the outcomes of the classifier become more fair, *i.e.*, $D_{FPR} \to 0$, while D_{FNR} remains close to zero (the invariance of D_{FNR} may however change depending on the underlying distribution of the data); ii) ensuring lower values of disparate mistreatment leads to a larger drop in accuracy.

5.1.2 *Disparate mistreatment on* **both** *false positive rate and false negative rate*

In this section, we consider a more complex scenario, where the outcomes of the classifier suffer from disparate mistreatment with respect to *both* false positive rate and false negative rate, *i.e.*, both D_{FPR} and D_{FNR} are non-zero. This scenario can in turn be split into two cases:

I. D_{FPR} and D_{FNR} have *opposite signs*, *i.e.*, the decision boundary disproportionately *favors* subjects from a certain sensitive attribute value group to be in the positive class (even when such assignments are misclassifications) while disproportionately assigning the subjects from the other group to the negative class. As a result, false positive rate for one group is higher than the other, while the false negative rate for the same group is lower.

II. D_{FPR} and D_{FNR} have the *same sign*, *i.e.*, both false positive as well as false negative rate are higher for a certain sensitive attribute value group. These cases might arise in scenarios when a certain group is harder to classify than the other.

Next, we experiment with each of the above cases separately.

— **Case I:** To simulate this scenario, we first generate 2,500 samples from each of the following distributions:

$$p(\mathbf{x}|z=0,y=1) = \mathcal{N}([2,0],[5,1;1,5])$$
$$p(\mathbf{x}|z=1,y=1) = \mathcal{N}([2,3],[5,1;1,5])$$
$$p(\mathbf{x}|z=0,y=-1) = \mathcal{N}([-1,-3],[5,1;1,5])$$
$$p(\mathbf{x}|z=1,y=-1) = \mathcal{N}([-1,0],[5,1;1,5])$$

An unconstrained logistic regression classifier on this dataset attains an overall accuracy of 0.78 but leads to a false positive rate of 0.14 and 0.30 (*i.e.*, $D_{FPR} = 0.14 - 0.30 = -0.16$) for the sensitive attribute groups $z = 0$ and $z = 1$, respectively; and false negative rates of 0.31 and 0.12 (*i.e.*, $D_{FNR} = 0.31 - 0.12 = 0.19$). Finally, we train three different fair classifiers, with fairness constraints on (i) false positive rates—$g_{\boldsymbol{\theta}}(y, \mathbf{x})$ given by Eq. (11), (ii) false negative rates—$g_{\boldsymbol{\theta}}(y, \mathbf{x})$ given by Eq. (12) and (iii) on both false positive and false negative rates—separate constraints for $g_{\boldsymbol{\theta}}(y, \mathbf{x})$ given by Eq. (11) and Eq. (12).

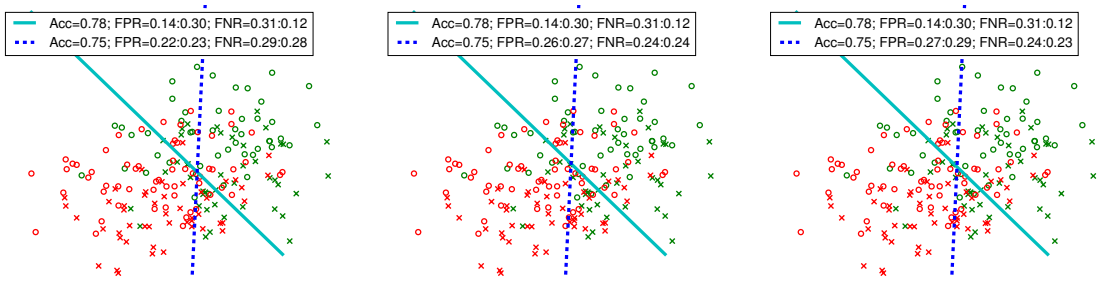
Results. Figure 3 summarizes the results for this scenario by showing the decision boundaries for the unconstrained classifier (solid) and the constrained fair classifiers. Here, we can observe several interesting patterns. First, removing disparate mistreatment on only false positive rate causes a rotation in the decision boundary to move previously *misclassified* examples with $z = 1$ into the negative class, *decreasing* their false positive rate. However, in the process, it also moves previously *well-classified* examples with $z = 1$ into the negative class, *increasing* their false negative rate. As a consequence, controlling disparate mistreatment on false positive rate (Figure 3(a)), also removes disparate mistreatment on false negative rate. A similar effect occurs when we control disparate mistreatment only with respect to the false negative rate (Figure 3(b)), and therefore, provides similar results as the constrained classifier for both false positive and false negative rates (Figure 3(c)). This effect is explained by the distribution of the data, where the centroids of the clusters for the group with $z = 0$ are shifted with respect to the ones for the group $z = 1$.

— **Case II:** To simulate the scenario where both D_{FPR} and D_{FNR} have the same sign, we generate 2,500 samples from each of the following distributions:

$$p(\mathbf{x}|z=0,y=1) = \mathcal{N}([1,2],[5,2;2,5])$$
$$p(\mathbf{x}|z=1,y=1) = \mathcal{N}([2,3],[10,1;1,4])$$
$$p(\mathbf{x}|z=0,y=-1) = \mathcal{N}([0,-1],[7,1;1,7])$$
$$p(\mathbf{x}|z=1,y=-1) = \mathcal{N}([-5,0],[5,1;1,5])$$

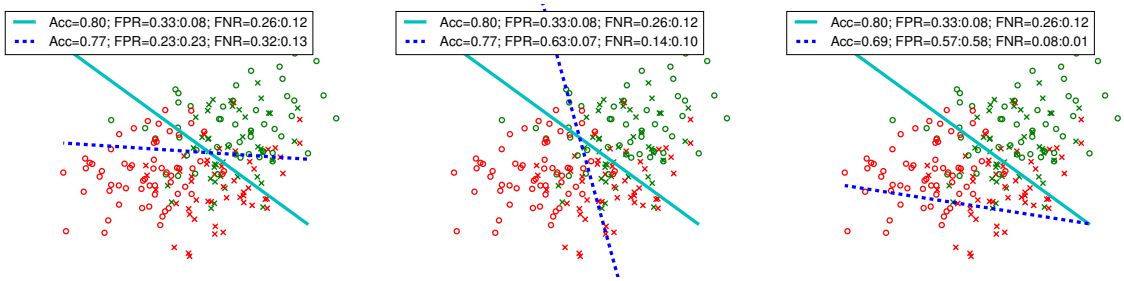
Then, we train an unconstrained logistic regression classifier on this dataset. It attains an accuracy of 0.80 but leads to $D_{FPR} = 0.33 - 0.08 = 0.25$ and $D_{FNR} = 0.26 - 0.12 = 0.14$, resulting in disparate mistreatment in terms of both false positive and negative rates. Then, similarly to the previous scenario, we train three different kind of constrained classifiers to remove disparate mistreatment on (i) false positive rate, (ii) false negatives rate, and (iii) both.

Results. Figure 4 summarizes the results by showing the decision boundaries for both the unconstrained classifiers (solid) and the fair constrained classifier (dashed) when controlling for disparate mistreatment with respect to false positive rate, false negative rate and both, respectively. We



(a) FPR constraints (b) FNR constraints (c) Both constraints

Figure 3: [Synthetic data] D_{FPR} and D_{FNR} have opposite signs. Removing disparate mistreatment on FPR can potentially help remove disparate mistreatment on FNR. Removing disparate mistreatment on both at the same time leads to very similar results.



(a) FPR constraints (b) FNR constraints (c) Both constraints

Figure 4: [Synthetic data] D_{FPR} and D_{FNR} have the same sign. Removing disparate mistreatment on FPR can potentially increase disparate mistreatment on FNR. Removing disparate mistreatment on both at the same time causes a larger drop in accuracy.

observe several interesting patterns. First, controlling disparate mistreatment for only false positive rate (false negative rate), leads to a minor drop in accuracy, but can exacerbate the disparate mistreatment on false negative rate (false positive rate). For example, while the decision boundary is moved to control for disparate mistreatment on false negative rate, that is, to ensure that more examples with $z = 0$ are well-classified in the positive class (reducing false negative rate), it also moves previously well-classified negative examples into the positive class, hence increasing the false positive rate. A similar phenomenon occur when controlling disparate mistreatment with respect to only false positive rate. As a consequence, controlling for both types of disparate mistreatment simultaneously brings D_{FPR} and D_{FNR} close to zero, but causes a large drop in accuracy.

5.1.3 Performance Comparison

In this section, we compare the performance of our scheme with two different methods on the synthetic datasets described above. In particular, we compare the performance of the following approaches:

***Our method*:** implements our scheme to avoid disparate treatment and disparate mistreatment *simultaneously*. Disparate mistreatment is avoided by using fairness constraints (as described in Sections 5.1.1 and 5.1.2). Disparate treatment is avoided by ensuring that sensitive attribute information is not used while making decisions, *i.e.*, by keeping user feature vectors ($\mathbf{x}$) and the sensitive features (z) disjoint. All the explanatory simulations on synthetic data shown earlier (Sections 5.1.1 and 5.1.2) implement this scheme.

***Our method*$_{sen}$:** implements our scheme to avoid disparate mistreatment only. The user feature vectors ($\mathbf{x}$) and the sensitive features (z) are not disjoint, that is, z is used as a learnable feature. Therefore, the sensitive attribute information is used for decision making, resulting in disparate treatment.

Hardt et al. [15]: operates by post-processing the outcomes of an unfair classifier (logistic regression in this case) and using different decision thresholds for different sensitive attribute value groups to achieve fairness. By construction, it needs the sensitive attribute information while making decisions, and hence cannot avoid disparate treatment.

***Baseline*:** tries to remove disparate mistreatment by introducing different penalties for misclassified data points with different sensitive attribute values during training phase. Specifically, it proceeds in two steps. First, it trains an (unfair) classifier minimizing a loss function (*e.g.*, logistic loss) over the training data. Next, it selects the set of misclassified data points from the sensitive attribute value group that presents the higher error rate. For example, if one wants to remove disparate mistreatment with respect to false positive rate and $D_{FPR} > 0$ (which means the false positive rate for points with $z = 0$ is higher than that of $z = 1$), it selects the set of misclassified data points in the training set having $z = 0$ and $y = -1$. Next, it iteratively re-trains the classifier with increasingly higher penalties on this set of data points until a certain fairness level is achieved in the training set (until $D_{FPR} \leq \epsilon$). The algorithm is summarized in Figure 5, particularized to ensure fairness in terms of false positive rate. This process can be intuitively ex-

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^N$, $\Delta > 0$ $\epsilon > 0$
Output: Fair baseline decision boundary $\boldsymbol{\theta}$
Initialize: Penalty $C = 1$
Train (unfair) classifier $\boldsymbol{\theta} = \text{argmin}_{\boldsymbol{\theta}} \sum_{\mathbf{d} \in \mathcal{D}} L(\boldsymbol{\theta}, \mathbf{d})$
Compute $\hat{y}_i = \text{sign}(d_{\boldsymbol{\theta}}(\mathbf{x}_i))$ and D_{FP} on $\mathcal{D}$.
if $D_{FP} > 0$ **then** $s = 0$
else $s = 1$
$\mathcal{P} = \{\mathbf{x}_i, y_i, z_i | \hat{y} \neq y_i, z_i = s\}$, $\bar{\mathcal{P}} = \mathcal{D} \setminus \mathcal{P}$.
while $D_{FP} > \epsilon$ **do**
  Increase penalty: $C = C + \Delta$.
  $\boldsymbol{\theta} = \text{argmin}_{\boldsymbol{\theta}} C \sum_{\mathbf{d} \in \mathcal{P}} L(\boldsymbol{\theta}, \mathbf{d}) + \sum_{\mathbf{d} \in \bar{\mathcal{P}}} L(\boldsymbol{\theta}, \mathbf{d})$
end

Figure 5: Baseline method for removing disparate mistreatment on false positive rates.

tended to account for fairness in terms of false negative rate or for *both* false positive rate and false negative rate. Like ***Our method***, the baseline does not use sensitive attribute information while making decisions.

Comparison results. Table 2 shows the performance comparison for all the methods on the three synthetic datasets described above. We can observe that, while all four methods mostly achieve similar levels of fairness, they do it at different costs in terms of accuracy. Both ***Our method$_{sen}$*** and ***Hardt et al.***—which use sensitive feature information while making decisions—present the best performance in terms of accuracy (due to the additional information available to them). However, as explained earlier, these two methods suffer from disparate treatment. On the other hand, the implementation of our scheme to simultaneously remove disparate mistreatment and disparate treatment (***Our method***) does so with further accuracy drop of only $\sim$5% with respect to the above two methods that cause disparate treatment. Finally, the ***baseline*** is sometimes unable to achieve fairness. When it does achieves fairness, it does so at a (sometimes much) greater cost in accuracy in comparison with the competing methods.

In summary, our method achieves the same performance as *Hardt et al.* when making use of the same information in the data, *i.e.*, non-sensitive as well as sensitive features. However, in contrast to *Hardt et al.*, it also allows us to simultaneously remove both disparate mistreatment and disparate treatment at a small additional cost in terms of accuracy.

5.2 Real world dataset: ProPublica COMPAS

In this section, we experiment with the COMPAS risk assessment dataset compiled by ProPublica [19] and show that our method can significantly reduce disparate mistreatment at a modest cost in terms of accuracy.

Dataset and experimental setup. ProPublica compiled a list of all criminal offenders screened through the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool[5] in Broward County, Florida during 2013-2014. The data includes information on the offenders' demographic features (gender, race, age), criminal history (charge for which the person was arrested, number of prior offenses) and the risk score assigned to the offender by COMPAS. ProPublica also collected the *ground truth* on whether or not these individuals actually recidivated within two years after the screening. For more information about the data collection, we point the reader to a detailed description [20]. Some of the follow-up discussion on this dataset can be found at [3, 11].

In this analysis, for simplicity, we only consider a subset of offenders whose race was either black or white. Recidivism rates for the two groups are shown in Table 3.

Using this ground truth, we build an unconstrained logistic regression classifier to predict whether an offender will (positive class) or will not (negative class) recidivate within two years. The set of features used in the classification task are described in Table 4.[6,7]

The (unconstrained) logistic regression classifier leads to an accuracy of 0.668. However, the classifier yields false positive rates of 0.35 and 0.17, respectively, for blacks and whites (*i.e.*, $D_{FPR} = 0.18$), and false negative rates of 0.31 and 0.61 (*i.e.*, $D_{FNR} = -0.30$). These results constitute a clear case of disparate mistreatment in terms of both false positive rate and false negative rate. The classifier puts one group (blacks) at relative disadvantage by disproportionately misclassifying negative (did not recidivate) examples from this group into positive (did recidivate) class. This disproportional assignment results in a significantly higher false positive rate for blacks as compared to whites. On the other hand, the classifier puts the other group (whites) on a relative advantage by disproportionately misclassifying positive (did recidivate) examples from this group into negative (did not recidivate) class (resulting in a higher false negative rate). Note that this scenario resembles our synthetic example Case I in Section 5.1.2.

Finally, we train logistic regression classifiers with three types of constraints: constraints on false positive rate, false negative rate, and on both.

Results. Table 2 (last block) summarizes the results by showing the trade-off between fairness and accuracy achieved by our method, the method by Hardt et al., and the baseline. Similarly to the results in Section 5.1.2, we observe that for all three mehtods, controlling for disparate mistreatment on false positive rate (false negative rate) also helps decrease disparate mistreatment on false negative rate (false positive rate). Moreover, all three methods are able to achieve similar accuracy for a given level of fairness.

Additionally, we observe that our method (as well as the baseline) does not completely remove disparate mistreatment, *i.e.*, it does not achieve zero D_{FPR} or/and D_{FNR} in any of the cases. This is probably due to the relatively small size of the dataset[8] (and hence a smaller ratio between number of training examples and number of learnable features), which hinders a robust estimate of misclassification covariance (Eqs. 11 and 12). This highlights the fact that our method can suffer from reduced performance on small datasets. In scenarios with sufficiently large training datasets, we expect more reliable estimates of covariance,

[5] COMPAS tries to predict the recidivism risk (on a scale of 1–10) of a criminal offender by analyzing answers to 137 questions pertaining to the offender's criminal history and behavioral patterns [2].

[6] Notice that goal of this section is not to analyze the best set of features for recidivism prediction, rather, we focus on showing that our method can effectively remove disparate mistreatment in a given dataset. Hence, we chose to use the same set of features as used by ProPublica for their analysis.

[7] Since race is one of the features in the learnable set, we additionally assume that *all* the methods have access to the sensitive attributes while making decisions.

[8] 2,639 examples in the training set.

		FPR constraints			FNR constraints			Both constraints		
		Acc.	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**
Synthetic setting 1 (Figure 2)	**Our method**	0.80	0.02	0.00	–	–	–	–	–	–
	Our method$_{sen}$	0.85	0.00	0.25	–	–	–	0.83	0.07	0.01
	Baseline	0.65	0.00	0.00	–	–	–	–	–	–
	Hardt et al.	0.85	0.00	0.21	–	–	–	0.80	0.00	0.02
Synthetic setting 2 (Figure 3)	**Our method**	0.75	−0.01	0.01	0.75	−0.01	0.01	0.75	−0.01	0.01
	Our method$_{sen}$	0.80	0.00	0.03	0.80	0.02	0.01	0.80	0.01	0.02
	Baseline	0.59	−0.01	0.15	0.59	−0.15	0.01	0.76	−0.04	0.03
	Hardt et al.	0.80	0.00	0.03	0.80	0.03	0.00	0.79	0.00	−0.01
Synthetic setting 3 (Figure 4)	**Our method**	0.77	0.00	0.19	0.77	0.55	0.04	0.69	−0.01	0.06
	Our method$_{sen}$	0.78	0.00	0.42	0.79	0.38	0.03	0.77	0.14	0.06
	Baseline	0.57	0.01	0.09	0.67	0.44	0.01	0.38	−0.43	0.01
	Hardt et al.	0.78	0.01	0.44	0.79	0.41	0.02	0.67	0.02	0.00
ProPuclica COMPAS (Section 5.2)	**Our method$_{sen}$**	0.660	0.06	−0.14	0.662	0.03	−0.10	0.661	0.03	−0.11
	Baseline	0.643	0.03	−0.11	0.660	0.00	−0.07	0.660	0.01	−0.09
	Hardt et al.	0.659	0.02	−0.08	0.653	−0.06	−0.01	0.645	−0.01	−0.01

Table 2: Performance of different methods while removing disparate mistreatment with respect to false positive rate, false negative rate and both.

Race	Yes	No	Total
Black	1,661(52%)	1,514(48%)	3,175(100%)
White	8,22(39%)	1,281(61%)	2,103(100%)
Total	2,483(47%)	2,795(53%)	5,278(100%)

Table 3: Recidivism rates in ProPublica COMPAS data for both races.

Feature	Description
Age Category	< 25, between 25 and 45, > 45
Gender	Male or Female
Race	White or Black
Priors Count	0–37
Charge Degree	Misconduct or Felony
2-year-rec. (target feature)	Whether (+ve) or not (-ve) the defendant recidivated within two years

Table 4: Description of features used from ProPublica COMPAS data.

and hence, a better performance from our method. On the other hand, the method by Hardt et al. is able to achieve both zero D_{FPR} and D_{FNR} while controlling for disparate mistreatment on both false positive and false negative rates (Table 2)—albeit at a considerable drop in terms of accuracy. Since this method operates on a data of much smaller dimensionality (the final classifier probability estimates), it is not expected to suffer as much from the small size of the dataset as compared to our method or the baseline (which depend on the misclassification covariance computed on the feature set).

6. DISCUSSION AND FUTURE WORK

As shown in Section 5, the method proposed in this paper provides a flexible tradeoff between disparate mistreatment-based fairness and accuracy. It also allows to avoid disparate mistreatment and disparate treatment *simultaneously*. This feature might be specially useful in scenarios when the sensitive attribute information is not available (*e.g.*, due to privacy reasons) or is prohibited from being used due to disparate treatment laws [5].

Although we proposed fair classifier formulations to remove disparate mistreatment only on false positive and false negative rates, as described in Section 3, disparate mistreatment can also be measured with respect to false discovery and false omission rates. Extending our current formulation to include false discovery and false omission rates is a non-trivial task due to computational complexities involved. A natural extension of this work would be to include these other measures of disparate mistreatment into our fair classifier formulation.

Finally, we would like to point out that the current formulation of fairness constraints may suffer from the following limitations. Firstly, the proposed formulation to train fair classifiers is not a convex program, but a disciplined convex-concave program (DCCP), which can be efficiently solved using heuristic-based methods [26]. While these methods are shown to work well in practice, unlike convex optimization, they do not provide any guarantees on the global optimality of the solution. Secondly, since computing the analytical covariance in fairness constraints is not a trivial task, we approximate it through Monte Carlo covariance on the training set (Eq. 9). While this approximation is expected to work well when a reasonable amount of training data is provided, it might be inaccurate for smaller datasets.

7. REFERENCES

[1] Stop-and-frisk in New York City. https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City.

[2] https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html, 2016.

[3] J. Angwin and J. Larson. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[5] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] A. Chouldechova. Fair Prediction with Disparate Impact:A Study of Bias in Recidivism Prediction Instruments. *arXiv preprint, arXiv:1610.07524*, 2016.

[8] K. Crawford. Artificial Intelligence's White Guy Problem. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

[9] C. Dwork, M. Hardt, T. Pitassi, and O. Reingold. Fairness Through Awareness. In *ITCSC*, 2012.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and Removing Disparate Impact. In *KDD*, 2015.

[11] A. W. Flores, C. T. Lowenkamp, and K. Bechtel. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.". 2016.

[12] S. Goel, J. M. Rao, and R. Shroff. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*, 2015.

[13] G. Goh, A. Cotter, M. Gupta, and M. Friedlander. Satisfying Real-world Goals with Dataset Constraints. In *NIPS*, 2016.

[14] J. M. Greg Ridgeway. Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops. *Journal of the American Statistical Association*, 2009.

[15] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, 2016.

[16] F. Kamiran and T. Calders. Classification with No Discrimination by Preferential Sampling. In *BENELEARN*, 2010.

[17] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, 2017.

[19] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. https://github.com/propublica/compas-analysis, 2016.

[20] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

[21] B. T. Luong, S. Ruggieri, and F. Turini. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*, 2011.

[22] C. Muñoz, M. Smith, and D. Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*, 2016.

[23] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware Data Mining. In *KDD*, 2008.

[24] J. Podesta, P. Pritzker, E. Moniz, J. Holdren, and J. Zients. Big Data: Seizing Opportunities, Preserving Values. *Executive Office of the President. The White House.*, 2014.

[25] A. Romei and S. Ruggieri. A Multidisciplinary Survey on Discrimination Analysis. *KER*, 2014.

[26] X. Shen, S. Diamond, Y. Gu, and S. Boyd. Disciplined Convex-Concave Programming. *arXiv:1604.02639*, 2016.

[27] L. Sweeney. Discrimination in Online Ad Delivery. *ACM Queue*, 2013.

[28] M. B. Zafar, I. V. Martinez, M. G. Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017.

[29] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In *ICML*, 2013.