Journal of Machine Learning Research 20 (2019) 1-42 Submitted 5/18; Revised 12/18; Published 3/19

Fairness Constraints: A Flexible Approach for Fair Classification

Muhammad Bilal Zafar MZAFAR@MPI-SWS.ORG
Bosch Center for Artificial Intelligence and Max Planck Institute for Software Systems
Saarbrücken, Germany

Isabel Valera ISABEL.VALERA@TUE.MPG.DE
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Manuel Gomez-Rodriguez MANUELGR@MPI-SWS.ORG
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Krishna P. Gummadi GUMMADI@MPI-SWS.ORG
Max Planck Institute for Software Systems
Saarbrücken, Germany

Editor: Moritz Hardt

Abstract

Algorithmic decision making is employed in an increasing number of real-world applications to aid human decision making. While it has shown considerable promise in terms of improved decision accuracy, in some scenarios, its outcomes have been also shown to impose an unfair (dis)advantage on people from certain social groups (*e.g.*, women, blacks). In this context, there is a need for computational techniques to limit unfairness in algorithmic decision making. In this work, we take a step forward to fulfill that need and introduce a flexible constraint-based framework to enable the design of fair margin-based classifiers. The main technical innovation of our framework is a general and intuitive measure of decision boundary unfairness, which serves as a tractable proxy to several of the most popular computational definitions of unfairness from the literature. Leveraging our measure, we can reduce the design of fair margin-based classifiers to adding tractable constraints on their decision boundaries. Experiments on multiple synthetic and real-world datasets show that our framework is able to successfully limit unfairness, often at a small cost in terms of accuracy.

Keywords: Supervised learning, margin-based classifiers, fairness, discrimination, disparate impact.

1. Introduction

Algorithmic decision making systems are assisting, or in some cases even replacing, human decision making in an increasing number of application domains. Examples of such domains include credit approval (FICO, 2017), criminal risk assessment (Perry, 2013), hir-

. A preliminary version of this work appeared in Zafar et al. (2017a,b).
. Open-source code implementation is available at: http://fate-computing.mpi-sws.org/.

ing (Posse, 2016; Taylor, 2016) and education (Romero and Ventura, 2011). By leveraging vast amounts of (training) data and modern machine learning methods, algorithmic decision systems are able to provide highly accurate predictions, often surpassing even human accuracy (Kleinberg et al., 2017a; Liu et al., 2017). However, recent studies by civil organizations (Bhandari, 2016), governments (Podesta et al., 2014; Muñoz et al., 2016), and researchers (Sweeney, 2013) alike have raised concerns about the unfairness of these algorithmic decision systems towards people from certain social groups (*e.g.*, women, blacks). Importantly, in various countries, these concerns are often grounded on anti-discrimination laws which prohibit unfair treatment of people based on one or more *sensitive features*, such as gender and race (Civil Rights Act, 1964; Barocas and Selbst, 2016).

To overcome the above concerns, a number of recent studies in the emerging field of ethical machine learning have proposed mechanisms to ensure that algorithmic decision systems do not lead to unfair outcomes (Kamiran and Calders, 2009; Corbett-Davies et al., 2017; Hardt et al., 2016; Feldman et al., 2015; Zafar et al., 2017b,a; Zemel et al., 2013; Dwork et al., 2012). In doing so, they have have measured the (un)fairness of a decision making process through the distribution of its outcomes among people from different sensitive feature groups (*e.g.*, men, women). More specifically, they have typically adopted one of the following three notions of (un)fairness: [1]

— ***Disparate treatment***: a decision making process suffers from disparate treatment if its outcomes change based on a change in the sensitive feature value with all other features being the same.

— ***Disparate impact***: a decision making process suffers from disparate impact if it grants disproportionately large fraction of beneficial (or positive classification) outcomes to certain sensitive feature groups (*e.g.*, men, women).

— ***Disparate mistreatment***: a decision making process suffers from disparate mistreatment if its accuracy (or error rate) is different for different sensitive feature groups.

However, the mechanisms proposed in prior studies typically lack flexibility with respect to one or more of the following aspects:

(i) They are specifically designed for only one of the above notions. As a consequence, they cannot accommodate more than one of them *simultaneously* (*e.g.*, disparate treatment and disparate impact).

(ii) They cannot ensure fairness with respect to multiple sensitive features *simultaneously* (*e.g.*, gender and race).

(iii) They are only limited to a narrow range of classification models (*e.g.*, logistic regression).

In this work, we propose a flexible framework to design a variety of fair classifiers that do not suffer from the above limitations. More specifically, for any convex boundary-based classifier, our framework defines an intuitive measure of decision boundary unfairness: the covariance between the sensitive features and the signed distance between the (non sensitive) feature vectors and the decision boundary of the classifier for a subset the subjects which depends on the fairness notion of interest. This measure can be readily incorporated into the classifier formulation in the form of convex of convex-concave constraints, one per sensitive feature or fairness notion, which can be efficiently solved using well-known meth-

1. As we later discuss in Section 2, the same notions are often referred to by different names by different studies.

ods (Boyd and Vandenberghe, 2004; Shen et al., 2016). Interestingly, our framework also allows for a dual formulation which maximizes fairness under accuracy constraints and, as a consequence, it ensures compliance with the business necessity clause of anti-discrimination doctrines (Barocas and Selbst, 2016), an aspect not considered by prior studies. Experiments on multiple synthetic and real-world datasets show that our framework is able to successfully limit disparate treatment, disparate impact and disparate mistreatment, often at a small cost in terms of accuracy, and it provides more flexibility than state-of-the-art methods (see Table 3).

The rest of the paper is organized as follows. First, we first revisit several well-known fairness notions from the literature and discuss the type of scenarios each notion is most suitable for (Section 2). Then, we formally state the problem of fairness-aware classification (Section 3) and describe our framework (Section 4). Later, we experiment with several datasets, including comparisons with related methodologies to highlight the effectiveness of our mechanism in controlling unfairness (Section 5). Finally, we conclude with a review of the related work on unfairness in algorithmic decision making and strategies proposed to mitigate unfairness (Section 6) as well as a discussion of future work (Section 7).

2. Background on different notions of (un)fairness

In this section, we revisit three of the most popular notions of fairness used in the machine learning literature: disparate treatment, disparate impact, and disparate mistreatment. More specifically, we first elaborate on each of these notions separately in the context of automated decision making systems and then highlight the differences between them.

— **Disparate treatment.** A decision making system suffers from disparate treatment if it provides different outputs for groups of people with the same (or similar) values of non-sensitive features but different values of sensitive features (Barocas and Selbst, 2016). In other words, (partly) basing the decision outcomes on the sensitive feature value amounts to disparate treatment. [2] This notion has been also referred to as *direct discrimination* (Pedreschi et al., 2008).

Figure 1 provides examples of binary classifiers with and without disparate treatment in a *stop-and-frisk* (Gelman et al., 2007) application. In all cases, the classifiers need to decide whether to stop a pedestrian on the suspicion of possessing an illegal weapon based on 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 deem classifiers $\mathbf{C_2}$ and $\mathbf{C_3}$ to be unfair due to disparate treatment since $\mathbf{C_2}$'s ($\mathbf{C_3}$'s) decisions for $Male$ 1 and $Female$ 1 ($Male$ 2 and $Female$ 2) are different even though they have the

2. Technically, the disparate treatment doctrine tries to counter *explicit* as well as *intentional* discrimination (Barocas and Selbst, 2016). It follows from the specification of disparate treatment that a decision maker with an intent to discriminate could try to disadvantage a group with a certain sensitive feature value (*e.g.*, a specific race group) not by *explicitly using the sensitive feature* itself, but by *intentionally basing decisions on a correlated feature* (*e.g.*, the non-sensitive feature location might be correlated with the sensitive feature race). This practice is often referred to as *redlining* in the US anti-discrimination law and also qualifies as disparate treatment (Gano, 2017). However, such hidden intentional disparate treatment maybe be hard to detect, and some authors argue that disparate impact might be a more suitable framework for detecting such covert discrimination (Siegel, 2014). Hence, in this paper, when discussing disparate treatment, we will focus only on *explicit* disparate treatment.

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 ($\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$) 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. $\mathbf{C_1}$ has no disparate treatment. It has disparate impact because it stops disproportional fractions of males and females (1.0 and 0.66, respectively). It has disparate mistreatment because it has different false negative rates (0.0 and 0.5, respectively) for males and females. $\mathbf{C_2}$ has disparate treatment since, of a male and a female having same non-sensitive attributes (Male 1 and Female 1), it stops only one of them. It has no disparate impact, since it stops equal fractions (0.66) of males and females. It has disparate mistreatment because of different false positive (0.0 and 1.0, respectively) and false negative rates (0.0 and 0.5, respectively) for males and females. $\mathbf{C_3}$ has disparate treatment and no disparate impact. It has no disparate mistreatment because both false positive rates (1.0) and false negative rates (0.5) for males and females are equal.

same values of non-sensitive features. On the other hand, classifier $\mathbf{C_1}$ does not suffer from disparate treatment since, with all feature values except gender being equal, its decisions are identical.

— **Disparate impact.** A decision making system suffers from disparate impact if it provides outputs that benefit (hurt) a group of people sharing a value of a sensitive feature more frequently than other groups of people (Barocas and Selbst, 2016). This notion has been also referred to as *statistical parity* (Corbett-Davies et al., 2017) or *demographic parity* (Dwork et al., 2012).

Similarly as with disparate treatment, Figure 1 provides examples of binary classifiers with and without disparate impact in a stop-and-frisk application. Under the assumption that a pedestrian benefits from a decision of not being stopped, we deem classifier $\mathbf{C_1}$ to be unfair due to disparate impact because the fraction of males and females that were stopped are different (1.0 and 0.66, respectively), where the latter benefit from a decision of not being stopped. On the other hand, classifiers $\mathbf{C_2}$ and $\mathbf{C_3}$ do not suffer from disparate impact because the fractions of males and females that were stopped are the same (0.66).

— **Disparate mistreatment.** A decision making system suffers from disparate mistreatment if it achieves different classification accuracy (or conversely, error rate) for groups of people sharing different values of a sensitive feature (Zafar et al., 2017a). This notion has been also referred to as *equality of opportunity* (Hardt et al., 2016) and *predictive equality* (Corbett-Davies et al., 2017). In addition to overall classification accuracy, this notion has been also particularized to *different types of misclassifications*, *i.e.*, false positives and

false negatives. In that context, a decision making system suffers from disparate mistreatment if individual misclassification rates (*e.g.*, false positive rate, false negative rate) are different for groups of people sharing different values of a sensitive feature.

In Figure 1, we deem classifiers $\mathbf{C_1}$ and $\mathbf{C_2}$ to be unfair due to disparate mistreatment since 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. Finally, classifier $\mathbf{C_3}$ does not suffer from disparate mistreatment because it has the same false negative and false positive rates for males and females.

Differences among (un)fairness notions. The above (un)fairness notions account for either *direct* (or *intentional*) and *indirect* (or *unintentional*) unfairness (Altman, 2016). More specifically, *disparate treatment accounts for direct unfairness*, *i.e.*, a situation where a decision making process directly (or intentionally) uses the sensitive feature information to put a group of people sharing a value of a sensitive feature on relative disadvantage. In this way, removing disparate treatment corresponds to a very intuitive notion of fairness: two otherwise similar persons should not be treated differently solely because of the difference in gender. On the other hand, *disparate impact and disparate mistreatment account for indirect unfairness*, *i.e.*, a situation where the decision making process can indirectly or unintentionally leverage the correlation between sensitive features and class labels to put a sensitive feature group at relative disadvantage (through low beneficial outcome rate under disparate impact and through high misclassification rate under disparate mistreatment).

Moreover, while disparate impact and disparate mistreatment both account for indirect unfairness, their application scenarios strongly differ. 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 (Barocas and Selbst, 2016). 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 (*e.g.*, *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 (Hardt et al., 2016; Zafar et al., 2017a).

3. Fairness in classification

In a binary classification task, [3] one aims to find a mapping function $f(\boldsymbol{x})$ between user feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. This task is achieved by utilizing a training set, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, to construct a mapping that works *well* on an *unseen* test set. For decision boundary-based classifiers, finding this mapping usually reduces to building a decision boundary in feature space that separates users in the training set according to their class labels. One typically looks for a decision boundary, defined by a set of parameters $\boldsymbol{\theta}^*$, that achieves the greatest classification accuracy in a test set, by minimizing a loss function over a training set $L(\boldsymbol{\theta})$, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. Then, given an *unseen* feature vector $\boldsymbol{x}_i$ from the test set, the classifier predicts the label $\hat{y}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i) \geq 0$ and $\hat{y}_i = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from the feature vector $\boldsymbol{x}$ to the decision boundary.

In the context of fairness in binary classification, each user also has an associated sensitive feature $z \in \{0, 1\}^4$ and the goal is finding a mapping (or decision boundary) that provides both accurate predictions and fairness guarantees. More formally, we can express the absence of disparate treatment, disparate impact and disparate mistreatment in a binary classifier as follows:

1. No disparate treatment. A binary classifier does not suffer from disparate treatment if the probability that the classifier outputs a specific value of $\hat{y}$ given a feature vector $\boldsymbol{x}$ does not change after observing the sensitive feature z, *i.e.*,

$$P(\hat{y}|\boldsymbol{x}, z) = P(\hat{y}|\boldsymbol{x}). \tag{3.1}$$

2. No disparate impact. A binary classifier does not suffer from disparate impact 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, *i.e.*,

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1). \tag{3.2}$$

Notice that neither disparate treatment nor disparate impact depend on the subjects' ground truth label (y).

3. No 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. [5] Consequently, the absence of disparate mistreatment in a binary classification task

3. For simplicity, we consider binary classification tasks in this work. However, our ideas can be easily extended to m-ary classification.
4. For ease of exposition, we assume z to be unidimensional and binary, however, our setup can be easily generalized to categorical as well as multiple sensitive features.
5. 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

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	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*.

can be specified with respect to the different misclassification measures as follows:

— ***Overall misclassification rate (OMR)***:

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1), \tag{3.3}$$

— ***False positive rate (FPR)***:

$$P(\hat{y} \neq y | y = -1, z = 0) = P(\hat{y} \neq y | y = -1, z = 1), \tag{3.4}$$

— ***False negative rate (FNR)***:

$$P(\hat{y} \neq y | y = 1, z = 0) = P(\hat{y} \neq y | y = 1, z = 1), \tag{3.5}$$

— ***False omission rate (FOR)***:

$$P(\hat{y} \neq y | \hat{y} = -1, z = 0) = P(\hat{y} \neq y | \hat{y} = -1, z = 1), \tag{3.6}$$

— ***False discovery rates (FDR)***:

$$P(\hat{y} \neq y | \hat{y} = 1, z = 0) = P(\hat{y} \neq y | \hat{y} = 1, z = 1). \tag{3.7}$$

Remarks. We would like to note that the formal criteria to avoid disparate impact and disparate mistreatment, as given by Eqs.(3.2–3.7), represent equality in certain *group-conditional probabilities* for the two sensitive feature groups, *i.e.*, $P(\cdot|z = 0)$ and $P(\cdot|z = 1)$. For example, the disparate impact criterion in Eq. (3.2) requires the group-conditional probabilities of positive class outcomes to be the same for users with $z = 0$ and $z = 1$, whereas the disparate mistreatment criterion in Eq. (3.3) requires the group-conditional

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.

probabilities of users being misclassified to be the same. In Section 4, we will show that these group-conditional equalities can be relaxed by various decision boundary covariances to facilitate efficient training of fair classifiers.

Finally, we would also like to highlight that some recent works have explored the cost of achieving fairness in a decision making process and pointed out some inherent tensions between fairness and accuracy (Chouldechova, 2016; Corbett-Davies et al., 2017; Kleinberg et al., 2017b; Menon and Williamson, 2017). In fact, this cost can become prohibitively high if one aims to achieve multiple fairness criteria *simultaneously*. For example, Chouldechova (2016) and Kleinberg et al. (2017b) have recently shown 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", or null, classifier that assign all examples to a single class). Pleiss et al. (2017) further expand on this impossibility result. These results suggest that satisfying all five criteria of disparate mistreatment (Table 1) simultaneously is impossible when the underlying distribution of data is different for different groups. Similarly, Kleinberg et al. (2017b) show that when the fraction of users with positive class labels differ between members of different sensitive attribute value groups, it is impossible to satisfy disparate impact and disparate mistreatment simultaneously—where disparate mistreatment is defined in terms of false positive and false negative rates, in terms of false discovery and false omission rates, or both. However, in practice, it may still be interesting to explore the best, even if imperfect, extent of fairness a classifier can achieve.

4. Our fair classification framework

In this section, we present our framework to design boundary-based classifiers which are free of disparate treatment, disparate impact and disparate mistreatment, as defined in Section 3.

4.1. Fairness criteria as constraints during training

To design a *fair* convex boundary-based classifier, one can think of including fairness constraints during training. More specifically, minimizing the corresponding loss function under fairness constraints, *i.e.*,

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) && \big\} \text{ Classifier loss function} \\ \text{subject to} \quad & P(.|z=0) = P(.|z=1) && \big\} \text{ Fairness constraints,} \end{aligned} \tag{4.1}$$

where the probabilities in the constraint(s) can be replaced with the respective disparate impact and disparate mistreatment criteria in Eqs. (3.2–3.7). Here, note that, if $z \notin \boldsymbol{x}$ (*i.e.*, $\boldsymbol{x}$ and z consist of disjoint feature sets), the resulting classifier does not suffer from disparate treatment since z is not used during test (*i.e.*, at decision time).

Due to its flexibility, the above formulation exhibits several advantages:

(i) it can satisfy both disparate impact and (any version of) disparate mistreatment by including the corresponding constraints. Disparate treatment can be achieved by excluding the sensitive features z from $\boldsymbol{x}$ so that they are not used during test.

(ii) it can accommodate any convex decision boundary based classifier.
(iii) it can ensure fairness with respect to multiple sensitive features (*e.g.*, gender, race) by including constraints for each sensitive feature separately.

Unfortunately, solving the formulation given by Eq. (4.1) is very challenging. First, for many such classifiers (*e.g.*, SVM), the probabilities in Eqs. (3.2–3.7) are a non-convex function of the classifier parameters $\boldsymbol{\theta}$, therefore leading to non-convex formulations, which are difficult to solve efficiently. Second, as long as the user feature vectors lie on the same side of the decision boundary, the probabilities are invariant to changes in the decision boundary. In other words, the probabilities are functions having saddle points. The presence of saddle points furthers complicate the procedure for solving non-convex optimization problems (Dauphin et al., 2014).

To overcome these challenges, we next introduce a relaxation of the group-conditional probability constraints given by Eqs. (3.2–3.7) using a novel covariance measure of decision boundary unfairness.

4.2. Designing fair classifiers using decision boundary covariances

In this section, we first introduce our covariance measure of decision boundary unfairness in the context of disparate impact, use this measure to design classifiers free of disparate impact, and then generalize our measure to disparate mistreatment.

4.2.1. Disparate impact

We measure the decision boundary (un)fairness due to disparate impact by means of the covariance between the users' sensitive attribute z and the signed distance from the users' feature vectors to the decision boundary $d_{\boldsymbol{\theta}}(\boldsymbol{x})$, *i.e.*,

$$\text{Cov}_{DI}(z, d_{\boldsymbol{\theta}}(\boldsymbol{x})) = \mathbb{E}[(z - \bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}[(z - \bar{z})]\bar{d_{\boldsymbol{\theta}}}(\boldsymbol{x}) \approx \frac{1}{N} \sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z - \bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}), \qquad (4.2)$$

where $\mathbb{E}[(z - \bar{z})]\bar{d_{\boldsymbol{\theta}}}(\boldsymbol{x})$ cancels out since $\mathbb{E}[(z - \bar{z})] = 0$. Note that, if a decision boundary satisfies Eq. (3.2), *i.e.*, $P(d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0|z = 0) = P(d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0|z = 1)$, then the (empirical) covariance defined above will be (approximately) zero (for a sufficiently large training set)[6]. Moreover, in contrast to the group-conditional probabilities given by Eq. (3.2) the decision boundary covariance in Eq. (4.2) is a convex function with respect to the decision boundary parameters $\boldsymbol{\theta}$ because $d_{\boldsymbol{\theta}}(\boldsymbol{x})$ is convex with respect to $\boldsymbol{\theta}$ for all linear, convex boundary-based classifiers. [7] Hence, it can be easily included in the formulation of these classifiers without increasing the complexity of their training.

6. Note that the converse is not true, that is why we call our covariance measure a proxy.
7. For non-linear convex boundary-based classifiers like non-linear SVM, the equivalent of $d_{\boldsymbol{\theta}}(\boldsymbol{x})$ (via Representer Theorem) is still convex in the corresponding reproducing kernel Hilbert space as we will discuss shortly.

More specifically, to train a classifier free of disparate impact, one can replace the (intractable) constraint in Eq. (4.1) by an alternative constraint including the decision boundary covariance constraint as follows:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}) \leq c, \\ & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq -c, \end{array} \tag{4.3}$$

where $c \in \mathbb{R}^+$ is a given threshold, which trades off accuracy and unfairness due to disparate impact. Furthermore, note that since the above optimization problem is convex, our scheme ensures that the trade-off between the classifier loss function and decision boundary covariance is Pareto optimal.

When considering *multiple sensitive features* (*e.g.*, gender, race), one can include constraints for each sensitive feature separately. For *polyvalent sensitive features* having $k \geq 2$ values, one can first convert the sensitive feature into k binary sensitive features using one hot encoding, and then add constraints for each of the k sensitive features. To avoid "fairness-gerrymandering" (Kearns et al., 2018) when considering multiple sensitive features, one could construct all possible combinations of the sensitive feature values (*e.g.*, white man, black women) and add constraints for each combination separately.

4.2.2. Disparate Mistreatment

We can naturally extend our covariance measure to (un)fairness due to disparate mistreatment. More specifically, for the case disparate mistreatment with respect to the overall misclassification rate, we compute 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.*,

$$\text{Cov}_{OMR}(z, g_{\boldsymbol{\theta}}(y,\boldsymbol{x})) = \mathbb{E}[(z-\bar{z})(g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) - \bar{g}_{\boldsymbol{\theta}}(y,\boldsymbol{x}))] \approx \frac{1}{N}\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}), \tag{4.4}$$

where $g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) = \min(0, y d_{\boldsymbol{\theta}}(\boldsymbol{x}))$ and the term $\mathbb{E}[(z-\bar{z})]\bar{g}_{\boldsymbol{\theta}}(\boldsymbol{x})$ cancels out since $\mathbb{E}[(z-\bar{z})] = 0$. As in the case of disparate impact, if a decision boundary satisfies Eq. (3.3), then the (empirical) covariance defined above will be (approximately) zero (for a sufficiently large training set) and we can train a classifier free of disparate mistreatment with respect to overall misclassification rate by replacing the (intractable) constraint in Eq. (4.1) by an alternative constraint as follows:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c, \\ & \frac{1}{N}\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{array} \tag{4.5}$$

where $c \in \mathbb{R}^+$ is a given threshold, which trades off accuracy and unfairness due to disparate mistreatment. Similarly, we can define the above covariance measure for disparate mistreatment with respect to false positive rates, false negative rates, false omission rates

or false discovery rates. For example, for false positive rates, one needs to consider the set of *misclassified* users with (ground-truth) negative labels ($\mathcal{D}^-$), *i.e.*,

$$\text{Cov}_{FPR}(z, g_{\boldsymbol{\theta}}(y, \boldsymbol{x})) \approx \frac{1}{N^-} \sum_{(\boldsymbol{x},y,z)\in\mathcal{D}^-} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \boldsymbol{x}), \tag{4.6}$$

where N^- represents the size of $\mathcal{D}^-$.

In contrast with the covariance measure in the case of disparate impact, defined by Eq. (4.2), the above covariance measures are not convex. Fortunately, the covariance constraints for disparate mistreatment with respect to overall misclassification rates, false positive rates and false negative rates can be easily converted into convex-concave constraints, which can be handled efficiently (Shen et al., 2016), as follows. Consider the constraints in Eq. (4.5), *i.e.*,

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

where $\sim$ denotes '$\geq$' and '$\leq$' and, without loss of generality, we just left out the constant term $\frac{1}{N}$. Then, we can split the sum in the above expression into two terms:

$$\sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} (0 - \bar{z})\, g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) + \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} (1 - \bar{z})\, g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) \sim c, \tag{4.7}$$

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 Eq. (4.7) as:

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

which, given that $g_{\boldsymbol{\theta}}(y, \boldsymbol{x})$ is convex in $\boldsymbol{\theta}$, results into a convex-concave (or, difference of convex) function. Finally, we can rewrite the problem defined by (4.5) as:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) + \frac{N_0}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) \leq c \\ & \frac{-N_1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) + \frac{N_0}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) \geq -c, \end{array} \tag{4.8}$$

which is a Disciplined Convex-Concave Program (DCCP) for any convex loss $L(\boldsymbol{\theta})$, and can be efficiently solved using well-known heuristics (Shen et al., 2016). Note that the non-convexity of the formulation in Eq. (4.8) implies that the resulting tradeoff between the classifier loss function and decision boundary covariance is not guarenteed to be Pareto optimal. This is in contrast with the convex formulation in Eq. (4.3). However, as we show via comparisons with related methods in Section 5.2.2, Eq. (4.8) can still lead to competitive results in comparison with the state-of-the-art.

Proceeding similarly, we can convert the covariance constraints for disparate mistreatment with respect to false positive rates and false negative rates to convex-concave constraints. For example, Eq. (4.1) can be rewritten to impose equality in false positive rates

as:

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \frac{-N_1^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c \\ & \frac{-N_1^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{aligned} \tag{4.9}$$

where $\mathcal{D}_i^-$ is the subset of the training data with $z = i$ and $y = -1$, and $N_i^- = |\mathcal{D}_i^-|$. Note that unlike in Zafar et al. (2017a), we define the false positive rate covariance (Eq. (4.6)) only over the ground truth negative dataset instead of the whole dataset. In cases where $\frac{N_0^-}{N_0} \neq \frac{N_1^-}{N_1}$ (or in other words, the base-rates are different for the two sensitive feature groups), the false positive rate covariance as defined by Zafar et al. (2017a) would not fully remove disparate mistreatment.

Finally, while the covariance constraints for disparate mistreatment with respect to false omission and false discovery rates can be readily defined, the corresponding constraints cannot be easily converted into convex-concave constraints. Handling such constraints efficiently is left as an interesting venue for future work.

4.3. Accounting for the business necessity clause

In the previous section, we have used covariance constraints to design classifiers that maximize accuracy under fairness constraints. However, if the underlying correlation between the class labels and the sensitive attributes in the training set is very high, enforcing these constraints may results in underwhelming performance (accuracy) and thus be unacceptable in terms of business objectives. This is particularly concerning in the case of disparate impact, where a "business necessity" clause has been argued for—an employer would need to ensure that the decision making causes *least possible* disparate impact under the given performance (accuracy) constraints (Barocas and Selbst, 2016).

Fortunately we can account for the above mentioned "business necessity" clause in disparate impact using an alternative formulation that maximizes fairness (minimizes disparate impact) subject to accuracy constraints. More specifically, we can find the decision boundary parameters $\boldsymbol{\theta}$ by minimizing the corresponding (absolute) decision boundary covariance over the training set under constraints on the classifier loss function, *i.e.*:

$$\begin{aligned} \text{minimize} \quad & \left|\tfrac{1}{N} \textstyle\sum_{i=1}^{N} (z_i - \bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| \\ \text{subject to} \quad & L(\boldsymbol{\theta}) \leq (1+\gamma) L(\boldsymbol{\theta}^*), \end{aligned} \tag{4.10}$$

where $L(\boldsymbol{\theta}^*)$ denotes the optimal loss over the training set provided by the unconstrained classifier and $\gamma \geq 0$ specifies the maximum additional loss with respect to the loss provided by the unconstrained classifier. Here, we can ensure maximum fairness with no loss in accuracy by setting $\gamma = 0$.

Remarkably, in many classifiers, including logistic regression and SVMs, the loss function (or the dual of the loss function) is additive over the points in the training set, *i.e.*, $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta})$, where $L_i(\boldsymbol{\theta})$ is the individual loss associated with the i-th point in the training set. Moreover, the individual loss $L_i(\boldsymbol{\theta})$ typically tells us how *close* the predicted label $f(\boldsymbol{x}_i)$ is to the true label y_i, by means of the signed distance to the decision boundary. Therefore, one may think of incorporating loss constraints for a certain set of users, and consequently, prevent individual users originally classified as positive (by the unconstrained

classifier) from being classified as negative by the constrained classifier. To do so, we find the decision boundary parameters $\boldsymbol{\theta}$ as:

$$\begin{array}{ll} \text{minimize} & |\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)| \\ \text{subject to} & L_i(\boldsymbol{\theta}) \leq (1+\gamma_i) L_i(\boldsymbol{\theta}^*) \quad \forall i \in \{1,\ldots,N\}, \end{array} \tag{4.11}$$

where $L_i(\boldsymbol{\theta}^*)$ is the individual loss associated to the i-th user in the training set provided by the unconstrained classifier and $\gamma_i \geq 0$ is her allowed additional loss.

The problem formulation in Eq. (4.11) could be used to ensure that certain set of users who are correctly classified by the unconstrained classifier are not misclassified by the fairness-constrained classifier (*e.g.*, to make sure that applying fairness constraints does not lead to egregious misclassification of certain users). However, note that, in comparison with Eq. (4.10), Eq. (4.11) involves a tighter set of constraints—it aims to bound *individual* users' losses as opposed to the *aggregate* loss over all the users. As a result, Eq. (4.11) could lead to larger drops in accuracy for the same level of fairness (refer to Figure 10 for an example).

One could also think of extending the formulation of disparate mistreatment-free classification given by Eq. (4.8) to include a similar business necessity clause. However, such a formulation would result in an optimization problem—with a convex-concave objective and convex constraints—that is currently not supported by the standard convex-concave solvers (Shen et al., 2016). Extending these solvers to cater to such problems, or reformulating the optimization problem and solving it with alternative optimizers (*e.g.*, using evolutionary multi-objective optimization as in Quadrianto and Sharmanska (2017)) would be an interesting direction for future work.

4.4. Examples

In this section, we illustrate how to design fair logistic regression classifiers and linear and nonlinear SVMs using our covariance measures.

Logistic regression free of disparate impact. In logistic regression classifiers, one maps the feature vectors $\boldsymbol{x}_i$ to the class labels y_i by means of a probability distribution:

$$p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}, \tag{4.12}$$

where $\boldsymbol{\theta}$ is obtained by solving a maximum likelihood problem over the training set, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} -\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta})$. Thus, the corresponding loss function is given by $-\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta})$, and the problem defined by Eq. (4.3) adopts the following form:

$$\begin{array}{rl} \text{minimize} & -\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x} \leq c, \\ & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x} \geq -c. \end{array} \quad \begin{array}{l} \left.\vphantom{\begin{array}{l}a\\b\end{array}}\right\} \text{Logistic regression formulation} \\ \left.\vphantom{\begin{array}{l}a\\b\end{array}}\right\} \text{Disparate impact constraints} \end{array} \tag{4.13}$$

Linear SVM free of disparate mistreatment. A linear SVM distinguishes among classes using a linear hyperplane $\boldsymbol{\theta}^T\boldsymbol{x} = 0$. In this case, the parameter vector $\boldsymbol{\theta}$ of the *fair*

linear SVM can be found by solving the problem defined by Eq. (4.8), which becomes the following quadratic program with convex-concave constraints:

$$\left.\begin{array}{ll} \text{minimize} & \|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} & y_i\boldsymbol{\theta}^T\boldsymbol{x}_i \geq 1-\xi_i, \forall i \in \{1,\ldots,N\} \\ & \xi_i \geq 0, \forall i \in \{1,\ldots,N\}, \end{array}\right\} \text{SVM formulation}$$

$$\left.\begin{array}{l} \frac{-N_1}{N}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_0}\min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) + \frac{N_0}{N}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_1}\min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \leq c, \\ \frac{-N_1}{N}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_0}\min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) + \frac{N_0}{N}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_1}\min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \geq -c, \end{array}\right\} \begin{array}{c}\text{Disparate}\\ \text{mistreatment}\\ \text{constraints}\end{array} \tag{4.14}$$

where $\boldsymbol{\theta}$ and ξ are the variables, $\|\boldsymbol{\theta}\|^2$ corresponds to the boundary between the *support vectors* assigned to different classes, and $C\sum_{i=1}^{n}\xi_i$ penalizes the number of data points falling inside the boundary.

Nonlinear SVM free of disparate impact. In a nonlinear SVM, the decision boundary takes the form $\boldsymbol{\theta}^T\Phi(\boldsymbol{x}) = 0$, where $\Phi(\cdot)$ is a nonlinear transformation that maps every feature vector $\boldsymbol{x}$ into a higher dimensional transformed feature space. Similarly as in the case of a linear SVM, one may think of finding the parameter vector $\boldsymbol{\theta}$ by solving a constrained quadratic program, similar to the one defined by Eq. (4.14). However, the dimensionality of the transformed feature space can be large, or even infinite, making the corresponding optimization problem difficult to solve. Fortunately, we can leverage the *kernel trick* (Schölkopf and Smola, 2002) both in the original optimization problem and the fairness inequalities, and resort instead to the dual form of the problem, which can be solved efficiently. In particular, the dual form is given by (for conciseness, we use the dual form notation of Gentle et al. (2012)):

$$\left.\begin{array}{ll} \text{minimize} & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \\ \text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C, \\ & \boldsymbol{y}^T\boldsymbol{\alpha} = 0, \end{array}\right\} \text{SVM formulation}$$

$$\left.\begin{array}{l} \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\, d_{\boldsymbol{\alpha}}(\boldsymbol{x}) \leq c, \\ \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\, d_{\boldsymbol{\alpha}}(\boldsymbol{x}) \geq -c, \end{array}\right\} \begin{array}{c}\text{Disparate impact}\\ \text{constraints}\end{array} \tag{4.15}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$ are the dual variables, $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T$ are the class labels, $\boldsymbol{G}$ is the $N \times N$ Gram matrix with $\boldsymbol{G}_{i,j} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$ denotes the inner product between a pair of transformed feature vectors. Here, $d_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) = \sum_{j=1}^{N}\alpha_j y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can still be interpreted as a signed distance to the decision boundary in the transformed feature space.

5. Evaluation

In this section, we first experiment with synthetic datasets in which we simulate various kind of unfairness and analyze both quantitatively and qualitatively the effectiveness of our framework at designing fair classifiers. We then evaluate the performance of our framework on several real-world datasets in comparison with multiple baselines.

Across this section, we quantify disparate impact (Eq. 3.2) as the absolute difference between the positive class probability for the sensitive feature groups with $z = 0$ and $z = 1$, as various prior studies (Corbett-Davies et al., 2017; Calders and Verwer, 2010; Menon and Williamson, 2017), *i.e.*,

$$DI = \Big|P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)\Big|, \tag{5.1}$$

where a value of DI closer to zero denotes a smaller degree of disparate impact. Similarly, we quantify disparate mistreatment with respect to false positive rates and false negative rates (Eqs. (3.4-3.5)) as the difference between false positive (negative) rate probabilities, *i.e.*,

$$DM_{FPR} = P(\hat{y} \neq y|z = 0, y = -1) - P(\hat{y} \neq y|z = 1, y = -1), \tag{5.2}$$
$$DM_{FNR} = P(\hat{y} \neq y|z = 0, y = 1) - P(\hat{y} \neq y|z = 1, y = 1), \tag{5.3}$$

where the closer the values of DM_{FPR} and DM_{FNR} to 0, the lower the degree of disparate mistreatment. Note that unlike in the case of disparate impact in Eq. (5.1), we do not use the absolute difference while quantifying disparate mistreatment. As we later show in Section 5.1.2, the (in)equality in the signs of DM_{FPR} and DM_{FNR} carries significant consequences when considering disparate mistreatment w.r.t. false positive rate and false negative rate simultaneously. In such cases, the sign of the differences should also be taken into account.

5.1. Experiments on synthetic data

In this section, we first generate synthetic data where a classifier optimizing for accuracy would lead to disparate impact and then generate data where the accuracy-optimizing classifier would lead to disparate mistreatment. In both the cases, we *simultaneously* control for disparate treatment as well, that is, the classifiers do not leverage sensitive feature during decision time in either of the cases.

5.1.1. Mitigating disparate impact

To simulate different degrees of disparate impact in classification outcomes, we generate two synthetic datasets with different levels of correlation between a single, binary sensitive attribute and class labels. Specifically, we generate 4,000 binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from two different Gaussian distributions:

$$p(\boldsymbol{x}|y = 1) = \mathcal{N}([2; 2], [5,\ 1; 1,\ 5])$$
$$p(\boldsymbol{x}|y = -1) = \mathcal{N}([-2; -2], [10,\ 1; 1,\ 3]).$$

Then, we draw each user's sensitive attribute z from a Bernoulli distribution: $p(z = 1) = p(\boldsymbol{x}'|y = 1)/(p(\boldsymbol{x}'|y = 1) + p(\boldsymbol{x}'|y = -1))$, where $\boldsymbol{x}' = [\cos(\phi),\ -\sin(\phi); \sin(\phi),\ \cos(\phi)]\boldsymbol{x}$ is simply a rotated version of the feature vector, $\boldsymbol{x}$. We generate two datasets with different values for the parameter ϕ ($\pi/4$ and $\pi/8$), which controls the correlation between the sensitive attribute, z, and the class labels, y (and hence, the resulting degree of disparate impact). Here, the closer ϕ is to zero, the higher the correlation between z and y.

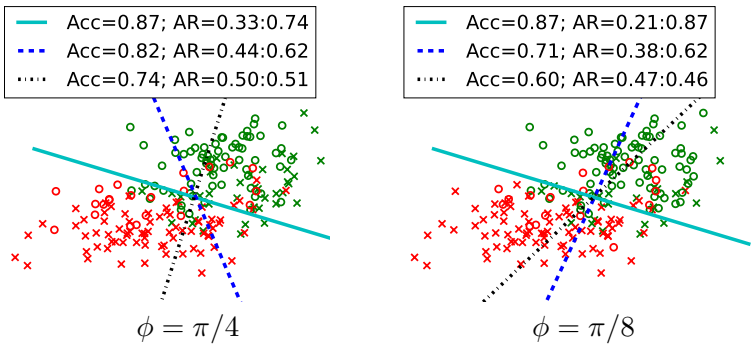


Figure 2: [Disparate impact constraints] Performance of different (unconstrained and constrained) classifiers along with their accuracy (Acc) and positive class acceptance rates (AR) for groups $z = 0$ (crosses) and $z = 1$ (circles). Green points represent examples with $y = 1$ and red points represent example with $y = -1$. The solid lines show the decision boundaries for logistic regression classifiers without fairness constraints. The dashed lines show the decision boundaries for logistic regression classifiers trained to maximize accuracy under fairness constraints (Eq. (4.13)). Each column corresponds to a dataset with different correlation value between sensitive attribute values and class labels.

Next, we train logistic regression classifiers optimizing for accuracy on both the datasets. The accuracy of the classifiers in both cases is 0.87 (note that the datasets only differ in terms of the correlation between z and y). However, the classifiers lead to $DI = |0.33 - 0.74| = 0.41$ and $DI = |0.21 - 0.87| = 0.66$ on datasets with $\phi = \pi/4$ and $\phi = \pi/8$, respectively. To overcome the unfairness, we train logistic regression classifiers with disparate impact constraints (Eq. 4.13) on both datasets.

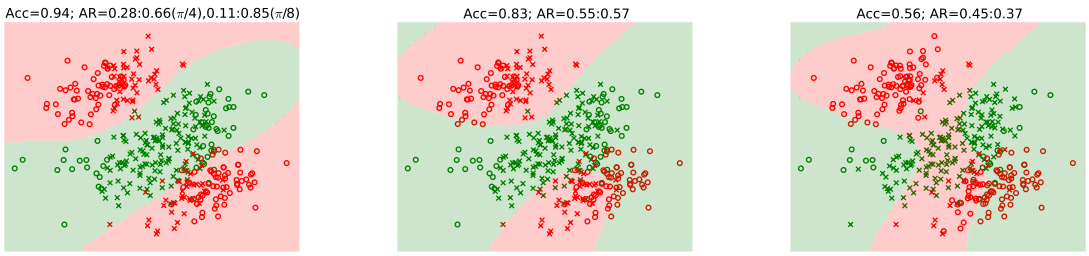
Figure 2 shows the decision boundaries provided by the classifiers for two (successively decreasing) covariance thresholds, c. We compare these boundaries against the unconstrained decision boundary (solid line). As expected, given the data generation process, fairness constraints map into a rotation of the decision boundary (dashed lines), which is greater as we decrease threshold value c or increase the correlation in the original data (from $\phi = \pi/4$ to $\phi = \pi/8$). This movement of the decision boundaries shows that our fairness constraints are successfully undoing (albeit in a highly controlled setting) the rotations we used to induce disparate impact in the dataset. Moreover, a smaller covariance threshold (a larger rotation) leads to a more fair solution, although, it comes at a larger cost in accuracy.

Next, we illustrate how the decision boundary of an non-linear classifier, a SVM with radial basis function (RBF) kernel, changes under disparate impact constraints (Eq. (4.15)). To this end, we generate 4,000 user binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from

$$p(\boldsymbol{x}|y = 1, \beta) = \beta\mathcal{N}([2; 2], [5\ 1; 1\ 5]) + (1 - \beta)\mathcal{N}([-2; -2], [10\ 1; 1\ 3])$$
$$p(\boldsymbol{x}|y = -1, \beta) = \beta\mathcal{N}([4; -4], [4\ 4; 2\ 5]) + (1 - \beta)\mathcal{N}([-4; 6], [6\ 2; 2\ 3])$$

where $\beta \in \{0, 1\}$ is sampled from Bernoulli(0.5). Then, we generate each user's sensitive attribute z by applying the same rotation as described earlier.

Figure 3 shows the decision boundaries provided by the SVM that maximizes accuracy under fairness constraints with $c = 0$ for two different correlation values: $\phi = \pi/4$ and



(a) Unconstrained (b) $\phi = \pi/4$ (c) $\phi = \pi/8$

Figure 3: [Disparate impact constraints] Decision boundaries for SVM classifier with RBF Kernel trained without fairness constraints (left) and with fairness constraints (middle and right) on two synthetic datasets. Also shown are the classification accuracy (Acc) and acceptance rate (AR) for each group.

$\phi = \pi/8$, in comparison with the unconstrained SVM. We observe that, in this case, the decision boundaries provided by the constrained SVMs are very different to the decision boundary provided by the unconstrained SVM, and are not just simple shifts or rotations of the latter.

5.1.2. Mitigating disparate mistreatment

In a manner similar to the previous section, we now experiment with synthetic datasets where training classifiers optimizing for accuracy would lead to disparate mistreatment. However, disparate mistreatment can arise in multiple different ways, as detailed below. To study these different situations, 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*.

Disparate mistreatment on *only* false positive rate *or* false negative rate. The first scenario considers a case where a classifier maximizing accuracy 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.*, $DM_{FPR} \neq 0$ and $DM_{FNR} = 0$ (or, alternatively, $DM_{FPR} = 0$ and $DM_{FNR} \neq 0$).

To simulate this scenario, we 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 ($\boldsymbol{x}$) 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):

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

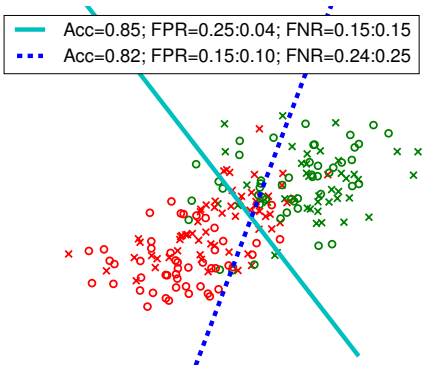


Figure 4: [Disparate mistreatment constraints] The figure shows the original decision boundary (solid line) and fair decision boundary (dashed line), along with corresponding accuracy and false positive rates for groups $z = 0$ (crosses) and $z = 1$ (circles). Fairness constraints cause the original decision boundary to rotate such that previously misclassified subjects with $z = 0$ are moved into the negative class (decreasing false positives), while well-classified subjects with $z = 1$ are moved into the positive class (increasing false positives), leading to equal false positive rates for both groups.

Next, we train a logistic regression classifier optimizing for accuracy 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 groups, it achieves $DM_{FNR} = 0.15 - 0.15 = 0$ and $DM_{FPR} = 0.25 - 0.04 = 0.21$, which constitutes a clear case of disparate mistreatment in terms of false positive rate. We then train a logistic regression classifier subject to fairness constraints on false positive rate, with a covariance threshold $c = 0$.

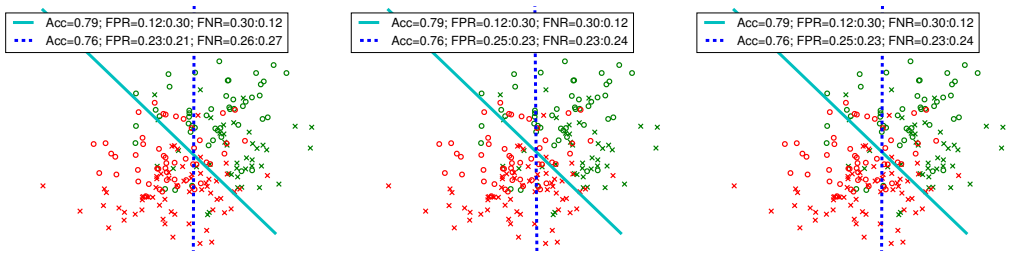
Figure 4 shows the decision boundaries for both the unconstrained classifier (solid) and the fairness-constrained classifier (dashed). We observe that applying the fairness constraint successfully causes the false positive rates for both groups ($z = 0$ and $z = 1$) to converge, and hence, the outcomes of the classifier become more fair, *i.e.*, $DM_{FPR} \to 0$, while DM_{FNR} remains close to zero. We note that the invariance of DM_{FNR} may however change depending on the underlying distribution of the data.

Disparate mistreatment on *both* false positive rate and false negative rate. In this part, 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 DM_{FPR} and DM_{FNR} are non-zero. This scenario can in turn be split into two cases:

I. DM_{FPR} and DM_{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. DM_{FPR} and DM_{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.



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

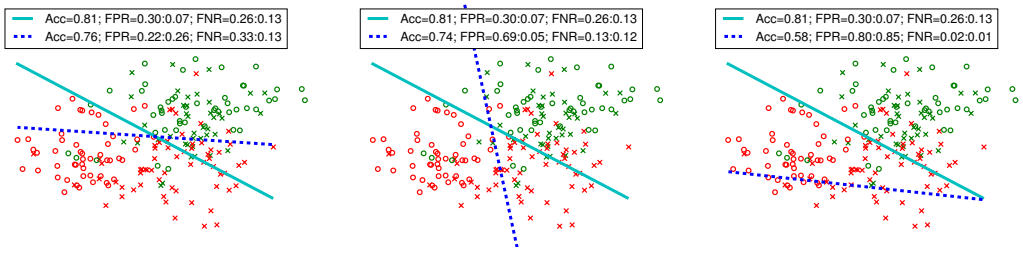
Figure 5: [Disparate mistreatment constraints] DM_{FPR} and DM_{FNR} have opposite signs. Removing disparate mistreatment on FPR can potentially help remove disparate mistreatment on FNR. Removing disparate mistreatment on both at the same time leads to very similar results.

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

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

An accuracy-maximizing logistic regression classifier on this dataset attains an overall accuracy of 0.79 but leads to a false positive rate of 0.12 and 0.30 (*i.e.*, $DM_{FPR} = 0.12 - 0.30 = -0.18$) for the sensitive attribute groups $z = 0$ and $z = 1$, respectively; and false negative rates of 0.30 and 0.12 (*i.e.*, $DM_{FNR} = 0.30 - 0.12 = 0.18$). To remove this disparate mistreatment, we train three different classifiers, with fairness constraints on (i) false positive rates (ii) false negative rates and (iii) on both false positive and false negative rates.

Figure 5 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* subjects with $z = 1$ into the negative class, *decreasing* their false positive rate. However, in the process, it also moves previously *well-classified* subjects with $z = 1$ into the negative class, *increasing* their false negative rate. As a consequence, controlling disparate mistreatment on false positive rate (Figure 5(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 5(b)), and therefore, provides similar results as the constrained classifier for both false positive and false negative rates (Figure 5(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 $z = 1$.



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

Figure 6: [Disparate mistreatment constraints] DM_{FPR} and DM_{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.

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

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

We then train a accuracy-optimizing logistic regression classifier on this dataset. It attains an accuracy of 0.81 but leads to $DM_{FPR} = 0.30 - 0.07 = 0.23$ and $DM_{FNR} = 0.26 - 0.13 = 0.13$, 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.

Figure 6 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 observe following noticeable patterns. First, controlling disparate mistreatment for only false positive rate (false negative rate), leads to a minor drop in accuracy, but in contrast to Case I, 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 subjects with $z = 0$ are well-classified in the positive class (reducing false negative rate), it also moves previously well-classified negative subjects 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 DM_{FPR} and DM_{FNR} close to zero, but causes a large drop in accuracy.

5.2. Real-world datasets

We now evaluate the effectiveness of our covariance framework in removing disparate impact and disparate mistreatment on several real-world datasets. In doing so, we also compare the performance of our framework to several methods from the fair machine learning literature.

In all the experiments, to obtain more reliable estimates of accuracy and fairness, we repeatedly split each dataset into a train (70%) and test (30%) set 5 times and report the average statistics for accuracy and fairness.

5.2.1. Mitigating disparate impact

Datasets and experimental setup. Here, we experiment with two real-world datasets: The Adult income dataset (Adult, 1996) and the Bank marketing dataset (Bank, 2014).

The Adult dataset contains a total of 45,222 subjects, each with 14 features (*e.g.*, age, educational level) and a binary label, which indicates whether a subject's incomes is above (positive class) or below (negative class) 50K USD. With the aim of experimenting with binary as well as non-binary (polyvalent) sensitive attributes, we consider the features gender and race to be sensitive. Here, gender (with feature values: men and women) serves as an example of binary sensitive attribute and race (with feature values: American-Indian, Asian, Black, White and Other) serves as an example of a non-binary sensitive attribute.

The Bank dataset contains a total of 41,188 subjects, each with 20 attributes (*e.g.*, marital status) and a binary label, which indicates whether the client has subscribed (positive class) or not (negative class) to a term deposit. In this case, we consider age as (binary) sensitive attribute, which is discretized to indicate whether the client's age is between 25 and 60 years. For detailed statistics about the distribution of different sensitive attributes in positive class in these datasets, we refer the reader to Appendix A.

For the sake of conciseness, while presenting the results for binary sensitive attributes, we refer to women and men, respectively, as protected and non-protected groups in Adult data. Similarly, in Bank data, we refer to users between age 25 and 60 as protected and rest of the users as non-protected group.

Methods. In our experiments, we also compare our approach to well-known competing method from fairness-aware machine learning literature (detailed in Section 6). More specifically, we consider the following methods:

- ***Our method (C-LR and C-SVM)*:** Implements our covariance constraints-based methods for controlling disparate impact with a logistic regression classifier (Eq. (4.13)) and a dual-form SVM classifier with a linear kernel (Eq. (4.15)). [8] On the datasets considered here, different choices of kernel (linear vs. RBF) lead to a very similar performance in terms of accuracy and disparate impact. This method does not use the sensitive feature information at decision time.
- ***Preferential sampling (PS-LR and PS-SVM)*:** Implements the data pre-processing technique of Kamiran and Calders (2010) on a logistic regression and a SVM classifier. Specifically, this method operates as follows: (i) We first train a standard (potentially

8. For the SVM classifier, the hyperparameter C (in Eq. (4.15)) is only cross-validated for the unconstrained classifier, and the same hyperparameter is used for the fairness-constrained classifiers. One could further optimize the classifier performance by cross-validating the value of C for each value of the covariance threshold separately.

unfair) classifier on the given dataset. (ii) Next, we move / replicate the protected group data points to / on the positive side of the decision boundary (and vice versa for the non-protected group) until the decision boundary leads to zero disparate impact, *i.e.*, until it satisfies Eq. (3.2). (iii) We then train the final (fair) classifier on the perturbed dataset. This method does not use the sensitive feature information at decision time.

- ***Fairness-regularized logistic regression (FR-LR)*:** The in-processing technique of Kamishima et al. (2011). This technique is only limited to the logistic regression classification model. This technique works by adding a fairness regularization term in the objective function that penalizes the mutual information between the sensitive feature and the classifier decisions. In this way, this method treats the mutual information as the unfairness proxy, as opposed to covariance in our case. This technique needs the sensitive feature information at decision time, hence cannot remove disparate treatment.
- ***Post-Processing (PP-LR and PP-SVM)*:** The post-processing technique discussed in Corbett-Davies et al. (2017). This method works by first training a standard logistic regression or SVM classifier on the given dataset and then finding a pair of positive class acceptance thresholds [9] such that the decisions based on those thresholds lead to maximum accuracy while having no disparate impact. This technique also requires the sensitive feature information at decision time so it cannot avoid disparate treatment.

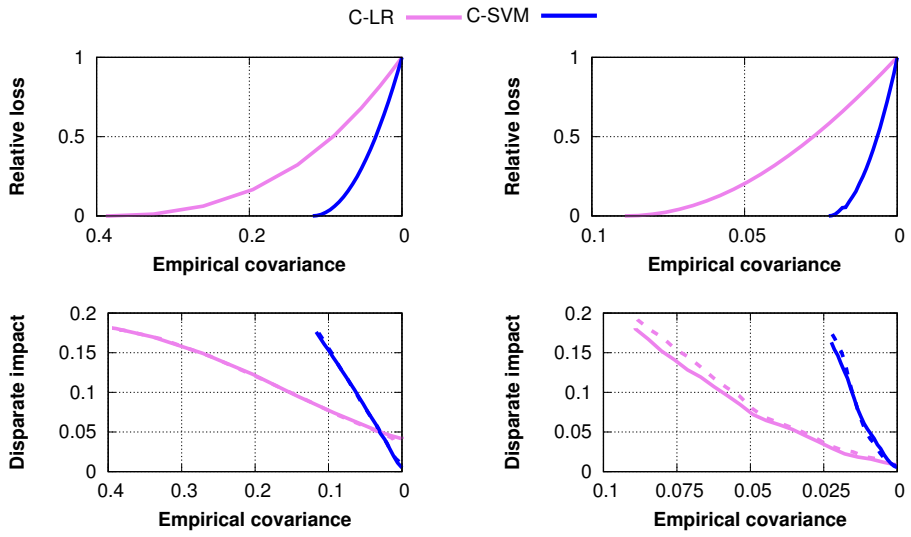
Results. First, we experiment with two standard (unconstrained) logistic regression and SVM classifiers. In the Adult dataset, the logistic regression classifier leads to an accuracy of 0.846. However, the classifier results in highly disparate positive class acceptance rates for protected and non-protected groups: 0.08 and 0.26. The SVM classifier leads to a similar accuracy (0.847) and disparity in positive class acceptance rates (0.08 vs 0.25). In the Bank dataset, the two classifiers lead to accuracies of 0.911 and 0.910, respectively, and acceptance rates of 0.06 vs. 0.25, and 0.05 vs. 0.23 respectively. The high disparity in acceptance rates over the two datasets clearly constitutes a case of disparate impact.

We then apply our framework to limit disparate impact with respect to a single binary sensitive attribute, gender and age, for respectively, the Adult and Bank datasets. For each dataset, we train several logistic regression and SVM classifiers (denoted by 'C-LR' and 'C-SVM', respectively), each subject to fairness constraints with different values of covariance threshold, c (Eqs.(4.13, 4.15)). Next, we study the effect of covariance constraints on the loss function value, level of disparate impact and accuracy of the classifier.

Figure 7 (top row) shows the empirical decision boundary covariance against the relative loss incurred by the classifier. The 'relative loss' is normalized between the loss incurred by an unconstrained classifier and by the classifier with a covariance threshold of 0. We notice that as expected, a decreasing value of empirical covariance results in an increasing loss. However, each pair of (covariance, loss) values is guaranteed to be Pareto optimal, since our problem formulation is convex.

The bottom row in Figure 7 investigates the correspondence between decision boundary covariance and disparate impact, as defined in Eq. (5.1), computed on the training set (solid

9. The acceptance threshold is zero for a standard logistic regression or SVM classifier.

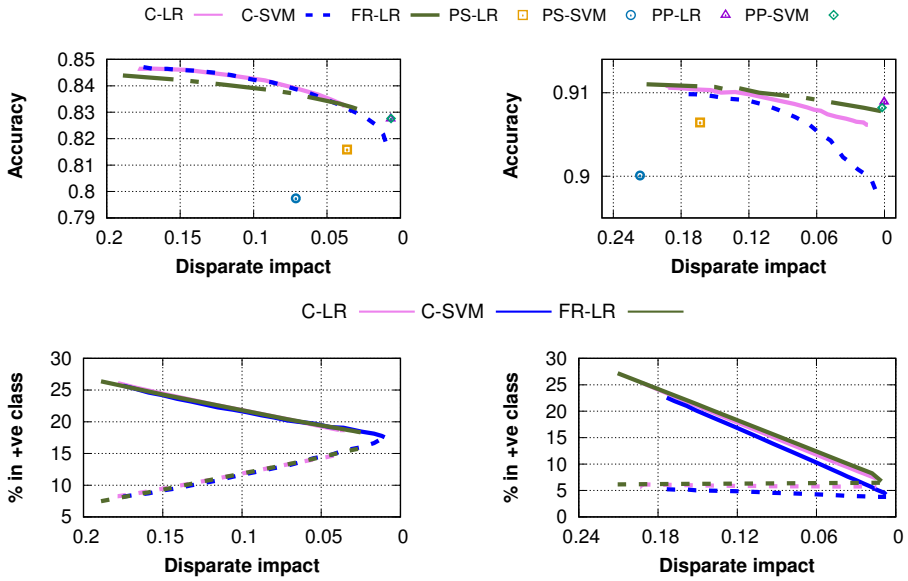


Adult data Bank data

Figure 7: [Disparate impact constraints: Single, binary sensitive attribute] Panels in the top row show the trade-off between the empirical covariance in Eq. (4.2) and the relative loss with respect to the unconstrained classifier, for the Adult and Bank (training) datasets. Here each pair of (covariance, loss) values is guaranteed to be Pareto optimal by construction. Panels in the bottom row show the correspondence between the empirical covariance and disparate impact in Eq. (5.1) for classifiers trained under fairness constraints. Solid lines correspond to the training data whereas dashed lines correspond to the test data. The figure shows that a decreasing empirical covariance leads to higher loss but lower disparate impact. Moreover, the classifiers do not lead to "fairness overfitting"—in fact, the dashed lines for the Adult data are barely visible due to strong correspondance between the performance on the training and test sets.

lines) as well as the test set (dashed lines). The figure shows that, as desired: (i) the lower the covariance, the lower the disparate impact of the classifier; (ii) zero disparate impact maps to roughly zero covariance; and, (iii) there is a strong correspondence between the covariance-fairness relationship across the training and the test set, that is, a fair classifier on the training data also leads to fair outcomes on the test data.

We next compare the performance of our constrained classifiers in terms of disparate impact–accuracy tradeoffs with the baselines methods mentioned above. The results presented in Figure 8, top row, show that: (i) the performance of our classifiers (C-LR, C-SVM) and fairness-regularized logistic regression (FR-LR) is comparable, ours are slightly better for Adult data (left column) while slightly worse for Bank data (right column); (ii) the preferential sampling presents the worst performance and results in high disparate impact; and, (iii) the post-processing technique leads to the best performance among all methods. However, we note that both FR-LR and PP-LR / PP-SVM use the sensitive feature information at decision time while the other two techniques do not use it.



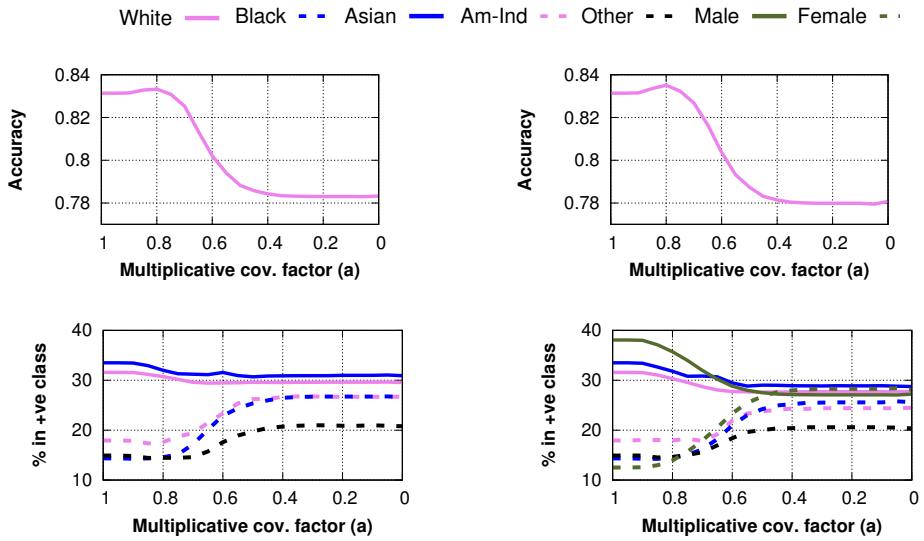
Adult data Bank data

Figure 8: [Disparate impact constraints: Single, binary sensitive attribute] The figure shows the accuracy against disparate impact in Eq. 5.1 (top) and the percentage of protected (dashed) and non-protected (solid) users in the positive class against the disparate impact value (bottom). For all methods, a decreasing degree of disparate impact also leads to a decreasing accuracy. The post-processing technique (PP-LR and PP-SVM) achieves the best disparate impact-accuracy tradeoff. However, this technique as well as FR-LR use the sensitive feature information at decision time (as opposed to C-LR, C-SVM, PS-LR and PS-SVM), and would hence result in disparate treatment.

For a more fair comparison, we also train our method with access to sensitive features at decision time. Specifically, we train constrained logistic regression classifiers (C-LR) under the same setup as above, with the exception that the non-sensitive ($\boldsymbol{x}$) and sensitive features (z) are not disjoint feature sets—that is, the classifier learns a non-zero weight for the sensitive feature z.

Under this setup, on the Adult dataset, our constrained logistic regression classifier (C-LR) achieves an accuracy of 0.839 and DI of 0.09, as compared to 0.828 accuracy and 0.01 DI achieved by the PP-LR classifier. In this case C-LR achieves a better accuracy than PP-LR, but does not remove DI as well as PP-LR. Next, we adjust the thresholds of PP-LR in a way that the resulting classifier has DI ≤ 0.9 (*i.e.*, it tries to match the DI of C-LR) while maximizing accuracy. Under these thresholds, PP-LR achieves an accuracy of 0.840 and DI of 0.07. On the Bank dataset, C-LR achieves an accuracy of 0.908 (0.909 for PP-LR) and DI of 0.01 (0.0 for PP-LR). On both Bank and Adult datasets, both methods achieve similar accuracy for a similar level of DI (with PP-LR performing marginally better).

The bottom row of Figure 8 shows the percentage of users from protected and non-protected groups in the positive class along with the degree of disparate impact. We note



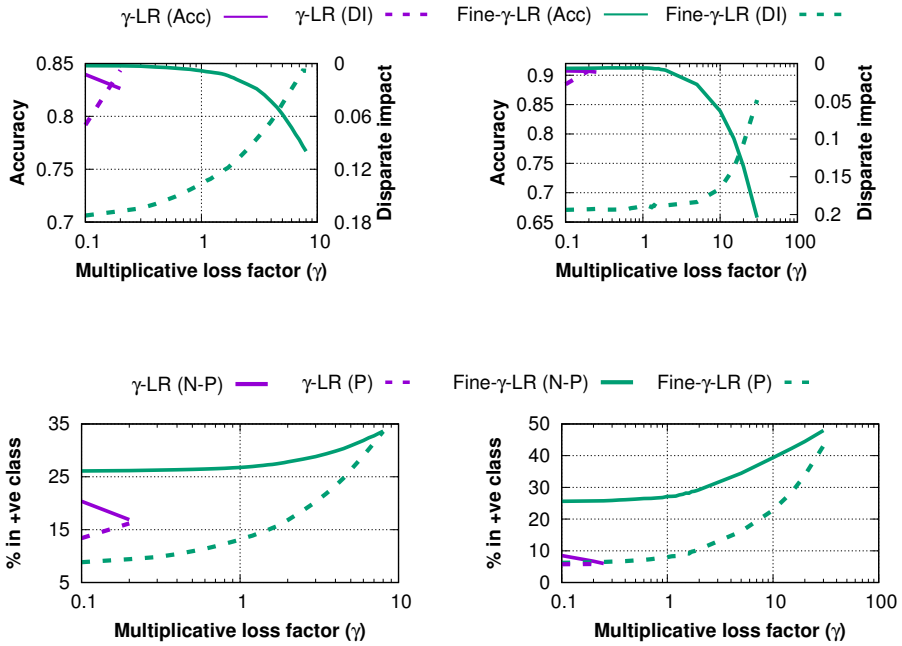
(a) Non-binary (polyvalent) sensitive attribute (b) Multiple sensitive attributes

Figure 9: [Disparate impact constraints: Non-binary and several sensitive attributes] The figure shows accuracy (top) and percentage of users in positive class (bottom) against a multiplicative factor $a \in [0, 1]$ such that $c = ac^*$, where c^* denotes the unconstrained classifier covariance.

that in the Adult data, all classifiers move non-protected users (men) to the negative class and protected users (women) to the positive class to remove disparate impact. In contrast, in the Bank data, they only move non-protected (young and old) users originally labeled as positive to the negative class since it provides a smaller accuracy loss. However, the latter can be problematic: from a business perspective, a bank may be interested in finding potential subscribers rather than losing existing customers. This observation could motivate the business necessity clause of the disparate impact doctrine. To counter such situations, one can use our alternative formulation in Section 4.3. We experiment with this formulation later in this section.

Finally, we apply our framework to control disparate impact with respect to non-binary (race) and several (gender and race) sensitive attributes in the Adult dataset. We do not compare with competing methods since they cannot handle non-binary or several sensitive attributes. Figure 9 summarizes the results by showing the accuracy and the percentage of subjects sharing each sensitive attribute value classified as positive against a multiplicative covariance factor $a \in [0, 1]$ such that $c = ac^*$, where c^* is the unconstrained classifier covariance[10] (note that disparate impact in Eq. (5.1) is only defined for a binary sensitive feature). As expected, as the value of c decreases, the percentage of subjects in the positive

10. For several sensitive features, we compute the initial covariance c_k^* for each of the sensitive feature k, and then compute the covariance threshold separately for each sensitive feature as ac_k^*.



Adult Bank

Figure 10: [Business necessity clause] Panels in top row show the accuracy (solid) and disparate impact (dashed) against γ. Panels in the bottom row show the percentage of protected (P, dashed) and non-protected (N-P, solid) users in the positive class against γ.

class from sensitive attribute value groups become nearly equal [11] while the loss in accuracy is modest.

Disparate impact's business necessity clause. We now experiment with our formulation for handling the business necessity clause (Section 4.3) to avoid scenarios where removing disparate impact leads to almost all the users being assigned the negative class label (Figure 8). Specifically, we demonstrate that our formulation in Section 4.3 can minimize disparate impact while precisely controlling loss in accuracy. We also demonstrate that our formulation can additionally provide guarantees for classifying certain users in the positive class while minimizing disparate impact.

To this end, we first train several logistic regression classifiers (denoted by 'γ-LR'), which minimize the decision boundary covariance subject to accuracy constraints over the entire dataset by solving problem (4.10) with increasing values of γ. Then, we train logistic regression classifiers (denoted by 'Fine-γ-LR') that minimize the decision boundary covariance subject to *fine-grained* accuracy constraints by solving problem (4.11). Here, we prevent the non-protected users that were classified as positive by the unconstrained logistic regression

11. The scarce representation of the race value 'Other' (only 0.8% of the data) hinders an accurate estimation of the decision boundary covariance and, as a result, the classifier does not reach perfect fairness with respect to this sensitive attribute value.

classifier from being classified as negative by constraining that their distance from decision boundary stays positive while learning the fair boundary. We then increase $\gamma_i = \gamma$ for the remaining users. In both cases, we increased the value of γ until we reach 0 disparate impact during training. Figure 10 summarizes the results for both datasets, by showing (a) the average accuracy (solid curves) and disparate impact (dashed curves) against γ, and (b) the percentage of non-protected (N-P, solid curves) and protected (P, dashed curves) users in the positive class against γ. We observe that, as we increase γ, the classifiers that constrain the overall training loss (γ-LR) remove non-protected users from the positive class and add protected users to the positive class, in contrast, the classifiers that prevent the non-protected users that were classified as positive in the unconstrained classifier from being classified as negative (Fine-γ-LR) add both protected and non-protected users to the positive class. As a consequence, the latter achieves lower accuracy for the same value of disparate impact.

5.2.2. Mitigating disparate mistreatment

Datasets and experimental setup. In this section, we experiment with two real-world datasets: COMPAS risk assessment dataset (Larson et al., 2016a) and the NYPD stop-question-and-frisk (SQF) dataset (Stop, Question and Frisk Data, 2017).

The ProPublica COMPAS dataset consists of data about $7,215$ pretrial criminal defendants, and contains a number of features such as age of the criminal defendant, number of prior criminal offenses *etc.*, and a class label indicating whether a person would recidivate within two years (positive class) or not (negative class). For more information about the data collection, we point the reader to a detailed description (Larson et al., 2016b) and some of the follow-up discussion on this dataset (Angwin and Larson; Flores et al., 2016). We designate race as the sensitive feature. Following ProPublica's analysis (Larson et al., 2016b), we only consider a subset of offenders whose race (the sensitive feature) is either black or white. Recidivism rates for the two groups are shown in Table 6 in Appendix A. For modeling the classification task, we use the same set of features as used by ProPublica (Larson et al., 2016b). [12] After performing the filtering described above, we obtain $5,287$ subjects and 5 features.

The NYPD SQF dataset consists of $84,868$ pedestrians who were stopped in the year 2012 on the suspicion of having a weapon. The dataset also contains over 100 features (*e.g.*, gender, height, reason for stop) and a binary label which indicates whether (negative class) or not (positive class) a weapon was discovered. For our analysis, we consider the race to be the sensitive feature with values blacks and whites. The classes in this dataset are highly imbalanced (97% of subjects in positive class), and as a result, a logistic regression classifier classifies almost all data points into the positive class. To counter this imbalance, we subsample the dataset to have equal number of subjects from each class. Information about weapon discovery rate for both races in included in Tables 7 and 8 in Appendix A.

12. 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. Moreover, 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.

Algorithm 1: Baseline method for removing disparate mistreatment w.r.t. FPR.

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

Furthermore, for training the classifiers, we consider a similar set of features as Goel et al. (2015). After performing these two filtering steps, we obtain 5,832 subjects and 19 features.

Methods. In our experiments, we compare our approach to two baseline methods. More specifically, we consider the following methods:

- ***Our method***: Implements our scheme to avoid disparate treatment and disparate mistreatment *simultaneously*. Disparate mistreatment is avoided by using fairness constraints on false positive and / or false negative rates. Disparate treatment is avoided by ensuring that sensitive attribute information is not used while making decisions, *i.e.*, by keeping user feature vectors ($\boldsymbol{x}$) and the sensitive features (z) disjoint.

- ***Our method$_{sen}$***: Implements our scheme to avoid disparate mistreatment only. The user feature vectors ($\boldsymbol{x}$) and the sensitive features (z) are not disjoint, that is, the classifier learns a non-zero weight for z. Therefore, the sensitive feature information is used for decision making, resulting in disparate treatment.

- ***Hardt et al.*** (Hardt et al., 2016): Operates by post-processing the outcomes of a possibly discriminatory classifier (logistic regression in this case) and using different decision thresholds for different sensitive feature value groups to remove disparate mistreatment. By construction, it needs the sensitive feature information while making decisions, and hence cannot avoid disparate treatment. This method is similar to the post-processing scheme discussed in Corbett-Davies et al. (2017).

- ***Donini et al.*** (Donini et al., 2018): This pre- / in-processing method aims to remove the discrepancy in the false negative rate of the two groups by adding constraints to a SVM classifier. We use the implementation provided by the authors[13] and train the method such that it has access to the sensitive feature information at decision time.

- ***Baseline***: Baseline introduced by us to facilitate a third comparison method. Tries to remove disparate mistreatment by introducing different penalties for misclassified

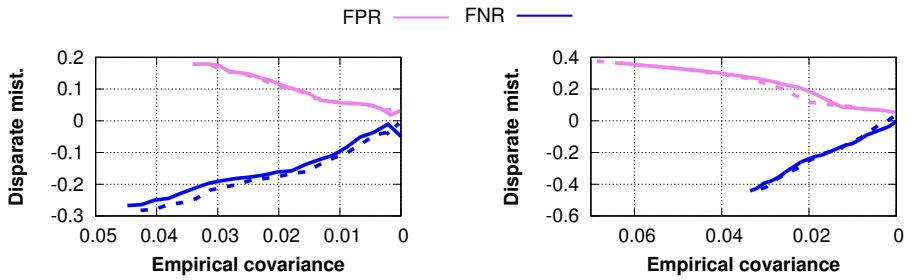
13. https://github.com/jmikko/fair_ERM

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 $DM_{FPR} > 0$ (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 $DM_{FPR} \leq \epsilon$). The algorithm is summarized in Figure 1, particularized to ensure fairness in terms of false positive rate. This process can be intuitively extended to account for fairness in terms of false negative rate or for *both* false positive rate and false negative rate. This method can be trained with or without using sensitive feature information while making decisions. We opt for the latter option.

Results. First, we experiment with a standard logistic regression classifier optimizing for accuracy on both datasets. For the COMPAS dataset, the (unconstrained) logistic regression classifier leads to an accuracy of 0.664. However, the classifier yields false positive rates of 0.35 and 0.17, respectively, for blacks and whites (*i.e.*, $DM_{FPR} = 0.18$), and false negative rates of 0.32 and 0.61 (*i.e.*, $DM_{FNR} = -0.29$). 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) subjects 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) subjects 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.

For the SQF data, the (unconstrained) logistic regression classifier leads to an accuracy of 0.751. However, the classifier yields false positive rates of 0.38 and 0.11, respectively, for blacks and whites (*i.e.*, $DM_{FPR} = 0.27$), and false negative rates of 0.19 and 0.31 (*i.e.*, $DM_{FNR} = -0.12$). Notice that unlike the COMPAS dataset, being classified positive in here is an advantageous outcome—positive class in this case is not being stopped whereas the positive class in the COMPAS dataset is being classified as being a recidivist. This scenario also resembles our synthetic example Case I in Section 5.1.2.

Next, we apply our framework on a logistic regression classifier to mitigate disparate mistreatment with respect to false positive rate, false negative rate, and on both. Figure 11 shows the link between the empirical decision boundary covariance when imposing the constraints to remove disparate mistreatment based on false positive rate or false negative rate on the ProPublica COMPAS and NYPD SQF datasets. The figure shows that: (i) a lower value of false positive or false negative rate covariance corresponds to a lower degree of disparate mistreatment with respect to the corresponding error rate, *i.e.*, Eqs. (5.2) and (5.3) respectively; and, (ii) a zero disparate mistreatment roughly corresponds to zero



ProPublica COMPAS data NYPD SQF data

Figure 11: [Disparate mistreatment constraints] The figure shows the effect of applying Our method$_{\text{sen}}$ to limit disparate mistreatment with respect to false positive rate and false negative rate, as defined by Eqs. (5.2) and (5.3)), respectively. Each type of constraint is applied separately. The plots show the correspondence between the empirical covariance (for each constraint type) and disparate mistreatment caused by the classifiers trained under fairness constraints on ProPublica COMPAS (left) and NYPD SQF (right) datasets. Solid lines correspond to the training data whereas dashed lines correspond to the test data. The figure shows that a decreasing empirical covariance leads to lower disparate mistreatment. Moreover, the statistics for training and test sets show a close correspondence, meaning that the constraints do not seem to be overfitting on the training data.

		FPR constraints			**FNR constraints**			**Both constraints**		
		Acc	**D_{FPR}**	**D_{FNR}**	**Acc**	**D_{FPR}**	**D_{FNR}**	**Acc**	**D_{FPR}**	**D_{FNR}**
ProPuclica COMPAS	**Our method$_{\text{sen}}$**	0.653	0.03	−0.10	0.656	−0.05	−0.01	0.654	−0.02	−0.03
	Baseline	0.631	0.01	−0.18	0.656	−0.03	−0.03	0.615	−0.19	0.13
	Hardt et al.	0.661	0.01	−0.08	0.654	−0.06	0.01	0.632	0.02	0.01
	Donini et al.	–	–	–	0.649	0.03	−0.09	–	–	–
NYPD SQF	**Our method**	0.633	0.06	−0.01	0.705	0.22	−0.07	0.642	0.05	0.04
	Our method$_{\text{sen}}$	0.727	0.08	0.07	0.743	0.18	0.00	0.726	0.07	0.07
	Baseline	0.527	0.02	−0.08	0.734	0.14	0.01	0.435	−0.71	0.95
	Hardt et al.	0.725	0.03	0.12	0.734	0.14	0.04	0.722	0.02	0.06
	Donini et al.	–	–	–	0.737	0.21	−0.04	–	–	–

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

covariance (though the correspondence is not completely strict on the ProPublica COMPAS data).

Next, we compare the performance of our scheme with the three alternative methods. While controlling for disparate mistreatment with respect to FPR and FNR simultaneously, the method of Hardt et al. can be interpreted as finding the optimal point that minimizes the loss on the average of the two group-conditional ROC curves (one curve for each sensitive feature group), or the one that minimizes the loss on the point-wise minimum of the two

curves. The optimal point in both cases lies on the point-wise minimum of the two curves. Both variants lead to similar performance, hence we report the results for the former.

Table 2 summarizes the results by showing the trade-off between fairness and accuracy achieved by our method, the methods by Hardt et al. and Donini et al., and the baseline. Similarly to the results in Section 5.1.2, we observe that for all four methods, controlling for disparate mistreatment on false positive rate (false negative rate) also helps decrease disparate mistreatment on false negative rate (false positive rate), at least to a limited extent. Moreover, our method and the methods by Hardt et al. and Donini et al. achieve similar accuracy for a given level of fairness when provided with the same amount of information (sensitive attribute information). The baseline on the other hand leads to the worst performance (especially on the NYPD SQF data.) We also note that the baseline tends to be somewhat unstable and fails to converge to a fair solution in some cases (*e.g.*, both FPR and FNR constraints on COMPAS and SQF datasets).

6. Related Work

6.1. Algorithmic decision making and evidence of unfairness

Algorithmic (or, automated) decision making has been used in applications involving human subjects since several decades. For example, previous works have studied the factors determining whether a criminal defendant would recidivate if granted bail (Burgess, 1928), or whether a loan applicant is likely to return their loan or not (Furletti, 2002). However, with the advent of complex learning methods, and convenient accessibility of 'big data' in several domains, automated decision making has permeated into a large number of human-centric applications, *e.g.*, job screening (Posse, 2016; Taylor, 2016), community safety (Perry, 2013; Lowenkamp, 2009; Kleinberg et al., 2017a), product personalization (Covington et al., 2016; Gomez-Uribe and Hunt, 2015) and online ad delivery (Google AdSense).

Several recent studies have shown that algorithmic decision making systems can potentially lead to (unintentional) unfairness against certain legally "protected" groups (Civil Rights Act, 1964). For example, Sweeney (2013) showed that Google ad platform is more likely to link typically African-American names with having criminal records as compared to typically white names. A recent study by ProPublica (Larson et al., 2016a) found that a criminal risk assessment tool used in Broward County, Florida was disproportionately marking African-American defendants as high risk (as compared to whites), even when they did not recidivate. Since the fairness of a decision making process is often a legally mandated criterion (Barocas and Selbst, 2016; Civil Rights Act, 1964), in the face of potential unfairness by automated decision making systems, a number of studies by regulatory authorities have called for taking fairness into account while designing algorithmic decision systems. (Podesta et al., 2014; Muñoz et al., 2016; Ramirez et al., 2016).

While automated decision making spans a variety of tasks including classification, ranking and recommendations, in this paper, we only focus on fairness in the context of classification. Next, we discuss the state-of-the-art in the context of the notions of unfairness discussed in Section 1.

6.2. Avoiding unfairness in classification

In this section, we will discuss techniques that aim to remove disparate treatment, disparate impact or disparate mistreatment from classification outcomes.

The first study on fair classification dates back to 2008 when Pedreschi et al. (2008) proposed techniques to avoid unfairness in classification rule mining. In the years that followed, a number of studies proposed techniques to remove unfairness from classification outcomes. Especially, last year or so has seen a flurry of methods proposed to control unfairness in classification. These studies operate by first specifying one or more measures of unfairness that they aim to control, *i.e.*, disparate treatment, disparate impact or disparate mistreatment, and then propose techniques to control for the selected measure(s).

These techniques can be divided into three different categories: *pre-processing*, *in-processing* and *post-processing*. Below, we discuss each of these categories separately.

6.2.1. Pre-processing

This technique consists of pre-processing the training data that would later be fed to a training algorithm Kamiran and Calders (2010); Luong et al. (2011); Feldman et al. (2015); Calmon et al. (2017). The goal is to pre-process the training data such that *any* classification algorithm trained on this data would generate unfairness-free outcomes. This strategy can be roughly divided into two different sub-categories. Below, we briefly discuss these subcategories:

The first sub-category involves changing the values of class labels for certain data points Kamiran and Calders (2010); Luong et al. (2011). For example, Kamiran and Calders (2010) propose a pre-processing technique that operates by first training an unconstrained classifier, and then moving / duplicating the data points from the group with lower acceptance rate (as compared to the other group) until the classification outcomes are free of disparate impact.

The second sub-category involves perturbing the non-sensitive features Feldman et al. (2015), or mapping the data to a transformed space Calmon et al. (2017). For example, building on ideas in the area of privacy-preserving data analysis (specifically t-closeness), Feldman et al. (2015) "repair" the non-sensitive features such that it is impossible to predict the sensitive features from non-sensitive features (which in turn means that the classifier trained on this data will not incur disparate impact), while ensuring that the resulting distribution is close to the original data distribution.

On the plus side, the pre-processing techniques have an advantage that the transformed dataset can be used to train any downstream algorithm.

However, these techniques also suffer from some disadvantages. First, since these techniques are not optimized for any specific classification model, and treat the learning algorithm as a black box, as a consequence, the pre-processing can lead to unpredictable loss in accuracy or may not remove unfairness on the test data (as we saw in Section 5.2.1). Furthermore, transforming the dataset might also affect the explainability of the classifier—*e.g.*, since the feature values were transformed during pre-processing, the feature weights of a linear classifiers might not be interpretable anymore.

Method	Type	DT	DI	DM	BN	Polyvalent sens.	Multiple sens.	Range of classifiers
Our framework	In	✓	✓	✓	✓	✓	✓	Any convex margin-based
Kamiran and Calders (2010)	Pre	✓	✓	✗	✗	✗	✗	Any score-based
Calders and Verwer (2010)	In/Post	✓	✓	✗	✗	✗	✗	Naive Bayes
Kamiran et al. (2010)	In	✓	✓	✗	✗	✗	✗	Decision tree
Luong et al. (2011)	Pre	✓	✗	✗	✗	✗	✗	Any
Kamishima et al. (2011)	In	✗	✓	✗	✗	✗	✗	Logistic regression
Zemel et al. (2013)	Pre/In	✓	✓	✗	✗	✗	✗	Log loss
Feldman et al. (2015)	Pre	✓	✓	✗	✗	✓	✓	Any (only numerical features)
Edwards and Storkey (2016)	Pre/In	✓	✓	✗	✗	✓	✓	MLPs
Goh et al. (2016)	In	✓	✓	✓	✗	✓	✓	Ramp loss
Hardt et al. (2016)	Post	✗	✗	✓	✗	✓	✓	Any score-based
Corbett-Davies et al. (2017)	Post	✗	✓	✓	✗	✓	✓	Any score-based
Woodworth et al. (2017)	In	✗	✗	✓	✗	✗	✗	Any convex linear
Quadrianto and Sharmanska (2017)	In	✓	✓	✓	✗	✗	✗	Hinge loss
Calmon et al. (2017)	Pre	✓	✓	✗	✗	✓	✓	Any
Dwork et al. (2018)	In/Post	✗	✓	✓	✗	✓	✓	Any score-based
Menon and Williamson (2018)	Post	✗	✓	✓	✗	✗	✗	Any score-based
Madras et al. (2018)	Pre/In	✓	✓	✓	✗	✗	✗	MLPs
Agarwal et al. (2018)	In/Post	✓	✓	✓	✗	✓	✓	Any score-based
Donini et al. (2018)	Pre/In	✓	✗	✓	✗	✓	✓	SVM

Table 3: Capabilities of different methods in mitigating disparate treatment (DT), disparate impact (DI) and disparate mistreatment (DM). We also show the type of each method: pre-processing (pre), in-processing (in) and post-processing (post). None of the prior methods addresses disparate impact's business necessity (BN) clause. Many of the methods do not generalize to multiple (*e.g.*, gender and race) or polyvalent sensitive features (*e.g.*, race, that has more than two values). The strategy by Feldman et al. (2015) is limited to only numerical non-sensitive features.

6.2.2. In-processing

The second strategy consists of modifying the training procedure of the classifier. Examples of this scheme include Calders and Verwer (2010); Kamishima et al. (2011); Goh et al. (2016); Woodworth et al. (2017); Kamiran et al. (2010); Quadrianto and Sharmanska (2017). Our proposed covariance constraints also fall under this category.

For example, the technique by Kamishima et al. (2011)—which is only limited to a logistic regression classifier—works by adding a regularization term in the objective that penalizes the mutual information between the sensitive feature and the classifier decisions. The method of Kamiran et al. (2010), which is limited to a decision tree classifier, operates by changing the splitting or the leaf node labeling criterion of the tree learning phase to remove disparate impact.

Goh et al. (2016), Woodworth et al. (2017) and Quadrianto and Sharmanska (2017) on the other hand suggest adding constraints similar to ours to the classification model. However, their works are only limited to a single specific loss function (Goh et al., 2016; Quadrianto and Sharmanska, 2017) or to a single notion of unfairness (Woodworth et al., 2017).

Zemel et al. (2013), building on Dwork et al. (2012), combined pre-processing and in-processing by jointly learning a 'fair' representation of the data and the classifier parameters. The joint representation is learnt using a multi-objective loss function that ensures that (i) the resulting representations do not lead to disparate impact, (ii) the reconstruction loss from the original data and intermediate representations is small and (iii) the class label can be predicted with high accuracy. This approach has two main limitations: i) it leads to a non-convex optimization problem and does not guarantee optimality, and ii) the accuracy of the classifier depends on the dimension of the fair representation, which needs to be chosen rather arbitrarily. Inspired by Zemel et al. (2013), the methods of Edwards and Storkey (2016) and Madras et al. (2018) also aim at learning fair representations of the data.

6.2.3. Post-processing

The third and final strategy consists of post-processing the classifier scores such that the new outcomes contain no disparate impact or disparate mistreatment (Hardt et al., 2016; Corbett-Davies et al., 2017; Dwork et al., 2018; Menon and Williamson, 2018).

This approach usually involves learning different decision thresholds for a given score function to remove unfairness (specifically, disparate impact or disparate mistreatment). However, since these strategies require the sensitive feature information at the decision time, they cannot be used in cases where sensitive feature information is unavailable (*e.g.*, due to privacy reasons) or prohibited from being used due to disparate treatment laws (Barocas and Selbst, 2016).

Dwork et al. (2018) combine the in-processing and post-processing scheme by first training a number of classifiers for each group (with each classifier having different acceptance rate for the given group), and then selecting the group-conditional classifiers that minimize a certain loss function. The loss function is formulated as a combination of the loss in accuracy and a penalty term penalizing the deviation from the fairness criterion. Like Hardt et al. (2016) and Corbett-Davies et al. (2017), this method too requires access to the sensitive feature information at the decision time.

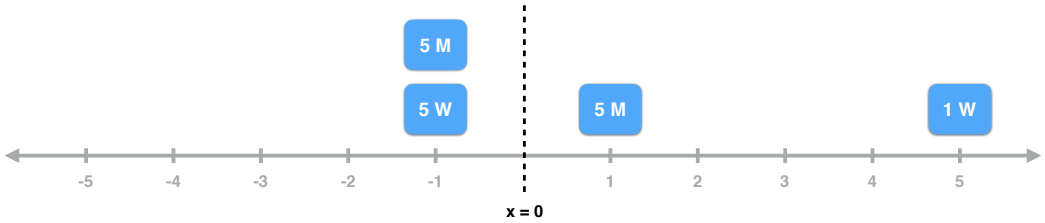


Figure 12: Covariance constraints may perform unfavorably in the presence of outliers. The figure shows a hypothetical dataset with just one feature (x) with values ranging from -5 to 5. Data points belong to two groups: men (M) or women (W). Each box shows the number of subjects of from a certain group (M or W) with that feature value. The decision boundary is at $x = 0$. The decision boundary covariance in this case is 0, yet the disparity in positive class outcome rates between men and women (0.5 for men and 0.17 for women) is very high. This situation is caused by one woman with feature value 5—this outlier point cancels out the effect of five normal examples (W with feature value -1) while computing the covariance.

In addition to the issues discussed above, most prior studies suffer from one or more of the following limitations: (i) they only accommodate a single, binary sensitive feature, (ii) they are restricted to a narrow range of classifiers, and, (iii) they cannot accommodate multiple unfairness notions *simultaneously*. Table 3 compares the capabilities of different methods in meeting different fairness criteria.

Finally, some recent studies (Kusner et al., 2017; Kilbertus et al., 2017) focus on detecting and removing unfairness by leveraging causal inference techniques. However, these studies often require access to causal graphs specifying causal relationships between different features, which can be quite challenging to obtain in practice.

7. Discussion, limitations and future work

In this work, we introduced a constraint-based framework to design fair margin-based classifiers that do not suffer from disparate treatment, disparate impact and disparate mistreatment. The key technical innovation of our framework is a general and intuitive measure of decision boundary unfairness, which serves as a tractable proxy for the above mentioned unfairness notions. Evaluation on various real-world datasets shows that our constrained-based framework is effective in limiting various forms of unfairness from classification outcomes, often at a small cost in terms of accuracy.

However, we also note that covariance mechanism only serves as a *proxy* for the unfairness measure under consideration (*i.e.*, disparate impact or disparate mistreatment). As a result, one may encounter cases where a zero value of the covariance proxy still results in a non-zero value of the unfairness measure. Such situations can arise due to various reasons. Below, we (non-exhaustively) list some of the reasons and leave a detailed formal analysis of such reasons for future work.

First, since our mechanism relies on empirically estimating the decision boundary covariance, very small presence of a certain group in the dataset can lead to poor estimate of the covariance and might not fully remove unfairness. Moreover, while the post-processing

schemes to remove unfairness (Hardt et al., 2016; Corbett-Davies et al., 2017) operate on the data of dimensionality 1 (that is, the scalar score assigned to each item by the classifier), our method operates by using all the features used in classification in order to compute the decision boundary covariance. As a result, our method is expected to suffer more from the data sparsity problem.

Second, we also notice that our method might not perform well in the presence of outliers. Consider for instance the example shown in Figure 12, where an outlier point causes the decision boundary covariance with respect to disparate impact (Eq. (4.2)) to be zero, even when the disparity in positive class outcomes caused by the corresponding decision boundary is very high. However, such outliers can in fact deteriorate the performance of any learning task (Bishop, 2006), even when no other constraints are applied, and one might wish to remove such outliers before training any classification model.

Third, while we observed that a decreasing covariance threshold corresponds to a more fair classifier (with respect to disparate impact or disparate mistreatment), the relation between the two is only empirically observed. A precise mapping between covariance and the precise value of the fairness notion under consideration is quite challenging to derive analytically, since it depends on the specific classifier and the dataset being used. Such a theoretical analysis would be an interesting future direction.

Our framework also opens many avenues for future work. For example, one could include fairness constraints in other supervised (*e.g.*, regression, recommendation) as well as unsupervised (*e.g.*, set selection, ranking) learning tasks. Our fairness constraints can be solved using standard convex or convex-concave optimizers (*e.g.*, SLSQP and DCCP, respectively) and we faced no scalability issues during the experiments conducted in this paper. However, extending these solvers to scale to cases when datasets or resulting optimization problems are too large to fit into memory would be an interesting future direction.

Appendix A. Additional dataset details

In this section, we show the distribution of sensitive features and class labels in the real-world datasets used in the evaluation (Section 5).

Gender	Low income (-ve)	High income (+ve)	Total
Males	20,988(69%)	9,539(21%)	30,527(100%)
Females	13,026(88%)	1,669(12%)	14,695(100%)
Total	34,014(75%)	11,208(25%)	45,222(100%)

Race	Low income (-ve)	High income (+ve)	Total
American-Indian/Eskimo	382(88%)	53(12%)	435(100%)
Asian/Pacific-Islander	934(72%)	369(28%)	1,303(100%)
White	28,696(74%)	10,207(26%)	38,903(100%)
Black	3,694(87%)	534(13%)	4,228(100%)
Other	308(87%)	45(13%)	353(100%)
Total	34,014(75%)	11,208(25%)	45,222(100%)

Table 4: [Adult dataset] High ($>$ 50K USD) and low income ($\leq$ 50K USD) rates in for gender and race groups.

Age	Yes (+ve)	No (-ve)	Total
$25 \leq$ age ≤ 60	3,970(10%)	35,240(90%)	39,210(100%)
age < 25 or age > 60	670(34%)	1,308(66%)	1,978(100%)
Total	4,640(11%)	36,548(89%)	41,188(100%)

Table 5: [Bank dataset] Term deposit subscription rates for the two race groups.

Race	Yes (+ve)	No (-ve)	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 6: [ProPublica COMPAS dataset] Recidivism rates both races.

Race	Yes (-ve)	No (+ve)	Total
Black	2,113(3%)	77,337(97%)	79,450
White	803(15%)	4,616(85%)	5,419
Total	2,916(3%)	81,953(97%)	84,869

Table 7: Persons found to be in possession of a weapon in 2012 NYPD SQF dataset (original).

Race	Yes (-ve)	No (+ve)	Total
Black	2,113(43%)	2,756(57%)	4,869
White	803(83%)	160(17%)	963
Total	2,916(50%)	2,916(50%)	5,832

Table 8: Persons found to be in possession of a weapon in 2012 NYPD SQF dataset (class-balanced).

References

Adult. http://tinyurl.com/UCI-Adult, 1996.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification. In *ICML*, 2018.

Andrew Altman. Discrimination. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Julia Angwin and Jeff 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.

Bank. http://tinyurl.com/UCI-Bank, 2014.

Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

Esha Bhandari. Big Data Can Be Used To Violate Civil Rights Laws, and the FTC Agrees, 2016. https://www.aclu.org/blog/free-future/big-data-can-be-used-violate-civil-rights-laws-and-ftc-agrees.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. 2004.

Ernest W. Burgess. Factors Determining Success or Failure on Parole. *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois*, 1928.

Toon Calders and Sicco Verwer. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*, 2010.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized Pre-Processing for Discrimination Prevention. In *NIPS*, 2017.

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

Civil Rights Act. Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *KDD*, 2017.

Paul Covington, Jay Adams, and Emre Sargin. Deep Neural Networks for YouTube Recommendations. In *RecSys*, 2016.

Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and Attacking the Saddle Point Problem in High-dimensional Non-convex Optimization. In *NIPS*, 2014.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical Risk Minimization under Fairness Constraints. In *NIPS*. 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, and Omer Reingold. Fairness Through Awareness. In *ITCSC*, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark DM Leiserson. Decoupled Classifiers for Group-fair and Efficient Machine Learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. In *ICLR*, 2016.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *KDD*, 2015.

FICO. FICO Score, 2017. https://en.wikipedia.org/wiki/Credit_score_in_the_United_States.

Anthony W. Flores, Christopher T. Lowenkamp, and Kristin 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.

Mark J. Furletti. An Overview and History of Credit Reporting, 2002. http://dx.doi.org/10.2139/ssrn.927487.

Alex Gano. Disparate impact and mortgage lending: A beginner's guide. *U. Colo. L. Rev.*, 88:1109, 2017.

Andrew Gelman, Jeffrey Fagan, and Alex Kiss. An analysis of the new york city police department's stop-and-frisk policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007.

James E Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of Computational Statistics: Concepts and Methods*. Springer Science & Business Media, 2012.

Sharad Goel, Justin M. Rao, and Ravi Shroff. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*, 2015.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael Friedlander. Satisfying Real-world Goals with Dataset Constraints. In *NIPS*, 2016.

Carlos A. Gomez-Uribe and Neil Hunt. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.*, 2015.

Google AdSense. https://www.google.com/adsense.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, 2016.

Faisal Kamiran and Toon Calders. Classifying without Discriminating. In *IC4*, 2009.

Faisal Kamiran and Toon Calders. Classification with No Discrimination by Preferential Sampling. In *BENELEARN*, 2010.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, 2010.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *ICML*, 2018.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. In *NIPS*. 2017.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. 2017a. Working paper.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, 2017b.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *NIPS*. 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. https://github.com/propublica/compas-analysis, 2016a.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016b.

Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Greg S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv preprint arXiv:1703.02442*, 2017.

Christopher T Lowenkamp. The Development of an Actuarial Risk Assessment Instrument for US Pretrial Services. *Fed. Probation*, 2009.

Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*, 2011.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *ICML*, 2018.

Aditya Krishna Menon and Robert C. Williamson. The Cost of Fairness in Classification. *arXiv:1705.09055*, 2017.

Aditya Krishna Menon and Robert C Williamson. The Cost of Fairness in Binary Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

Cecilia Muñoz, Megan Smith, and DJ Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*, 2016.

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware Data Mining. In *KDD*, 2008.

Walt L Perry. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation, 2013.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In *NIPS*. 2017.

John Podesta, Penny Pritzker, Ernest Moniz, John Holdren, and Jeffrey Zients. Big Data: Seizing Opportunities, Preserving Values. *Executive Office of the President. The White House.*, 2014.

Christian Posse. Cloud Jobs API: Machine Learning Goes to Work on Job Search and Discovery, 2016. https://cloud.google.com/blog/big-data/2016/11/cloud-jobs-api-machine-learning-goes-to-work-on-job-search-and-discovery.

Novi Quadrianto and Viktoriia Sharmanska. Recycling Privileged Learning and Distribution Matching for Fairness. In *NIPS*, 2017.

Edith Ramirez, Julie Brill, Maureen K. Ohlhausen, and Terrell McSweeny. Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. *Federal Trade Commission (FTC) Report*, 2016.

Cristóbal Romero and Sebastián Ventura. Preface to the Special Issue on Data Mining for Personalised Educational Systems. *User Modeling and User-Adapted Interaction*, 2011.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. Disciplined Convex-Concave Programming. *arXiv:1604.02639*, 2016.

Reva B Siegel. Race-Conscious but Race-Neutral: The Constitutionality of Disparate Impact in the Roberts Court. *Ala. L. Rev.*, 66:653, 2014.

Stop, Question and Frisk Data. http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page, 2017.

Latanya Sweeney. Discrimination in Online Ad Delivery. *ACM Queue*, 2013.

Tess Taylor. Recruiting will be an elementary task for Watson, says IBM, 2016. http://www.hrdive.com/news/recruiting-will-be-an-elementary-task-for-watson-says-ibm/427692/.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In *COLT*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017b.

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