From Parity to Preference-based Notions of Fairness in Classification

Muhammad Bilal Zafar\textsuperscript{1}, Isabel Valera\textsuperscript{2}, Manuel Gomez Rodriguez\textsuperscript{1}, Krishna P. Gummadi\textsuperscript{1}, and Adrian Weller\textsuperscript{2,3,4}

\textsuperscript{1}Max Planck Institute for Software Systems (MPI-SWS), \textsuperscript{2}University of Cambridge, \textsuperscript{3}Alan Turing Institute, \textsuperscript{4}Leverhulme Centre for the Future of Intelligence

Abstract

The adoption of automated, data-driven decision making in an ever expanding range of applications has raised concerns about its potential unfairness towards certain social groups. In this context, a number of recent studies have focused on defining, detecting, and removing unfairness from data-driven decision systems. However, the existing notions of fairness, based on parity (equality) in treatment or outcomes for different social groups, tend to be quite stringent, limiting the overall decision making accuracy. In this paper, we draw inspiration from the fair-division and envy-freeness literature in economics and game theory and propose preference-based notions of fairness—given the choice between various sets of decision treatments or outcomes, any group of users would collectively prefer its treatment or outcomes, regardless of the (dis)parity as compared to the other groups. Then, we introduce tractable proxies to design convex margin-based classifiers that satisfy these preference-based notions of fairness. Finally, we experiment with a variety of synthetic and real-world datasets and show that preference-based fairness allows for greater decision accuracy than parity-based fairness.

1 Introduction

As machine learning is increasingly being used to automate decision making in domains that affect human lives (e.g., credit ratings, housing allocation, recidivism risk prediction), there are growing concerns about the potential for unfairness in such algorithmic decisions [20, 22]. A flurry of recent research on fair learning has focused on defining appropriate notions of fairness and then designing mechanisms to ensure fairness in automated decision making [11, 12, 15, 16, 17, 18, 27, 28, 29].

Existing notions of fairness in the machine learning literature are largely inspired by the concept of discrimination in social sciences and law. These notions call for parity (i.e., equality) in treatment, in impact, or both. To ensure parity in treatment (or treatment parity), decision making systems need to avoid using users’ sensitive attribute information, i.e., avoid using the membership information in socially salient groups (e.g., gender, race), which are protected by anti-discrimination laws [4, 9]. As a result, the use of group-conditional decision making systems is often prohibited. To ensure parity in impact (or impact parity), decision making systems need to avoid disparity in the fraction of users belonging to different sensitive attribute groups (e.g., men, women) that receive beneficial decision outcomes. A number of learning mechanisms have been proposed to achieve parity in treatment [11, 21], parity in impact [7, 15, 18] or both [12, 14, 17, 27, 28, 29].
Figure 1: A fictitious decision making scenario involving two groups: men (M) and women (W). Feature $f_1$ (x-axis) is highly predictive for women whereas $f_2$ (y-axis) is highly predictive for men. Green (red) quadrants denote the positive (negative) class. Within each quadrant, the points are distributed uniformly and the numbers in parenthesis denote the number of subjects in that quadrant. The left panel shows the optimal classifier satisfying parity in treatment. This classifier leads to all the men getting classified as negative. The middle panel shows the optimal classifier satisfying parity in impact (in addition to parity in treatment). This classifier achieves impact parity by misclassifying women from positive class into negative class, and in the process, incurs a significant cost in terms of accuracy. The right panel shows a classifier consisting of group-conditional classifiers for men (purple) and women (blue). Both the classifiers satisfy the preferred treatment criterion since for each group, adopting the other group’s classifier would lead to a smaller fraction of beneficial outcomes. Additionally, this group-conditional classifier is also a preferred impact classifier since both groups get more benefit as compared to the impact parity classifier and the overall accuracy is better.

However, these mechanisms pay a significant cost in terms of the accuracy (or utility) of their predictions. In fact, there exist some inherent tradeoffs (both theoretical and empirical) between achieving high prediction accuracy and satisfying treatment and/or impact parity [8, 10, 13, 19].

In this work, we introduce, formalize and evaluate new notions of fairness that are inspired by the concepts of fair division and envy-freeness in economics and game theory [5, 26]. Our work is motivated by the observation that, in certain decision making scenarios, the existing parity-based fairness notions may be too stringent, precluding more accurate decisions, which may also be desired by every sensitive attribute group. To relax these parity-based notions, we introduce the concept of a user group’s preference for being assigned one set of decision outcomes over another. Given the choice between various sets of decision outcomes, any group of users would collectively prefer the set that contains the largest fraction (or the greatest number) of beneficial decision outcomes for that group. ¹ More specifically, our new preference-based notions of fairness, which we formally define in the next section, use the concept of user group’s preference as follows:

— From Parity Treatment to Preferred Treatment: To offer preferred treatment, a decision making system should ensure that every sensitive attribute group (e.g., men and women) prefers the set of decisions they receive over the set of decisions they would have received had they collectively presented themselves to the system as members of a different sensitive group.

The preferred treatment criterion represents a relaxation of treatment parity. That is, every decision making system that achieves treatment parity also satisfies the preferred treatment condition,

¹ Although it is quite possible that certain individuals from the group may not prefer the set that maximizes the benefit for the group as a whole.
which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred treatment condition is at least as high as the one achieved under treatment parity. Additionally, preferred treatment allows group-conditional decision making (not allowed by treatment parity), which is necessary to achieve high decision accuracy in scenarios when the predictive power of features varies greatly between different sensitive user groups, as shown in Figure 1.

While preferred treatment is a looser notion of fairness than treatment parity, it retains a core fairness property embodied in treatment parity, namely, envy-freeness at the level of user groups. Under preferred treatment, no group of users (e.g., men or women, blacks or whites) would feel that they would be collectively better off by switching their group membership (e.g., gender, race). Thus, preferred treatment decision making, despite allowing group-conditional decision making, is not vulnerable to being characterized as “reverse discrimination” against, or "affirmative action" for certain groups.

— From Parity Impact to Preferred Impact: To offer preferred impact, a decision making system needs to ensure that every sensitive attribute group (e.g., men and women) prefers the set of decisions they receive over the set of decisions they would have received under the criterion of impact parity.

The preferred impact criterion represents a relaxation of impact parity. That is, every decision making system that achieves impact parity also satisfies the preferred impact condition, which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred impact condition is at least as high as the one achieved under impact parity. Additionally, preferred impact allows disparity in benefits received by different groups, which may be justified in scenarios where insisting on impact parity would only lead to a reduction in the beneficial outcomes received by one or more groups, without necessarily improving them for any other group. In such scenarios, insisting on impact parity can additionally lead to a reduction in the decision accuracy, creating a case of tragedy of impact parity with a worse decision making all round, as shown in Figure 1.

While preferred impact is a looser notion of fairness compared to impact parity, by guaranteeing that every group receives at least as many beneficial outcomes as they would have received under impact parity, it retains the core fairness gains in beneficial outcomes that discriminated groups would have achieved under the fairness criterion of impact parity.

Finally, we note that our preference-based fairness notions, while having may attractive properties, are not the most suitable notions of fairness in all scenarios. In certain cases, parity fairness may well be the eventual goal [3] and the more desirable notion.

In the remainder of this paper, we formalize our preference-based fairness notions in the context of binary classification (Section 2), propose tractable and efficient proxies to include these notions in the formulations of convex margin-based classifiers in the form of convex-concave constraints (Section 3), and show on several real world datasets that our preference-based fairness notions can provide significant gains in overall decision making accuracy as compared to parity-based fairness (Section 4).

2 Defining preference-based fairness for classification

In this section, we will first introduce two useful quality metrics—utility and group benefit—in the context of fairness in classification, then revisit parity-based fairness definitions in the light of these quality metrics, and finally formalize the two preference-based notions of fairness introduced Section 1 from the perspective of the above metrics. For simplicity, we consider binary classification
tasks, however, the definitions can be easily extended to m-ary classification.

**Quality metrics in fair classification.** In a fair (binary) classification task, one needs to find a mapping between the user feature vectors \( \mathbf{x} \in \mathbb{R}^d \) and class labels \( y \in \{-1, 1\} \), where \((\mathbf{x}, y)\) are drawn from an (unknown) distribution \( f(\mathbf{x}, y) \). This is often achieved by finding a mapping function \( \theta : \mathbb{R}^d \to \mathbb{R} \) such that given a feature vector \( \mathbf{x} \) with an unknown label \( y \), the corresponding classifier predicts \( \hat{y} = \text{sign}(\theta(\mathbf{x})) \). However, this mapping function also needs to be fair with respect to the values of a user sensitive attribute \( z \in \mathbb{Z} \subseteq \mathbb{Z}_{\geq 0} \) (e.g., sex, race), which are drawn from an (unknown) distribution \( f(z) \) and may be dependent of the feature vectors and class labels, i.e., \( f(\mathbf{x}, y, z) = f(\mathbf{x}, y|z)f(z) \neq f(\mathbf{x}, y)f(z) \).

Given the above problem setting, we introduce the following quality metrics, which we will use to define and compare different fairness notions:

I. **Utility (\( U \))**: overall profit obtained by the decision maker using the classifier. For example, in a loan approval scenario, the decision maker is the bank that gives the loan and the utility can be the overall accuracy of the classifier, i.e.:

\[
U(\theta) = \mathbb{E}_{\mathbf{x}, y}[\mathbb{I}\{\text{sign}(\theta(\mathbf{x})) = y\}],
\]

where \( \mathbb{I}(\cdot) \) denotes the indicator function and the expectation is taken over the distribution \( f(\mathbf{x}, y) \). It is in the decision maker’s interest to use classifiers that maximize utility. Moreover, depending on the scenario, one can attribute different profit to true positives and true negatives—or conversely, different cost to false negatives and false positives—while computing utility. For example, in the loan approval scenario, marking an eventual defaulter as non-defaulter may have a higher cost than marking a non-defaulter as defaulter. For simplicity, we assume in this paper that both the costs are the same.

II. **Group benefit (\( B_z \))**: the fraction of beneficial outcomes received by users sharing a certain value of the sensitive attribute \( z \) (e.g., blacks, hispanics, whites). For example, in a loan approval scenario, the beneficial outcome for a user may be receiving the loan and the group benefit for each value of \( z \) can be defined as:

\[
B_z(\theta) = \mathbb{E}_{\mathbf{x}|z}[\mathbb{I}\{\text{sign}(\theta(\mathbf{x})) = 1\}],
\]

where the expectation is taken over the conditional distribution \( f(\mathbf{x}|z) \) and the bank offers a loan to a user if \( \text{sign}(\theta(\mathbf{x})) = 1 \). Moreover, as suggested by some recent studies in fairness-aware learning [15, 19, 28], the group benefits can also be defined as the fraction of beneficial outcomes conditional on the true label of the user. For example, in a recidivism prediction scenario, the group benefits can be defined as the fraction of eventually non-offending defendants sharing a certain sensitive attribute value getting bail, that is:

\[
B_z(\theta) = \mathbb{E}_{\mathbf{x}|z,y=1}[\mathbb{I}\{\text{sign}(\theta(\mathbf{x})) = 1\}],
\]

where the expectation is taken over the conditional distribution \( f(\mathbf{x}|z, y = 1) \), \( y = 1 \) indicates that the defendant does not re-offend, and bail is granted if \( \text{sign}(\theta(\mathbf{x})) = 1 \).

**Parity-based fairness.** A number of recent studies [7, 12, 15, 18, 27, 28, 29] have considered a classifier to be fair if it satisfies the impact parity criterion. That is, it ensures that the group benefits for all the sensitive attribute values are equal, i.e.,

\[
B_z(\theta) = B_{z'}(\theta) \quad \text{for all } z, z' \in \mathbb{Z}.
\]
In this context, different definitions of beneficial outcome have lead to different naming, e.g., disparate impact [12, 28], statistical parity [19, 29], disparate mistreatment [27], or equality of opportunity [15]. However, all of these benefit definitions invariably focus on achieving impact parity.

Although not always explicitly sought, most of the above studies propose classifiers that also satisfy treatment parity in addition to impact parity, i.e., they do not use the sensitive attribute $z$ in the decision making process. However, some of them [7, 15, 18] do not satisfy treatment parity since they resort to group-conditional classifiers, i.e., $\theta = \{\theta_z\}_{z \in \mathcal{Z}}$. In such case, we can rewrite the above parity condition as:

$$B_z(\theta_z) = B_{z'}(\theta_{z'}) \text{ for all } z, z' \in \mathcal{Z}. \tag{2}$$

**Fairness beyond parity.** Given the above quality metrics, we can now formalize the two preference-based fairness notions introduced in Section 1.

— **Preferred treatment:** if a classifier $\theta$ resorts to group-conditional classifiers, i.e., $\theta = \{\theta_z\}_{z \in \mathcal{Z}}$, it is a preferred treatment classifier if each group sharing a sensitive attribute value $z$ benefits more from its corresponding group-conditional classifier $\theta_z$ than it would benefit if it would be classified by any of the other group-conditional classifiers $\theta_{z'}$, i.e.,

$$B_z(\theta_z) \geq B_z(\theta_{z'}) \text{ for all } z, z' \in \mathcal{Z}. \tag{3}$$

Note that, if a classifier $\theta$ does not resort to group-conditional classifiers, i.e., $\theta_z = \theta$ for all $z \in \mathcal{Z}$, it will be always be a preferred treatment classifier. If, in addition, such classifier ensures impact parity, it is easy to show that its utility cannot be larger than a preferred treatment classifier consisting of group-conditional classifiers.

— **Preferred impact:** a classifier $\theta$ offers preferred impact over a classifier $\theta'$ ensuring impact parity if it achieves higher group benefit for each sensitive attribute value group, i.e.,

$$B_z(\theta) \geq B_z(\theta') \text{ for all } z \in \mathcal{Z}. \tag{4}$$

One can also rewrite the above condition for group-conditional classifiers, i.e., $\theta = \{\theta_z\}_{z \in \mathcal{Z}}$ and $\theta' = \{\theta'_z\}_{z \in \mathcal{Z}}$, as follows:

$$B_z(\theta_z) \geq B_z(\theta'_{z}) \text{ for all } z \in \mathcal{Z}. \tag{5}$$

Note that, given any classifier $\theta'$ ensuring impact parity, it is easy to show that there will always exist a preferred classifier $\theta$ with equal or higher utility.

**On individual-level preferences.** Notice that preferred treatment and preferred impact notions are defined based on the group preferences, i.e., whether a group as a whole prefers (or, gets more beneficial outcomes from) a given set of outcomes over another set. It is quite possible that a set of outcomes preferred by the group collectively is not preferred by certain individuals in the group. Consequently, one can extend these notions to account for individual preferences as well, i.e., a set of outcomes is preferred over another if all the individuals in the group prefer it. In the remainder of the paper, we focus on preferred treatment and preferred impact in the context of group preferences, and leave the case of individual preferences and its implications on the cost of achieving fairness to be explored thoroughly in a future study.
III. The group benefit

which is a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved using well-known heuristics such as the one proposed by Shen et al. [25]. For example, if we particularize the above formulation to group-conditional (standard) logistic regression classifiers $\theta_z$ and $\theta_z^T x$, and $L_2$-norm regularizer, then, Eq. 9 adopts the following form:

\[
\begin{align*}
\min_{\theta_z} & \quad -\frac{1}{N} \sum_{(x,y,z) \in D} \ell_\theta(x,y) + \sum_{z \in Z} \lambda_z \Omega(\theta_z) \\
\text{subject to} & \quad \sum_{x \in D_z} \max(0, \theta_z^T x) \geq \sum_{x \in D_z} \max(0, \theta_z^T x) \quad \forall z \in Z,
\end{align*}
\]

which is a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved using well-known heuristics such as the one proposed by Shen et al. [25]. For example, if we particularize the above formulation to group-conditional (standard) logistic regression classifiers $\theta_z$ and $\theta_z^T x$, and $L_2$-norm regularizer, then, Eq. 9 adopts the following form:

\[
\begin{align*}
\min_{\theta_z} & \quad -\frac{1}{N} \sum_{(x,y,z) \in D} \log p(y|x, \theta_z) + \sum_{z \in Z} \lambda_z ||\theta_z||^2 \\
\text{subject to} & \quad \sum_{x \in D_z} \max(0, \theta_z^T x) \geq \sum_{x \in D_z} \max(0, \theta_z^T x) \quad \forall z \in Z.
\end{align*}
\]

II. The utility function $U$ is defined as the overall accuracy of the group-conditional classifiers, i.e.,

\[
U(\theta) = \mathbb{E}_{x,y} \left[ \mathbb{I}\{\text{sign}(\theta(x)) = y\} \right] = \sum_{z \in Z} \mathbb{E}_{x,y,z} \left[ \mathbb{I}\{\text{sign}(\theta_z^T x) = y\} \right] f(z).
\] (6)

III. The group benefit $B_z$ for users sharing the sensitive attribute value $z$ is defined as their average probability of being classified into the positive class, i.e.,

\[
B_z(\theta) = \mathbb{E}_{x,z} \left[ \mathbb{I}\{\text{sign}(\theta(x)) = 1\} \right] = \mathbb{E}_{x,z} \left[ \mathbb{I}\{\text{sign}(\theta_z^T x) = 1\} \right].
\] (7)

Preferred impact classifiers. Given a impact parity classifier with decision boundary parameters $\{\theta_z\}_{z \in Z}$, one could think of finding the decision boundary parameters $\{\theta_z\}_{z \in Z}$ of a preferred classifier that maximizes utility by solving the following optimization problem:

\[
\begin{align*}
\min_{\theta_z} & \quad -\frac{1}{N} \sum_{(x,y,z) \in D} \mathbb{I}\{\text{sign}(\theta_z^T x) = y\} \\
\text{subject to} & \quad \sum_{x \in D_z} \mathbb{I}\{\text{sign}(\theta_z^T x) = 1\} \geq \sum_{x \in D_z} \mathbb{I}\{\text{sign}(\theta_z^T x) = 1\} \quad \forall z \in Z,
\end{align*}
\] (8)

where $D_z = \{(x_i, y_i, z_i) \in D | z_i = z\}$ denotes the set of users in the training set sharing sensitive attribute value $z$, the objective uses an empirical estimate of the utility, defined by Eq. 6, and the preferred impact constraints, defined by Eq. 5, use empirical estimates of the group benefits, defined by Eq. 7. Here, note that the right hand side of the inequalities does not contain any variables and can be precomputed, i.e., the impact parity classifiers $\{\theta_z\}_{z \in Z}$ are given.

Unfortunately, it is very challenging to solve the above optimization problem since, for nontrivial linear-classifiers (e.g., logistic regression, SVMs), both the objective and constraints are non-convex. To overcome this difficulty, we minimize instead a convex loss function $\ell_\theta(x,y)$, which is classifier dependent [6], and approximate the group benefits using a ramp (convex) function $r(x) = \max(0, x)$, i.e.,

\[
\begin{align*}
\min_{\theta_z} & \quad -\frac{1}{N} \sum_{(x,y,z) \in D} \ell_\theta(x,y) + \sum_{z \in Z} \lambda_z \Omega(\theta_z) \\
\text{subject to} & \quad \sum_{x \in D_z} \max(0, \theta_z^T x) \geq \sum_{x \in D_z} \max(0, \theta_z^T x) \quad \forall z \in Z,
\end{align*}
\] (9)

which is a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved using well-known heuristics such as the one proposed by Shen et al. [25]. For example, if we particularize the above formulation to group-conditional (standard) logistic regression classifiers $\theta_z$ and $\theta_z^T x$, and $L_2$-norm regularizer, then, Eq. 9 adopts the following form:

\[
\begin{align*}
\min_{\theta_z} & \quad -\frac{1}{N} \sum_{(x,y,z) \in D} \log p(y|x, \theta_z) + \sum_{z \in Z} \lambda_z ||\theta_z||^2 \\
\text{subject to} & \quad \sum_{x \in D_z} \max(0, \theta_z^T x) \geq \sum_{x \in D_z} \max(0, \theta_z^T x) \quad \forall z \in Z.
\end{align*}
\] (10)

\(^2\)Exploring the relaxations of these assumptions is a very interesting avenue for future work.
where $p(y = 1 | x, \theta_z) = \frac{1}{1+e^{-\theta_z^T x}}$. One can similarly particularize the formulation for other convex boundary-based classifiers like squared loss, linear / non-linear SVMs, etc.

**Preferred treatment classifiers.** One could think of finding the decision boundary parameters $\{\theta_z\}_{z \in Z}$ of a preferred treatment classifier that maximizes utility by solving the following optimization problem:

$$\min_{\{\theta_z\}} -\frac{1}{N} \sum_{(x,y,z) \in D} \mathbb{I}\{\text{sign}(\theta_z^T x) = y\}$$

subject to

$$\sum_{x \in D_z} \mathbb{I}\{\text{sign}(\theta_z^T x) = 1\} \geq \sum_{x \in D_{z'}} \mathbb{I}\{\text{sign}(\theta_{z'}^T x) = 1\} \quad \text{for all } z, z' \in Z,$$

(11)

where $D_z = \{(x_i, y_i, z_i) \in D | z_i = z\}$ denotes the set of users in the training set sharing sensitive attribute value $z$, the objective uses an empirical estimate of the utility, defined by Eq. 6, and the preferred treatment constraints, defined by Eq. 3, use empirical estimates of the group benefits, defined by Eq. 7. Here, note that both the left and right hand side of the inequalities contain optimization variables.

However, the objective and constraints in the above problem are non-convex and thus we adopt a similar strategy as in the case of preferred impact classifiers. More specifically, we solve instead the following tractable problem:

$$\min_{\{\theta_z\}} -\frac{1}{N} \sum_{(x,y,z) \in D} \ell_{\theta_z}(x, y) + \sum_{z \in Z} \lambda_z \Omega(\theta_z)$$

subject to

$$\sum_{x \in D_z} \max(0, \theta_z^T x) \geq \sum_{x \in D_{z'}} \max(0, \theta_{z'}^T x) \quad \text{for all } z \in Z,$$

(12)

which is also a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved.

### 4 Evaluation

In this section, we compare the performance of preferred treatment and preferred impact classifiers against unconstrained, treatment parity and impact parity classifiers on a variety of synthetic and real-world datasets. More specifically, we consider the following classifiers, which we train to maximize utility subject to the corresponding constraints:

- **Uncons:** an unconstrained classifier that resorts to group-conditional classifiers. It violates treatment parity—it trains a separate classifier per sensitive attribute group value—and potentially violates impact parity—it may lead to different benefits for different groups.

- **Parity:** a parity classifier that does not use the sensitive attribute group information in the decision making, but only during the training phase, and is constrained to satisfy both treatment parity—its decisions do not change based on the users’ sensitive attribute value as it does not resort to group-conditional classifiers—and impact parity—it ensures that the benefits for all groups are the same. We train this classifier using the methodology proposed by Zafar et al. [28].

- **Preferred treatment:** a classifier that resorts to group-conditional classifiers and is constrained to satisfy preferred treatment—each group gets the highest benefit with its own classifier than any other group’s classifier.

- **Preferred impact:** a classifier that resorts to group-conditional classifiers and is constrained to be preferred over the Parity classifier.
— **Preferred both**: a classifier that resort to group-conditional classifiers and is constrained to satisfy both preferred treatment and preferred impact.

In all our experiments, we use logistic regression classifiers with $L_2$-norm regularization, randomly split the corresponding dataset into 50%-50% train-test folds 5 times, and report the average accuracy and group benefits in the test folds. Here, we compute utility ($U$) as the overall accuracy of a classifier and group benefits ($B_z$) as the fraction of users sharing sensitive attribute $z$ that are classified into the positive class. Moreover, the sensitive attribute is always binary, i.e., $z \in \{0, 1\}$.

### 4.1 Experiments on synthetic data

**Experimental setup.** Following Zafar et al. [28], we generate a synthetic dataset in which the unconstrained classifier (Uncons) offers different benefits to each sensitive attribute group. In particular, we generate 20,000 binary class labels $y \in \{-1, 1\}$ uniformly at random along with their corresponding two-dimensional feature vectors sampled from the following Gaussian distributions: $p(x|y = 1) = N([2; 2], [5, 1; 1, 5])$ and $p(x|y = -1) = N([-2; -2], [10, 1; 1, 3])$. Then, we generate each sensitive attribute from the Bernoulli distribution $p(z = 1) = p(x'|y = 1)/(p(x'|y = 1)+p(x'|y = -1))$, where $x'$ is a rotated version of $x$, i.e., $x' = [\cos(\pi/8), -\sin(\pi/8); \sin(\pi/8), \cos(\pi/8)]$. Finally, we train the five classifiers described above and compute their overall (test) accuracy and (test) group benefits.

**Results.** Figure 2 shows the trained classifiers, along with their overall accuracy and group benefits. We can make several interesting observations:

The Uncons classifier leads to a reasonable accuracy of 0.87, however, the group-conditional boundaries and high disparity in treatment for the two groups (0.18 vs. 0.84) mean that it satisfies neither treatment parity nor impact parity. Moreover, although it satisfies preferred treatment—neither of the groups would get higher benefit by adopting the classifier of the other group—this will not always be the case, as we will later show in the experiments on real data.
The **Parity** classifier satisfies both treatment and impact parity, however, it does so at a large cost in terms of accuracy, which drops from 0.87 for **Uncons** to 0.58 for **Parity**.

The **Preferred impact** classifier, by making use of the relaxed parity-fairness conditions, provides higher benefits for both groups at a much smaller cost in terms of accuracy than the **Parity** classifier (0.78 vs. 0.58). Note that, while the **Parity** classifier achieved equality in benefits by misclassifying *negative examples from group-0* into the positive class and misclassifying *positive examples from group-1* into the negative class, the **Preferred impact** classifier only incurs the former type of misclassifications. However, the outcomes of the **Preferred impact** classifier do not satisfy the preferred treatment criterion: group-1 would attain higher benefit if it used the classifier of group-0 (0.95 as compared to 0.84).

Finally, the classifier that satisfies preferred treatment and preferred impact (**Preferred both**) achieves an accuracy and benefits at par with the **Preferred impact** classifier. However, this will not be always the case, as we will later show in the experiments on real data.

### 4.2 Experiments on real data

**Dataset description and experimental setup.** We experiment with three real-world datasets: the COMPAS recidivism prediction dataset compiled by ProPublica [20], the Adult income dataset from UCI machine learning repository [2], and the New York Police Department (NYPD) Stop-question-and-frisk (SQF) dataset made publicly available by NYPD [1]. These datasets have been used by a number of prior studies in the fairness-aware machine learning literature [12, 24, 27, 29, 28].

In the COMPAS dataset, the classification task is to predict whether a criminal defendant would recidivate within two years (negative class) or not (positive class); in the Adult dataset, the task is to predict whether a person earns more than 50K USD per year (positive class) or not; and, in SQF dataset, the task is to predict whether a pedestrian should be stopped on the suspicion of having an illegal weapon or not (positive class). In all datasets, we assume being classified as positive to be the beneficial outcome. Additionally, we divide the subjects in each dataset into two sensitive attribute value groups: women (group-0) and men (group-1) in the Adult dataset and blacks (group-0) and whites (group-1) in the COMPAS and SQF datasets. The supplementary material (Appendix A) contains more information on the sensitive and the non-sensitive features as well as the class distributions.

**Results.** Figure 3 shows the accuracy achieved by the five classifiers described above along with the benefits they provide for the three datasets. We can draw several interesting observations:

In all cases, the **Uncons** classifier, in addition to violating treatment parity (a separate classifier for each group) and impact parity (high disparity in group benefits), also violates the preferred treatment criterion (in all cases, at least one of group-0 or group-1 would benefit more by adopting the other group’s classifier). On the other hand, the **Parity** classifier satisfies the treatment parity and impact parity but it does so at a large cost in terms of accuracy.

---

3Since the SQF dataset is highly skewed in terms of class distribution (97% points in the positive class) resulting in a trained classifier predicting all points as positive (yet having 97% accuracy), we sub-sample the dataset to have equal class distribution. Another option was to use penalties proportional to the size of the class, but we observe that an unconstrained classifier with class penalties gives similar predictions as compared to a sub-sampled dataset. We decided to experiment with the sub-sampled dataset since the accuracy drops in this dataset are easily interpretable.

4The unfairness in the SQF dataset is different from what one would expect [23]—an unconstrained classifier gives more benefits to blacks as compared to whites. This is due to the fact that a larger fraction of stopped whites were found to be in possession on an illegal weapon (Tables 3 and 4 in Appendix A).
The **Preferred treatment** classifier provides a much higher accuracy than the **Parity** classifier—its accuracy is at par with that of the **Uncons** classifier—while satisfying the preferred treatment criterion. However, it does not meet the preferred impact criterion. The **Preferred impact** classifier meets the preferred impact criterion but does not satisfies preferred treatment. Moreover, it also leads to a better accuracy then **Parity** classifier in all cases. However, the gain in accuracy is more substantial in COMPAS and SQF datasets as compared to the Adult dataset.

The classifier satisfying preferred treatment and preferred impact (**Preferred both**) has an underwhelming performance in terms of accuracy. While the performance of this classifier is still better than the **Parity** classifier in the COMPAS dataset, it is worse for the other two datasets. In fact, for the SQF dataset, both preferred treatment and preferred impact constraints are only satisfied by classifying nearly all the users into the positive class, essentially leading to a **trivial** classifier.
In summary, the above results show that ensuring either preferred treatment or preferred impact is less costly in terms of accuracy loss than ensuring parity-based fairness, however, ensuring both preferred treatment and preferred impact can lead to great accuracy loss in certain datasets. We hypothesize that this loss in accuracy may be partly due to splitting the number of available samples into groups during training—each group-conditional classifier use only samples from the corresponding sensitive group—hence decreasing the effectiveness of empirical risk minimization.

5 Conclusion

In this paper, we introduced two preference-based notions of fairness—preferred treatment and preferred impact—establishing a previously unexplored connection between fair machine learning, and the economics and game theory concepts of fair-division and envy-freeness. Then, we proposed tractable proxies to design convex margin-based classifiers satisfying these fairness notions and experimented with a variety of synthetic and real-world datasets, showing that preference-based fairness often allows for greater decision accuracy than existing parity-based fairness notions.

Preference-based fairness have many attractive properties, however, we acknowledge it may not always be the most appropriate notion, e.g., in some scenarios, parity-based fairness may very well present the eventual goal and be more desirable [3].

Acknowledgments

AW acknowledges support by the Alan Turing Institute under EPSRC grant EP/N510129/1, and by the Leverhulme Trust via the CFI.

References

[1] Stop, Question and Frisk Data. http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page, 2017.
[2] Adult data. https://archive.ics.uci.edu/ml/datasets/adult, 1996.
[3] A. Altman. Discrimination. In The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/win2016/entries/discrimination/.
[4] S. Barocas and A. D. Selbst. Big Data’s Disparate Impact. California Law Review, 2016.
[5] M. Berliant and W. Thomson. On the Fair Division of a Heterogeneous Commodity. Journal of Mathematics Economics , 1992.
[6] C. M. Bishop. Pattern recognition. Springer, 2006.
[7] T. Calders and S. Verwer. Three Naive Bayes Approaches for Discrimination-Free Classification. Data Mining and Knowledge Discovery, 2010.
[8] A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. arXiv preprint, arXiv:1610.07524, 2016.
[9] Civil Rights Act. Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.
[10] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic Decision Making and the Cost of Fairness. In KDD, 2017.
[11] C. Dwork, M. Hardt, T. Pitassi, and O. Reingold. Fairness Through Awareness. In *ITSCSC*, 2012.
[12] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and Removing Disparate Impact. In *KDD*, 2015.
[13] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv:1609.07236*, 2016.
[14] G. Goh, A. Cotter, M. Gupta, and M. Friedlander. Satisfying Real-world Goals with Dataset Constraints. In *NIPS*, 2016.
[15] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, 2016.
[16] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in Learning: Classic and Contextual Bandits. In *NIPS*, 2016.
[17] F. Kamiran and T. Calders. Classification with No Discrimination by Preferential Sampling. In *Proceedings of Machine Learning conference of Belgium and The Netherlands*, 2010.
[18] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.
[19] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, 2017.
[20] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. https://github.com/propublica/compas-analysis, 2016.
[21] B. T. Luong, S. Ruggieri, and F. Turini. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*, 2011.
[22] C. Muñoz, M. Smith, and D. Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*, 2016.
[23] NYCLU. Stop-and-frisk data. https://www.nyclu.org/en/stop-and-frisk-data, 2017.
[24] R. S. Sharad Goel, Justin M. Rao. Precinct or Prejudice? Understanding Racial Disparities in New York City’s Stop-and-Frisk Policy. *Annals of Applied Statistics*, 2015.
[25] X. Shen, S. Diamond, Y. Gu, and S. Boyd. Disciplined Convex-Concave Programming. *arXiv:1604.02639*, 2016.
[26] H. R. Varian. Equity, Envy, and Efficiency. *Journal of Economic Theory*, 1974.
[27] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*, 2017.
[28] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017.
[29] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In *ICML*, 2013.
A Dataset details

For the ProPublica COMPAS dataset, we use the same non-sensitive features as used by Zafar et al. [27]. The non-sensitive features include number of prior offenses, the degree of the arrest charge (misdemeanor or felony), etc. The class and sensitive attribute distribution in the dataset is in Table 1.

Table 1: Recidivism rates in ProPublica COMPAS data for both races.

| Race | Yes (-ve) | No (+ve) | Total |
|------|-----------|----------|-------|
| Black | 1,661(52%) | 1,514(48%) | 3,175 |
| White | 8,222(39%) | 1,281(61%) | 2,103 |
| Total | 2,483(47%) | 2,795(53%) | 5,278 |

For Adult dataset [2], we use the same non-sensitive features as a number of prior studies [12, 28, 29] on fairness-aware learning. The non-sensitive features include educational level of the person, number of working hours per week, etc. The class and sensitive attribute distribution in the dataset is as follows in Table 2.

Table 2: High income ($\geq$ 50K USD) in Adult data for both genders.

| Gender | Yes (+ve) | No (-ve) | Total |
|--------|-----------|----------|-------|
| Males  | 9,539(31%) | 20,988(69%) | 30,527 |
| Females| 1,669(11%)  | 13,026(89%) | 14,695 |
| Total  | 34,014(75%) | 11,208(25%) | 45,222 |

For the NYPD SQF dataset [1], we use the same prediction task and non-sensitive features as used by Goel et al. [24]. We only use the stops made in 2012. The prediction task is, whether a pedestrian stopped on the suspicion of having a weapon actually possesses a weapon or not. The non-sensitive features include proximity to a crime scene, age/build of a person, and so on. Finally, as explained in Section 4, since the original dataset (Table 3) is highly skewed towards the positive class we subsample the majority class (positive) to match the size of the minority (negative) class.

Table 3: Persons found to be in possession of a weapon in 2012 NYPD SQF dataset (original).

| 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 4: Persons found to be in possession of a weapon in 2012 NYPD SQF dataset (class-balanced).

| 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 |