Mirror mirror on the wall... Who's the fairest of them all? *

Ng Keng Hwee | kenghwee@comp.nus.edu.sg

Abstract

Automated machine-learning (ML) systems are ubiquitously deployed in different domains of society, thus rendering the need for us to adopt a more critical viewpoint towards these solutions. In this paper, we will examine how ML classification systems can produce decisions which discriminate against certain salient social groups, where "discrimination" can be operationalized as disparity in the treatment, impact and mistreatment between these groups. After which, we proceed to understand the landscape of "fair machine learning", as we seek to train fairer ML classifiers which produce parity-based decisions. We then discuss another interesting companion to the parity-based fairness literature, which is the proposal of preference-based notions of fairness. The motivation for this preference-based notion is due to observations concluding that parity-based fairness notions are too stringent, and lead to underwhelming tradeoffs between fairness and accuracy. Drawing inspiration from envy-freeness literature in economics, we would hope to design Pareto-optimal [1] group-based classifiers for each group such that we can synthesise classifiers with high accuracy, while collectively satisfying all relevant social groups. In light of this multitude of different fairness notions, we hope to provide insights as to the applicability of these notions to different situations (thus the title). Finally, we experiment with a variety of synthetic and real-world datasets and understand these different fairness notions from a computational perspective, and summarise limitations as well as future research directions.

1 Introduction:

"I am a robot, ... just by reading the internet, and now I can write this column". [2] That was quoted from an article written entirely by GPT-3, a new language generator by OpenAI. GPT-3 is one of many innovations powered by artificial intelligence (A.I), which mainstream media platforms portray to be "omnipotent". This unfortunately undermines the darker, albeit equally relevant flipside of A.I, which is the potential of making unfair decisions. In this paper, we

*My reference papers are: [1]: "Fairness constraints: Mechanisms for fair classification", [2]: "Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification without Disparate Mistreatment", and [3]: "From Parity to Preference-based Notions of Fairness in Classification"

[1]The Pareto Optimal Solution refers to a solution, around which there is no way of improving any objective without degrading at least one other objective.

[2]Retrieved from "A robot wrote this entire article. Are you scared yet, human?", The Guardian. 8 Sep 2020

will be mainly discussing one popular notion of unfairness - "discrimination", which is the act of disadvantaging certain individuals based on their membership in some salient social group characterised by "sensitive attributes". [3] Notably, discrimination can disproportionately hurt or benefit particular groups of individuals, and this signifies a need for us to seek intuitive computational perspectives to better understand discrimination.

1.1 Contextualising the binary classification setting:

It is noteworthy to point out that the approaches illustrated in this paper are centred around the binary classification setting with margin-based classifiers: logistic regression (LR) and support vector machine (SVM) models. For a given unseen feature vector $\vec{x} \in \mathbb{R}^d$, we can train a margin-based classifier by minimizing a usually convex loss function, i.e: $\theta^* = \text{argmin}_\theta \, \mathsf{L}(\theta)$ over a training set: $\mathcal{D} : (\vec{x}, y)$ where $y \in \{-1, 1\}$. We can subsequently predict for an unseen sample from a testing set with a **class label** $\hat{y}$ to be 1 if $d_{\theta^*}(\vec{x})$, the **signed distance** between the feature vector $\vec{x}$ to the learned decision boundary is non-negative. Mathematically speaking, $\hat{y} = 1$ if $d_{\theta^*}(\vec{x}) \geq 0$, otherwise $\hat{y}$ = -1. Also, recall that we have discussed the concept of sensitive attributes in Section 1, and we term these attributes **z**. Without loss of generality, we assume that **z** is binary: i.e $z \in \{0, 1\}$. Note that the ideas discussed in this paper can be easily extended to m-ary classification with polyvalent sensitive attributes.

2 A computational perspective of unfairness in machine-learning:

After contextualizing our binary classification setting, we now seek to understand how to understand "discrimination" from a computational perspective. Essentially, a classifier is considered to be unfair, if there is the presence of **disparity** as summarised in the following notions highlighted in the below table, Table 1.

Notion	Mathematical Representation	Intuition
Disparate Treatment [1]	$P(\hat{y}\|x, z) \neq P(\hat{y}\|x)$	Disparity in the decision outcome if z changes, ceteris paribus.
Disparate Impact [1]	$P(\hat{y} = 1\|z = 0)$ $\neq P(\hat{y} = 1\|z = 1)$	Disparity in proportion of positive decisions for each z value.
Disparate Mistreatment [2] (Misclassification Rate)	$P(\hat{y} = y\|z = 0, y \in \{-1, 1\})$ $\neq P(\hat{y} = y\|z = 1, y \in \{-1, 1\})$	Disparity in overall misclassification rates, for each z value.

Table 1: A summary of the different operationalisable notions of "unfairness"

[3] These attributes were formalised in the Civil Rights Act, 1964

2.1 What causes discrimination in ML models?

We discuss two main reasons which explain the manifestation of the aforementioned disparity notions in our ML systems. Firstly, note that our classifiers are often trained on past historical data, which may contain sampling bias and labelling bias.[4] Hence, during the training phrase of the ML classifier, these biases would be "learnt" by the classifier. This consequently leads to the propagation of these erroneous biases throughout future predictions.

Secondly, the decision of using a particular loss function during the training phrase of the ML classifier may actually contain implicit bias in itself. This is because the chosen loss function may be just a utilitarian minimisation of the overall prediction errors, and disregards the proportion of misclassifications among various salient social groups. We therefore need to note the distinction that minimising the sum of overall errors made on an individual level is not equivalent to having equal misclassification rates across different social groups.

2.2 Why do we need different notions of unfairness?

In Table 1, we have introduced three widely-accepted notions of discrimination. We will seek to understand each of these notions individually, and also how they complement each other to eventually produce fairer classification systems based on different contexts.

2.2.1 The notion of disparate treatment:

Anti-discrimination laws established by the U.S Equal Employment Opportunity Commission are centred around this notion of unfairness - disparate treatment. Simply put, there is disparate treatment when the classifier decisions produced for an individual user change, according to variations in his or her sensitive attribute information ($\mathbf{z}$) while holding the other non-sensitive attribute information constant.[5] Intuitively, we should hope that two individuals who have the same values for all other non-sensitive features but differ in their sensitive attribute, should still receive the same classification outcome. This notion is represented succinctly as below:

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

A naïve thought experiment would possibly be to exclude the use of these sensitive attributes during decision making, i.e, $\{x_i\}_{i=1}^N$ and $\{z_i\}_{i=1}^N$ consist of disjoint feature sets. This idea motivated the implementation of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system.[6] With COMPAS, blacks and whites with the same

[4] Sampling bias often refers to the under-representation of a minority social group w.r.t certain sensitive attributes, while labelling bias refers to the persistence of biased human judgements as "labels" in the training set

[5] We usually term this **ceteris paribus**, where one assumes that all other variables except those under immediate consideration are held constant.

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

non-sensitive features get the same decision outcomes, which implied that we have parity in treatment. However, it was soon realised that we did not deal with another necessary problem: disproportionality in beneficial decisions produced across the different social groups.

2.2.2 The notion of disparate impact:

The disproportionality in the beneficial decisions produced by the classifier across the different salient social groups is exactly captured by the notion of disparate impact. If we look at the following equation:

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

For simplicity, we say that $\hat{y} = 1$ is considered a beneficial decision. We can interpret this as possibly the bank agreeing to a loan application. Ideally, we want a parity in the proportion of beneficial decisions across different values of some particular sensitive attribute. Notably, anti-discrimination laws also take into consideration this notion of disparate impact. These laws avoid disparate impact through adhering to a stipulated quantitative measure, known as the **"p% rule"** [4]. In summary, the p% rule states that the ratio between the proportion of individuals having a certain sensitive attribute assigned a beneficial decision, and the proportion of individuals not having that sensitive attribute also assigned a beneficial decision, should be no less than p:100. We illustrate **"p% rule"** in Figure 1 below - an ideal classifier should provide equal chances of attaining a beneficial decision across different groups. Note that disparate impact assumes that we do not know the correctness of decisions (i.e no ground truth).

$$\min\left(\frac{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=1)}{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=0)}, \frac{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=0)}{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=1)}\right) \geq \frac{p}{100}$$

Figure 1: A quantification of *disparate impact* stipulated by anti-discrimination laws: p% rule

Suppose instead if we know the correctness of the classifier's decisions, disproportionality in outcomes could actually be justified based on valid reasons, such as education qualifications. In that case, the notion of *disparate impact* may risk reverse-discrimination, i.e the more qualified subjects may envy the lower boundary threshold for the lower-educated to get a beneficial decision [5]. This then motivates us to consider an alternative notion, *disparate mistreatment*.

2.2.3 The notion of disparate mistreatment:

As discussed, if we happen to know the ground truth or the correctness of the historical decisions, the notion of disparate impact is actually justified and no longer serves as a meaningful notion for discrimination. We shall consider disparate mistreatment in terms of the (i) overall misclassification rate, (ii) false positive rate, (iii) false negative rate between different groups.

Disparate Mistreatment using overall misclassification rate:

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

Disparate Mistreatment using false positive rate:

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

Disparate Mistreatment using false negative rate:

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

We can see that Equation (3) is actually the sum of mutually exclusive cases: Equation (4) and Equation (5). In addition to Equations 3-5, we will briefly describe other less-known, albeit relevant error rates (such as false discovery and omission rates) in Figure 2 as below. We also provide a real-world example [7] to summarise the three aforementioned notions in **Appendix A**.

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

Figure 2: A matrix illustrating the 2 main types of error rates: (1) Fractions over the class distribution in the ground truth labels: False Positive and Negative Rates, and (2) Fractions over the class distributions in the predicted labels: False Discovery and Omission Rates

3 Synthesis of parity-based measures to deal with unfairness:

Upon reviewing relevant literature, the proposed approaches can be categorised into 3 domains: (1) Pre-processing the training data, (2) In-processing the ML models and (3) Post-processing the predictions. The baseline solutions we present in this paper belong to the "in-processing" domain, where constraints involving the sensitive attributes z are included in the optimisation objective. A more detailed overview of the three different domains can be found in **Appendix B**.

[7] We use the New York Police Department (NYPD) Stop-question and- frisk (SQF) dataset made publicly available at http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page during 2017.

Generally speaking, the fairness measures introduced in this paper often involve a tradeoff between maximising for the relevant notion of fairness and accuracy. In the next few subsections, we summarise proposed fairness measures to address each of the aforementioned notions of unfairness. We provide some guiding questions for the reader to promote better understanding:

Q1: What is the proposed proxy to the particular notion of unfairness?

Q2: What is the intuition behind the proposed proxy as in Q1?

Q3: How do we incorporate this proxy with our original optimisation objective?

3.1 Dealing with disparate treatment:

The notion of disparate treatment is naturally the easiest to deal with. As discussed in Section 2.2.1, we know that we can ensure parity in treatment, as long as we exclude the use of these sensitive attributes, $\{z_i\}_{i=1}^N$, during decision making. Specifically, the non-sensitive user features $\{x_i\}_{i=1}^N$ and the sensitive attributes $\{z_i\}_{i=1}^N$ are mutually disjoint. We adopt the ideas from the "in-processing" solution domain in fair machine learning literature, and only make use of sensitive attributes as constraints during the formulation of the optimisation objective.

3.2 Dealing with disparate impact:

3.2.1 What is the proposed proxy to model disparate impact?

Recall that in Figure 1, we ideally want our ML classifiers to follow the p% rule, where p is usually significant enough (i.e $\geq 80\%$). We note that it is challenging to directly incorporate the p% rule into our convex loss function, given that the p% rule is non-convex with respect to the classifier parameters θ. Another issue will be that the p% rule is a function having saddle points, thus rendering the procedure for solving this non-convex optimisation problem intractable. [8]

A novel *convex* proxy to efficiently design classifiers satisfying a given p% rule, was then introduced in the paper by Zafar et.all [1]. This proxy: **"decision boundary covariance"** can be represented as below, and has shown itself to empirically work with large enough datasets.

$$Cov(z, d_\theta(x)) = \mathbb{E}[(z - \bar{z})d_\theta(x)] - \mathbb{E}[(z - \bar{z})]\bar{d_\theta}(x) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \tag{6}$$

[8] As long as the user feature vectors lie on the same side of the decision boundary, the p% rule is invariant to changes in the decision boundary.

3.2.2 Intuition behind Equation (6) - "decision boundary covariance":

Intuitively, we note that a smaller covariance will lead to a higher "p% ratio". This is because a smaller covariance implies that a change in $z \in \{0, 1\}$ only varies $d_\theta(x)$ slightly; therefore, the signed distance to the decision boundary would be more likely the same for both values of z. We note that the **decision boundary covariance** is a convex function with respect to the loss function parameters, θ. Consequently, Equation (6) can be included as a fairness constraint to the relevant margin-based convex loss objective without increasing the training complexity. Furthermore, the convexity of **decision boundary covariance** allows for our solutions to be Pareto-optimal.

3.2.3 How to incorporate Equation (6) into our objectives?

With this convex proxy, we are able to achieve two relevant objectives: (1) maximising accuracy under fairness constraints, and (2) maximising fairness under accuracy constraints. [9]

Maximising accuracy under fairness constraints using Equation (6)

$$min_\theta \;\; L(\theta) \;\; such \;\; that \;\; |\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)| \leq c \tag{7}$$

Maximising for fairness under accuracy constraints using Equation (6)

$$min_\theta \;\; |\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)| \;\; such \;\; that \;\; L(\theta) \leq (1+\gamma)L(\theta^*) \tag{8}$$

Note that in Equation (7), we bound our **decision boundary covariance"** by a hyper-parameter *c*, such that we are able to have a small enough covariance thus implying a high p% value. As for Equation (8), we are interested to minimise our convex **decision boundary covariance"**, while trading off a loss value up to $(1 + \gamma)$ times of the unconstrained optimal loss determined by θ^*.

3.3 Dealing with disparate mistreatment:

3.3.1 What is the proposed proxy to model disparate mistreatment?

Recalling from Section 2.2.3, we consider disparate mistreatment with respect to the 3 error rates: overall misclassification and the corresponding false positive and negative rates. In similar vein to measures addressing disparate impact, we also consider measuring the **decision boundary covariance**. However, the difference is that the covariance takes into account the user's sensitive attributes and the signed distance between the feature vectors of misclassified users

[9] This objective was coined in the reference paper as a "business necessity clause", as we often want a minimum working level of accuracy for our ML systems.

and the decision boundary, as opposed to the previous requirement for all users. We also note that $g_\theta(y, x)$ differs for the aforementioned error rates, as shown in Equations 10 - 12.

$$Cov(z, g_\theta(y, x)) = \mathbb{E}[(z - \bar{z})g_\theta(y, x)] - \mathbb{E}[(z - \bar{z})]\bar{g}_\theta(x) \approx \frac{1}{N} \sum_{(x,y,z)\in\mathcal{D}} (z - \bar{z})g_\theta(y, x) \quad (9)$$

We note that the covariance illustrated in Equation (9) is actually non-convex in θ. Therefore, we subsequently rewrite this as a disciplined convex-concave program (DCCP), which can be efficiently solved using recent advances [6]. Given a training set $\mathcal{D}$, $\mathcal{D}_0$ and $\mathcal{D}_1$ are the subsets with a z value of 0 and 1 respectively. We then define $|\mathcal{D}_0| = N_0$, $|\mathcal{D}_1| = N_1$ and $|\mathcal{D}| = N$.

We rewrite Equation (9) as a DCCP in Equation (10), with detailed workings in **Appendix C**.

$$\frac{1}{N} \sum_{(x,y,z)\in\mathcal{D}} (z - \bar{z})g_\theta(y, x) = -\frac{N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y, x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y, x) \quad (10)$$

For *overall misclassification rate:* $g_\theta(y, x) = min(0, y * d_\theta(x))$ (11)

For *false positive rate:* $g_\theta(y, x) = min(0, \frac{1-y}{2} * y * d_\theta(x))$ (12)

For *false negative rate:* $g_\theta(y, x) = min(0, \frac{1+y}{2} * y * d_\theta(x))$ (13)

3.3.2 Intuition behind Equation 10:

As explained above, the smaller the empirical covariance listed in Equation (10) is, the less $g_\theta(y, x)$ will vary across different values for z. We now seek to understand the relationship between the different representations of $g_\theta(y, x)$ and how they address disparate mistreatment with respect to the different error rates.

We can see that in Equation 11, $g_\theta(y, x) = y * d_\theta(x)$ during only two cases: Case 1: $\{y = -1, d_\theta \geq 0\}$ & Case 2: $\{y = 1, d_\theta < 0\}$. Specifically, Case 1 (false positives) and Case 2 (false negatives) produces non-zero values for Equations 12 and 13 respectively. In essence, we want to minimise Equation (10) for the pertinent group of misclassified users.

3.3.3 How to incorporate Equation 10 into the relevant objectives?

Given the proxy as outlined in Equation 10, we want to be able to minimise our loss function:

$$min_\theta \ L(\theta) \ such \ that \ |-\frac{N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y, x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y, x)| \leq c \quad (14)$$

Remember that the covariance threshold, **c**, controls how strictly we want our classifier to adhere to parity in mistreatment, and we choose $g_\theta(y, x)$ according to a particular error-rate.

4 From parity-based to preference-based notions of fairness:

As mentioned, the introduction of the fairness constraints often lead to a tradeoff with the accuracy of the classification systems. So far, we have been working with a single parity-based classifier. A natural extension of the above-mentioned work will be to consider group-based decision boundaries. This idea draws inspiration from the two-person bargaining problem in the fair division literature [7]. Given some fixed resources, two groups try to divide these resources between themselves. In the case where resources cannot be divided equally, if one party is willing to accommodate a reasonably small amount of disparity [10], both groups will be able to obtain additional resources from the remaining resource pool. This compromise thus allows for each group to benefit more than the previous equality in the distributed resources, since they can still afford to obtain more resources from the remaining pool.

The above problem illustrates that parity-based fairness measures may be too strict, and we can instead consider more relaxed classifiers which are still able to produce Pareto-optimal solutions from each group's point of view. Intuitively, we want to construct classifiers which produce decisions, in which each distinct social group, characterised by z, will prefer as a collective whole. These classifiers are known as preference-based classifiers and are group-based, i.e: each group has their own distinct classifier.

4.1 Contextualising the two-person bargaining problem in our classification setting:

We now draw parallels from the above problem to our binary classification setting. Similar to each group wanting to maximise their allocated resoures, we usually want to maximise for some overall utility function in machine-learning settings. We formally define the **utility** function as:

$$\mathcal{U}(\theta) = \mathbb{E}_{x,y}[\mathbb{I}\{sign(\theta(x)) = y\}] \tag{15}$$

Note that $\mathbb{I}$ denotes the indicator function, and the expectation $\mathbb{E}$ is taken over the distribution $f(x, y)$.

We subsequently delve deeper into the utility function from an overall perspective down to a group perspective. We denote **group benefit** to be the fraction of beneficial outcomes received by users of a particular group (i.e sharing a certain value of the sensitive attribute z). For each value of z, we denote the **group benefit** as $\mathcal{B}_z(\theta)$:

$$\mathcal{B}_z(\theta) = \mathbb{E}_{x|z}[\mathbb{I}\{sign(\theta(x)) = 1\}] \tag{16}$$

In fact, we can actually represent parity in impact using Equation (16) to be:

$$B_z(\theta) = B_{z'}(\theta) \quad \forall z, z' \text{ and } z \neq z'$$

[10] The amount where one party is willing to compromise until, is termed the **disagreement point**.

4.2 Why should we care about group-based decision boundaries?

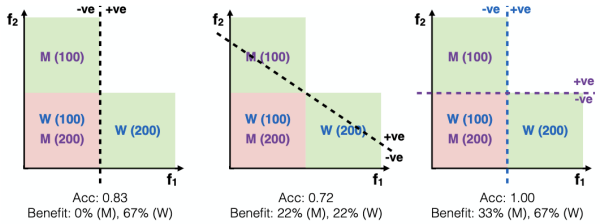


Figure 3: Consider a classification system involving gender as the sensitive attribute.

Figure 3 illustrates a classification setting based on two features, where feature **f1** is highly predictive for women whereas feature **f2** is highly predictive for men. Green (red) quadrants denote the positive (negative) class. The left panel shows a classifier satisfying treatment parity, but produces disparate impact as seen in the difference: $\mathcal{B}_{male}(\theta) = 0$ versus $\mathcal{B}_{female}(\theta) = 0.67$.

We then utilise Equation (7) to train a classifier which additionally satisfies parity in impact, as shown in the central panel. However, the achievement of impact parity was made possible through misclassifying women from positive class into negative class; therefore resulting in a marked decrease in our utility $\mathcal{U}(\theta)$ of interest (83% to 72% accuracy).

Can we do better? Yes, through the use of group-based classifiers! This is illustrated on the right panel of Figure 3, where the vertical blue boundary (horizontal purple boundary) is applicable to women (men). Ideally, we want our group-based classifiers to fulfil two distinct notions (i) preferred impact, (ii) preferred treatment, as fulfilled by the right panel classifier.

4.2.1 Preferred impact:

Referring to Figure 3, we observe that both men and women obtain higher **group benefit** values of 33% and 67% respectively when using group-based classifiers in the right panel. This is opposed to the lower **group benefit** value of 22% using the central panel classifier, which insists on parity in impact. We say that a classifier is a a preferred impact classifier, if it achieves higher **group benefit** for each sensitive attribute $z \in \mathbb{R}^+$ as opposed to the impact parity classifier. Formally:

$$B_z(\theta_z) \geq B_z(\theta'_z) \quad \forall z \in \mathbb{R}^+$$

4.2.2 Preferred treatment:

We also observe that men will collectively prefer a **group benefit** value of 33% using the horizontal group-based classifier on the right, as opposed to 0% using the single classifier on

the left. We say that a classifier is a preferred treatment classifier, if each group characterised by z, receives a higher **group benefit** value using their own group-based classifier θ_z as opposed to other group-based classifiers denoted generally by $\theta_{z'}$. Formally:

$$B_z(\theta_z) \geq B_z(\theta_{z'}) \quad \forall z, z' \in \mathbb{R}^+ \text{ and } z \neq z'$$

4.3 How do we train preferred classifiers?

In this section, our goal is to train classifiers which satisfy the aforementioned notions of "preferred treatment" and "preferred impact" group-conditional classifiers. Specifically, we want to find the optimal parameters: $\boldsymbol{\theta} = \{\theta_z\}_{z \in \mathbb{R}+}$, such that we can maximise a given utility function $\mathcal{U}(\boldsymbol{\theta})$ over a training dataset $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^N$.

4.3.1 Preferred impact classifiers:

Recall that we provided the theoretical utility function $\mathcal{U}(\boldsymbol{\theta})$ in Equation (15). In practice, we can train classifiers, which will minimise for a Monte-Carlo estimate of the negative-valued version of the utility function [11], given the preferred impact property listed in Section 4.2.1.

$$\begin{aligned} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} \quad & -\tfrac{1}{N} \textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T \boldsymbol{x}) = y\} \\ \text{subject to} \quad & \textstyle\sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T \boldsymbol{x}) = 1\} \geq \sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z'^T \boldsymbol{x}) = 1\} \quad \text{for all } z \in \mathcal{Z} \end{aligned}$$

Intuitively, we want to maximise for our utility function, given the preferred impact requirement on our classifier. This is represented by the constraint that the **group benefit** proportion produced by the optimal group-conditional classifier ($\boldsymbol{\theta_z}$), should be greater than or equal to that of the parity impact classifier ($\boldsymbol{\theta'_z}$), across all values of z.

4.3.2 Preferred treatment classifiers:

Similarly, we also minimise for the same negative-valued version of our utility function. However, we now constrain the objective with the preferred treatment property listed in Section 4.2.2.

$$\begin{aligned} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} \quad & -\tfrac{1}{N} \textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T \boldsymbol{x}) = y\} \\ \text{subject to} \quad & \textstyle\sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T \boldsymbol{x}) = 1\} \geq \sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_{z'}^T \boldsymbol{x}) = 1\} \quad \text{for all } z, z' \in \mathcal{Z} \end{aligned}$$

Similarly, we maximise our utility function, given the preferred treatment requirement on our classifier. For every group-conditional classifier, denoted by θ_z, we want the relevant group z to receive the most preferred amount of **group benefit** produced with θ_z as compared to a competing $\theta_{z'}$ characterising a classifier belonging to any other group, z'. We provide the exact formulations for Sections 4.3.1 and 4.3.2 in the logistic regression setting in **Appendix D**.

[11] We choose to minimise the negative value, since DCCP solvers are by default solving for minimisation problems.

5 So, who's the fairest of them all?

As the title of the paper implies, given this multitude of notions for fairness, how can we reconcile our knowledge and decide which measure of fairness is the most applicable? We provide an actionable flowchart based on required contexts in the following figure, Figure 4.

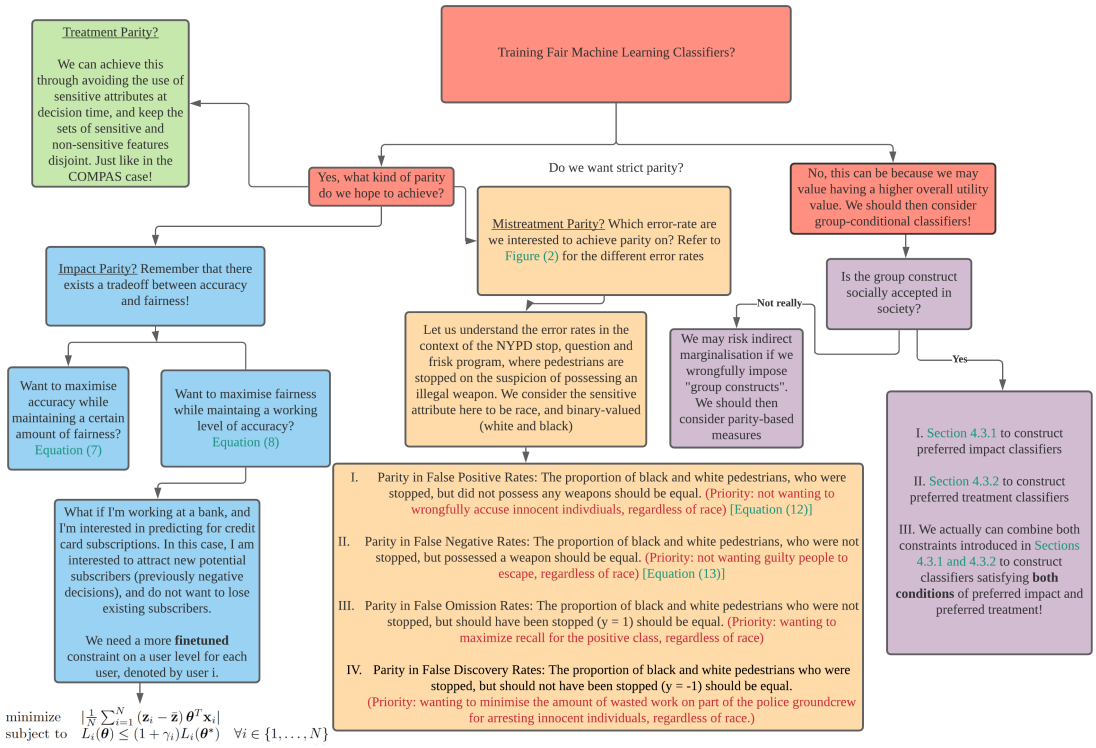


Figure 4: A comprehensive flowchart summarising aforementioned fairness measures.

6 Experiments:

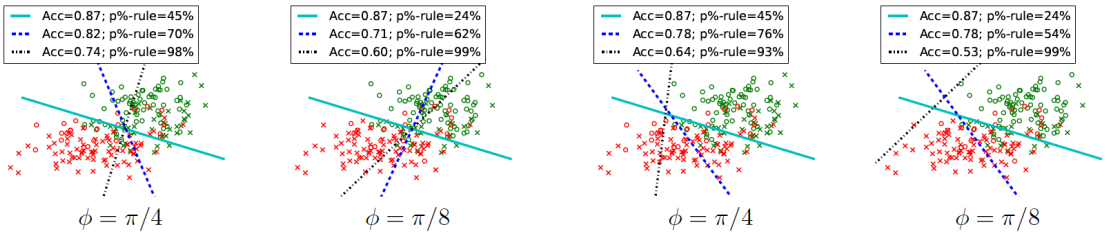
In this section, we train logistic regression classifiers with aforementioned fairness measures, and we note that this can be easily extended to other margin-based classifiers too.

6.1 Classifiers satisfying impact parity:

Specifically, we generated 4,000 binary class labels randomly using a uniform distribution, and assigned a 2-dimensional user feature vector per label by drawing samples from two distinct Gaussian distributions: $p(\boldsymbol{x}|y = 1) = N([2;2],[5,1;1,5])$ and $p(\boldsymbol{x}|y = -1) =$

$N([-2;-2],[10,1;1,3])$.

We then model each user's sensitive attribute using 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 $\boldsymbol{x}$. Note that as ϕ approaches 0, the higher the correlation is between the sensitive attributes and the user features. We then proceed to train two different classifiers which satisfy **Equations (7) and Equation (8)**, and visualise the classification space in the below figure: Figure 5.

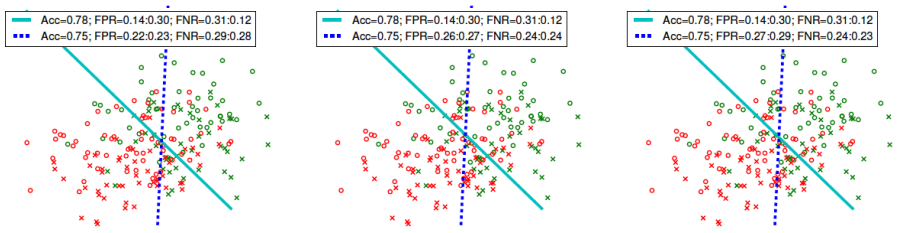


(a) Maximizing accuracy under fairness constraints (b) Maximizing fairness under accuracy constraints

Figure 5: In both panels, we immediately notice the tradeoff between accuracy and the p% rule. We also note that as ϕ gets closer to 0, our user features has higher discrimination ability for z. Thus, we have to incur a higher misclassificaiton rate to achieve the same fairness level.

6.2 Classifiers satisfying mistreatment parity:

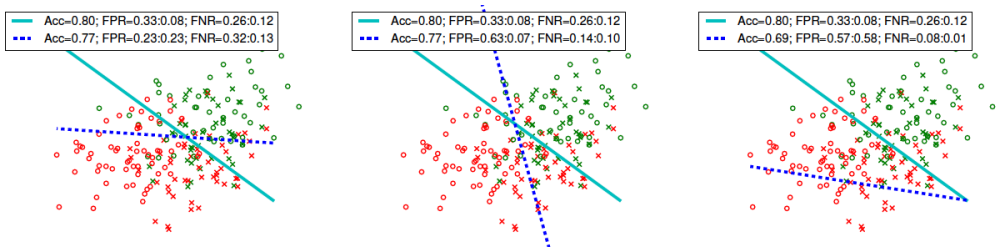
Case 1: We have that the disparity in false-positive rates (D_{FPR}) and false-negative rates (D_{FNR}) have opposite signs:



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

Figure 6: The removal of disparate mistreatment involves the rotation of the decision boundary. When D_{FPR} and D_{FNR} are of opposite signs, the previously misclassified positively-labeled circles ($z=1$) are now predicted as negative. This helps to decrease the false positive rate, but this in turn also increases the false negative rate for the class $z=1$.

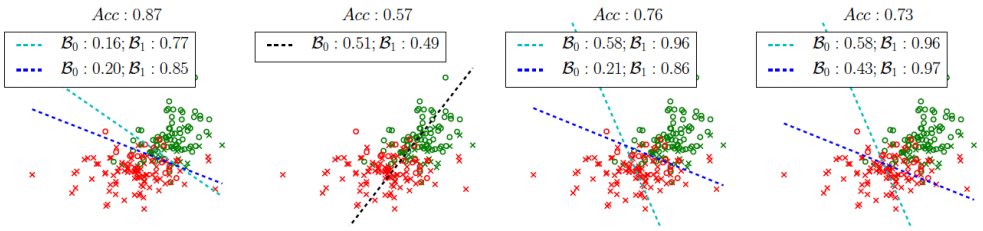
Case 2: We have that the disparity in false-positive rates (D_{FPR}) and false-negative rates (D_{FNR}) have the same signs:



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

Figure 7: When D_{FPR} and D_{FNR} have the same signs, the complementary effect between false-negative and false-positive rates as illustrated in Figure 6 does not hold water. In the left panel, when we try to control for disparate mistreatment for false positive rates (to get 0.23 for each group), we exacerbate the disparity in the false negative rate from (0.26 v.s 0.12) to (0.32 vs 0.13). If we do intend to achieve parity in both error-rates, it will come at a significant expense of our utility, as shown in the right panel classifier (a dip from 0.8 to 0.69 acccuracy).

6.3 Classifiers satisfying preferrence-based fairness notions:



(a) Uncons (b) Parity (c) Preferred impact (d) Preferred both

Figure 8: Crosses denote points belonging to group 0, while circles denote points belonging to group 1. Green (red) points are positive (negative) labels. Each panel shows the accuracy of the classifiers along with group benefits (B_0 for group 0 and B_1 for group 1 respectively). For group-conditional classifiers, cyan (blue) line denotes the decision boundary for the classifier of group 0 (group 1).

Contextualising the four different classifiers:

1. **Uncons:** Unconstrained group-based classifiers which are trained separately for each sensitive attribute value group. Note that this violates treatment parity, and possibly

impact parity too.

2. **Parity:** A single classifier which serves to resolve issues of disparate treatment and disparate impact in **Uncons**. This follows the ideas in Equations (7) and (8).

3. **Preferred impact:** Group-based classifiers which are trained with constraints that the provided group benefits should outweigh those provided by the **Parity** classifier.

4. **Preferred both:** Group-based classifiers which are trained to satisfy both preferred impact and preferred treatment conditions.

We have provided the source code for the various experiments done, which can be easily reproduced at `https://github.com/kenghweeng/fairest-of-them-all`.

7 Conclusion:

We would like to point out that the aforementioned fairness measures were also done on real-world datasets.[12] We would like to end this report with 2 concluding remarks.

Firstly, we note that for the error-rates: false omission rates and false discovery rates, we still do not have tractable convex or convex-concave proxies for them and we look forward to adopting them in future literature. Also, we note the preference-based notions are discussed in the context of a collective group. It would be worth revisiting the aforementioned notions in the context of individualised preferences instead, as we may interested for a more fine-tuned preference notion catering to specific individuals.

We would like to conclude that we, as researchers, should maintain informed neutrality between the different notions of discrimination (and should be as fair as possible, pun intended).

[12] Please refer to [1], [2], [3] for experiments on well-known datasets such as the Adult, Bank and COMPAS datasets.

8 Appendix A: The tale of the 3 notions of discrimination

We seek to understand the notions of disparate treatment (Disp. Treat.), disparate impact (Disp. Imp.) and disparate mistreatent (Disp. Mist.) with the following scenario in Figure 2.

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

Figure 9: Decisions of three trained classifiers (C1, C2 and C3) to stop a pedestrian on the suspicion of possessing an illegal weapon (1 for stop, 0 for no stop). 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.

The case of disparate treatment: We deem C_2 and C_3 to exhibit disparate treatment since the decisions of C_2 for Male 1 and Female 2 (Male 2 and Female 2 for C_3) are different even though they have the same values for the non-sensitive attributes: suspicious bulge in clothing and proximity to a crime scene.

The case of disparate impact: We deem C_1 unfair since the fraction of males ($\frac{3}{3}$) and females ($\frac{2}{3}$) that were stopped are different. This is opposed to C_2 and C_3, where $\frac{2}{3}$ in both the male and female groups were decided by the classifiers to stop.

The case of disparate mistreatment: We deem C_1 and C_2 unfair due to the differential rates in erroneous decisions. We note that for disparate mistreatment, we will have to make use of the ground truth information. C_1 disadvantages males, as the false negative rate for males is 0 given that the classifier always decides to stop males. C_2 has different false positive rates and false negative rates for males and females.

9 Appendix B: The tale of the 3 domains of fairness solutions

The first domain (pre-processing of training data) involves representational learning, in which the training data is mapped to a transformed space where the dependencies between sensitive attributes and class labels disappear. This pre-processing strategy is in fact a beautiful marriage with research on privacy of machine-learning [8], where we are interested in ideas of "data-anonymization" and we do not want the sensitive attributes to be easily identifiable with our feature space and should be as uncorrelated as possible. However, there are a few limitations we should highlight. Firstly, there is no common ground as to what the reduced dimensions of the representation space should be. This resonates similarly to the idea of hyper-parameter tuning in training deep neural nets. Secondly, we note that this approach generally treats the machine-learning models as a black-box, and this will not allow us to get a theoretical guarantee of the tradeoff bounds between fairness and accuracy.

The discussed approaches in our paper falls into this second domain, which involve the in-processing of usually convex or convex-concave fairness constraints coupled with our desired loss function in the ML setting. There are some relevant works by Kamishima et all. [9], where they introduce a regularization term to penalize discrimination in the formulation of the logistic regression classifier. Recent literature tend to mostly adopt this approach. Note that this domain is possible only if the constraints required to ensure fairness are also convex or convex-concave in nature. It is necessary to point out that most of our discussed notions fortunately are able to adopt convex or convex-concave constriants; however, there are still some error-rates such as false omission and discovery rates which still require further research on tractable proxies.

The final domain involves post-processing the probability estimates of an unfair classifier. This strategy [10] allows the classifier to learn varying decision thresholds for different sensitive attribute value groups, and applying these group-specific thresholds at decision making time. However, it is crucial to caution that the method requires knowledge of the sensitive attribute at decision time, and this then renders the issue of disparate treatment where we require the use of sensitive attributes. Also, if the sensitive attribute information is unavailable due to privacy or disparate treatment laws [11], this solution domain will then be rendered useless.

10 Appendix C: Formulating the DCCP equation

Solving the problem efficiently. While the constraints proposed can be an effective proxy for fairness, they are still non-convex, making it challenging to efficiently solve the optimization problem in (13). Next, we will convert these constraints into a Disciplined Convex-Concave Program (DCCP), which can be solved efficiently by leveraging recent advances in convex-concave programming [6].

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

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

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

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

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

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

11 Appendix D: Preferred impact and treatment classifiers in the logistic regression setting

We now formally present how to construct logistic regression classifiers, with the corresponding log likelihood function and sigmoid function, that satisfies preferred impact and preferred treatment properties.

$$\begin{aligned} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} \quad & -\tfrac{1}{N}\textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}_z) + \sum_{z\in\mathcal{Z}} \lambda_z ||\boldsymbol{\theta}_z||^2 \\ \text{subject to} \quad & \textstyle\sum_{\boldsymbol{x}\in\mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x}\in\mathcal{D}_z} \max(0, \boldsymbol{\theta}_z'^T \boldsymbol{x}) \quad \text{for all } z \in \mathcal{Z}. \end{aligned}$$

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

References

[1] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[2] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[3] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

[4] Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[5] Ann C McGinley. Ricci v. destefano: A masculinities theory analysis. *Harv. JL & Gender*, 33:581, 2010.

[6] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1009–1014. IEEE, 2016.

[7] John F Nash Jr. The bargaining problem. *Econometrica: Journal of the econometric society*, pages 155–162, 1950.

[8] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[9] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[10] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[11] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.