Short-term fairness constraints can lead to long-term harm: results from simulations

Adrien Morisot

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh

2019

Abstract

Machine learning is playing an increasingly large role in decision making, and therefore it is imperative that the decisions that machine learning systems make are fair. Multiple metrics that measure fairness have been proposed. These focus mainly on atemporal, short term aspects of fairness, which we dub synchronic fairness.

We run temporal, dynamic simulations in which these synchronic fairness metrics are optimised for, and contrast the results obtained to simulations in which these metrics are not optimised for. Comparing the results, we find that trying to optimise for synchronic fairness can, counterintuitively, lead to results that are less fair than what would be obtained without optimising for synchronic fairness.

We also find that optimising for synchronic fairness can cause the most harm when the initial distributions are particularly unfair, and explain why.

Acknowledgements

I would like to thank my supervisors John Pate and Gianluca Corrado for introducing me to, and expertly guiding me through, the field of fairness in machine learning, as well as for their feedback and support. Thanks also to my second supervisor Iain Murray, for his constructive comments on my project proposal, and for facilitating coordination between the University of Edinburgh and Amazon.

I would also like to thank my friends and coursemates –Tiffany, Viraat, Patrick, Alice, Azad– for being excellent sounding boards.

Special thanks go out to everyone who contributes to open source software, such as numpy and scipy, without which my dissertation would probably not have been possible.

Finally, to my parents and sister, thank you for your unwavering love and support.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Adrien Morisot*)

Table of Contents

Chapter 1

Introduction

Machine learning algorithms are increasingly making decisions that have the potential to significantly affect people's lives, such as whether someone ought to be hired or promoted, whether someone is targeted and interrogated by the police, and whether someone receives or is denied a loan. For example, some companies now use machine learning in automated resume scanning in order to determine which individuals get to the next round of interviews (ideal.com, 2018). Credit score assignment companies are starting to use machine learning to inform their scoring decisions (FICO, 2018). Mass video surveillance companies use machine learning to try to match people captured by CCTV cameras to databases of criminals (calipsa.io, 2018).

This shift towards using machine learning to automate certain tasks has many advantages: it allows companies to operate at a larger scale, make their products cheaper, and increase their products' performance.

However, the shift can also have unwelcome effects. More specifically, it can introduce, reinforce, or aggravate different forms of discrimination –such as racism, sexism, and homophobia– in the workings of software systems and products, rendering them unfair (Barocas et al., 2018).

This thesis studies these effects, and finds that designing machine learning systems that maximise fairness in the short run can lead to unfair outcomes in the long run, and vice versa.

1.1 How do machine learning systems learn to be unfair?

This is a broad question, with a plethora of different answers depending on the type of machine learning system, the context, and the definition of fairness one chooses to employ. In order to answer it, we first need to narrow our scope, and provide both a preliminary definition of fairness, as well as the context in which we wish to operate. We will expand upon these points in Chapter 2.

1.1.1 Preliminary definitions

For now, we set the context in which we wish to operate to be supervised learning, specifically *classification*, a task which involves assigning labels to subjects based on their characteristics. For a bank, the task might be making lending decisions, in which case the subjects would be loan applicants, their characteristics would be attributes such as credit score or employment status, and their labels would be “accept loan” or “decline loan”.

We also posit that a machine learning system is *unfair* if there is a significant difference in the system’s outputs for different natural groups. By *natural group*, we mean groups of people defined by properties that their members were born with, and cannot change, such as race, gender, sexuality, etc. For example, according to our definition, it is unfair that the credit scores of women are systematically lower than those of men.

1.1.2 The machine learning pipeline

In order to understand the ways in which unfairness can creep into decisions made by machine learning systems, it is helpful to understand the machine learning pipeline, the way in which machine learning systems are built, illustrated in Figure 1.1.

In order to train a classifier, one first needs to gather labelled data. One does this by sampling objects from the world, and then using humans to assign a label to each object in the sampled set. One then uses this labelled data to learn a model. Once the model is learned, it assigns new labels to objects that had not previously been given a label. The people using the model (e.g. banks, police departments, etc.) then make decisions

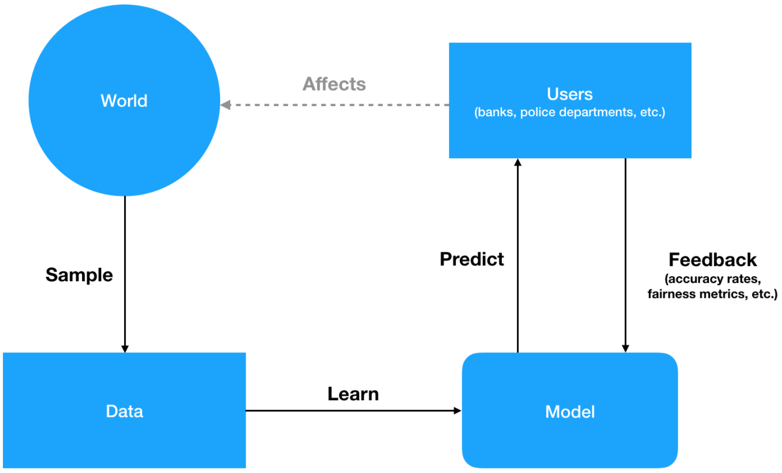


Figure 1.1: Machine learning pipeline, reproduced from Barocas et al. (2018).

based on those new labels. Those decisions have an effect on the world, and also on people using the model (the bank's profit goes up or down, the police department's neighbourhood crime rate goes up or down, etc.). The users then modify the model if they think they can improve it.

Assuming that the sampling method used is fixed, we identify two ways in which this system can produce unfair results.

- **Non-representative sampling**
  The model can be unfair because the dataset on which the model is trained systematically discriminates against a certain group. So even if the marginalised population's probability distribution over a property is equal to that of a non-marginalised population, the samples obtained from those groups disadvantage the marginalised population, and are thus unfair[1]. For example, suppose a police department tries to train a model that predicts whether or not someone is likely to be in possession of drugs. African-Americans currently represent 13% of the U.S. population, and 14% of its drug users (Mauer, 2007; Lum and Isaac, 2016). However, if the model is trained on a dataset in which half the people who were convicted of committing a drug crime are African-American, then the model will likely see a strong correlation between drug possession and race. As a consequence, the model will likely predict that a disproportionate number of African-Americans are likely to possess drugs, despite this not being grounded

[1] Throughout this thesis, we use "marginalised" population to refer to a population that is worse off than some other "non-marginalised" population. Our use of "marginalised" does not refer to the probabilistic/statistical notion of marginalisation.

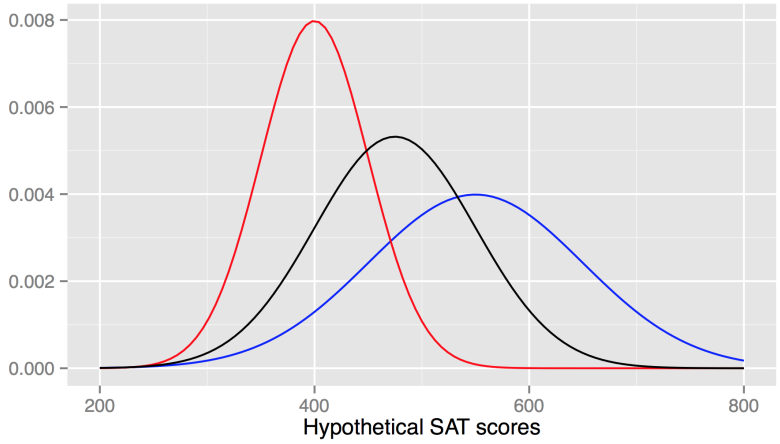


Figure 1.2: Hypothetical plot of probability distributions over SAT test scores. If the distributions were obtained through *non-representative sampling*, then the central black curve is the true distribution of both male and female scores. However, the data does not reflect this, and the leftmost red curve (representing the marginalised population's distribution) incorrectly has a much lower mean score than the rightmost blue curve (representing the non-marginalised population's distribution), even though they ought to overlap. If the distributions *truly are unequal*, then the red and blue curves correctly don't overlap. Figure reprinted from Feldman et al. (2015).

in reality.

- **The distributions are unequal**
  The model can be unfair because it reflects the reality of an unfair world, in which history and tradition play a major role in shaping the way in which humans behave. If a marginalised population's probability distribution over a property is truly different from that of a non-marginalised group, then the samples drawn from those distributions can be both statistically representative, in that they represent true samples drawn from the distribution, but also still unfair, according to our definition. For example, for a long time, and still today to some extent, tradition dictated that women ought to be homemakers. The effects of this tradition (and the sexism that accompanied it) are that women earn on average less than men and so their credit score distribution is lower than the distribution of men (Mottola, 2013), which is unfair according to our definition.

Figure 1.2 illustrates this further. These two different ways in which the machine learning system can produce unfair results, non-representative sampling and unequal initial distributions (or a combination of the two), lie at the heart of the issues we explore in this thesis.

1.2 Motivation for the thesis

Given this distinction between non-representative sampling and unequal distributions, we identify two ways in which we can attempt fair classification.

To address the non-representative sampling problem, we can try to constrain the classification algorithm in such a way that at prediction time, the classifier's predictions are as if the two initial distributions are actually very similar. It focuses on the Learn, Predict & Feedback arrows in Figure 1.1. This solution has been, and still is, a big focus of the fair classification literature, which has developed a wide variety of metrics to measure how well the modified classifier satisfies a certain set of constraints, such as equalised odds or equality of opportunity (see Chapter 2 for full definitions). We call this *synchronic fairness*, as it is a type of fairness concerned with the immediate effects of the classifier, outside of time.

We contrast this type of fairness to *diachronic fairness*. Diachronic fairness is a type of fairness concerned with the effects of a classifier over time, both on the populations that it is affecting, and on its own behaviour. We can use diachronic fairness to address the unequal distributions problem, by designing a classifier that takes into account the fact that its predictions have an effect over time to close the gap between the true distributions of the marginalised and non-marginalised populations over time. This solution focuses on the Affects & Learn arrows in Figure 1.1.

There is a certain amount of tension between these two types of fairness, since one is atemporal, whereas the other is defined by time. This tension gives rise to the question: how sharp a distinction should we draw between these two types of fairness, and how compatible are they?

1.3 What we show

- We show that for some situations in which the classifier has a large effect on the populations that it is classifying, using synchronic fairness –i.e. optimising for a short term fairness constraint– can lead to unfair results, and widen the gap between the marginalised and non-marginalised populations over one cycle of the machine learning pipeline, which we call a time-step.

- We show that this effect is exacerbated over multiple time-steps.
- We show that fairness constraints do not have a larger effect if the initial situation is unfair and the mean score of the marginalised population is much lower than that of the non-marginalised population.

1.4 Thesis outline

The remainder of the thesis is organised as follows. Chapter 2 formalises the task of classification, and defines the fairness metrics relevant to the thesis. Chapter 2 also provides an overview of the literature surrounding fair classification, focusing on the distinction between work on synchronic fairness and diachronic fairness, and on impossibility results, which show that machine learning systems cannot satisfy certain fairness criteria at the same time. Chapter 3 describes the way in which we use simulations to experiment with different techniques of fair classification, and the effects of those techniques on simulated populations. Chapter 4 presents the results of our experiments, and discusses their significance, dwelling particularly on our finding that maximising fairness in the short run can lead to unfair outcomes in the long run and vice versa. Finally, Chapter 5 concludes by summarising the results obtained, highlighting their limitations, and giving directions for future work.

Chapter 2

Background

In the previous Chapter, we provided a summary definition of fairness, outlined the machine learning pipeline, and drew a distinction between synchronic and diachronic fairness. This chapter formalises these notions, and uses this formalisation to review past work. We first introduce a distinction that Crawford (2017) draws between different types of unfairness leading to different types of harms, namely allocative harm and representational harm. We then review two case studies that suggest that allocative harm has the capacity to perpetuate itself over time. With this context in place, we study the ways in which the machine learning literature addresses the topic of fairness, focusing on the distinction between work primarily aimed at achieving fairness in the synchronic setting (in a single instant), and work aimed at achieving fairness in the diachronic setting (over time).

2.1 Formalisation of fair classification

The purpose of fair classification is, given information about someone, to assign them a label.

The formalisation for fair classification that we will be using throughout this thesis is as follows.

- x_i is a d-dimensional feature vector of an individual, composed of α sensitive features $s_{i1}, \ldots, s_{i\alpha}$, and β non-sensitive features $u_{i1}, \ldots, u_{i\beta}$, such that $\alpha + \beta = d$. For the sake of simplicity, we assume that there is only a single sensitive feature (i.e. $\alpha = 1$ and $s_1, \ldots, s_\alpha = s$), and that it is binary, such that $s \in \{0,1\}$.

- y_i is that individual's corresponding label, which we assume to be binary, i.e. $y_i \in \{0,1\}$.
- $g_0(\cdot)$ and $g_1(\cdot)$ are the probability distributions from which marginalised and non-marginalised individuals' characteristics and labels are respectively sampled from. μ_0 and μ_1 are their respective means. Note that $g_0(\cdot) = P(u_i \mid s_i = 0)$ and $g_1(\cdot) = P(u_i \mid s_i = 1)$.
- X and y are an $N \times d$ matrix of feature vectors and an $N \times 1$ column of labels, respectively.
- $f(\cdot)$ is a learned function that maps a individual's feature vector x_i to some score c_i (which can be a scalar or a vector).
- $d(\cdot)$ is a decision function that maps an individual's score to a label y_i^*. We assume that label to be binary, i.e. $y_i^* \in \{0,1\}$.
- We let y^* be a vector of all predicted labels y_i^*.

We dub the label $y = 0$ as the negative class (e.g. was denied a loan, was not allowed to leave prison early), and the label $y = 1$ as the positive class (e.g. was granted a loan, was allowed to leave prison early). We also dub the feature $s = 0$ as the feature defining the marginalised group (e.g. women in the credit score scenario), and the feature $s = 1$ as the feature defining the non-marginalised group (e.g. men in the credit score scenario).

Classification involves learning the mappings $f(\cdot)$ and $d(\cdot)$ that maximise test-set accuracy a. Here, a is the ratio of correct predictions (i.e. $y_i^* = y_i$) to total number of predictions, for individuals i in an unseen test set. In order to assess test-set accuracy, X and y are split into X_{train}, X_{test}, y_{train}, and y_{test}.

Fair classification preserves this initial classification setup, while adding an additional constraint: the classification must try to jointly maximise the accuracy a, as well as a measure of fairness e.

Finally, we formalise the concepts that were introduced in Chapter 1:

- **Non-representative sampling**

  $g_0(\cdot)$ and $g_1(\cdot)$ are equal to the same distribution, $g(\cdot)$, i.e. $g_0(\cdot) = g_1(\cdot) = g(\cdot)$. However, the procedure that samples from this distribution is non-representative, i.e. when $x_1, \ldots, x_n$ are sampled from $g(\cdot)$, the resulting probabilities are such that $P(x \mid s = 0) \neq P(x \mid s = 1)$, even though they are supposed to be equal. Instead, individuals x_i are being sampled from transformations of the initial $g(\cdot)$, call these transformations $\tau_0(\cdot)$ and $\tau_1(\cdot)$, where that transformation depends on the sensitive attribute. So marginalised individuals x_i will be sampled such that $x_i \sim \tau_0(g(\cdot))$, and non-marginalised individuals x_i will be sampled such that $x_i \sim \tau_1(g(\cdot))$.

- **Unequal distributions**
  $g_0(\cdot)$ and $g_1(\cdot)$ are unequal, with $\mu_0 < \mu_1$.

- **Diachronic fairness**
  Suppose $g_0(\cdot)$ and $g_1(\cdot)$ are equal, but $x_1, \ldots, x_n$ sampled from $g_0(\cdot)$ and $g_1(\cdot)$ are such that $P(x \mid s = 0) \neq P(x \mid s = 1)$, with $\mu_0 < \mu_1$. Diachronic fairness uses various pre-processing, intra-processing, and post-processing techniques that modify the data X, the predicted labels y^*, or the score function $f(\cdot)$, to ensure that the labels y^* were obtained as if $x_1, \ldots, x_n$ were obtained from equal initial distributions $g_0(\cdot)$ and $g_1(\cdot)$, while being reasonably accurate. Here, the fairness measure e is some measure of the *initial* distance between $P(x \mid s = 0)$ and $P(x \mid s = 1)$, and the goal is to minimise that distance.

- **Synchronic fairness**
  Suppose $g_0(\cdot)$ and $g_1(\cdot)$ are not equal, but that the labels y_i^* assigned to each individual i have an effect on the characteristics x_i of that individual. Synchronic fairness uses the same techniques as diachronic fairness, but with the goal of changing X, y^*, or $f(\cdot)$, in such a way that the $g_0(\cdot)$ and $g_1(\cdot)$ become less different in the future, while being reasonably accurate. Here, the fairness measure e is a measure of the *future* distance between $g_0(\cdot)$ and $g_1(\cdot)$, and the goal is to minimise that distance.

The assumptions underlying diachronic and synchronic fairness are different, and relate to two different types of harm caused by unfairness in machine learning systems discussed by Crawford (2017), which we discuss in the next section.

2.2 Two types of harm caused by unfair ML systems

According to Crawford (2017), unfairness in machine learning systems can cause two kinds of harm: representational harm and allocative harm. These are defined below.

2.2.1 Representational harm

Representational harm is harm that occurs when certain groups are stereotyped (by race, gender, religion, etc.) and grouped into some categories while being excluded from others.

For example, Bolukbasi et al. (2016) found that word embeddings trained on large free text datasets (in this case Google News) can pick up on gender imbalances within language. For example, they found that within the trained embedding, men were most closely associated with words such as “maestro”, “skipper”, and “protege”, whereas women were most closely associated with words such as “homemaker”, “nurse”, and “receptionist”. These imbalances in word embeddings can play out in the real world. In order to see this, Barocas et al. (2018) recommend using online neural network based translation tools –such as Google Translate– to translate from languages where pronouns do not have a gender (i.e., in languages where there is no distinction between the pronouns “he” and “she”) to languages where pronouns do have a gender, and back. For example, translating “She is a doctor. He is a nurse” into Turkish (which does not have gendered pronouns) returns “O bir doktor. O bir hemşire”. Translating “O bir doktor. O bir hemşire” back into English returns “He is a doctor. She is a nurse”, flipping the meaning back into the gendered stereotype that doctors are male and nurses are female. The pronoun flip also occurs in Hungarian, Persian and Malay. This is an example of representational harm, since it perpetuates gendered stereotypes about the jobs that men and women have.

2.2.2 Allocative harm

Allocative harm is harm that has a direct negative effect on a particular group.

For example, in the United States, some states use a machine learning tool named COMPAS to assign a risk score to prison inmates (Lum and Isaac, 2016; Perry, 2013).

A judge uses these scores to determine whether to grant parole to an inmate. An analysis performed by ProPublica, a nonprofit newsroom, suggested that the tool discriminates against African-American inmates (Angwin et al., 2016). Indeed, they found that COMPAS's false positive rate (i.e. the rate at which an inmate is classified as high risk, yet is released, and does not go on to recommit a crime) is nearly double for African-American inmates relative to Caucasian inmates: it is 23.5% for Caucasians and 44.9% for African-Americans. Conversely, COMPAS's false negative rate (i.e. the rate at which an inmate is classified as low risk, released, and does goes on to recommit a crime), is nearly double for Caucasians relative to African-Americans. It is 28% for African-Americans and 47.7% for Caucasians. In other words, COMPAS has a tendency to disproportionately label African-Americans as high-risk even though they won't go on to commit a crime, and to label Caucasians as low risk even though they will go on to commit a crime. This is an example of allocative harm, since it leads to African-Americans being kept in prison for longer than Caucasians.

We assume that there is no biological or natural explanation for which women would be more drawn to lower paying jobs than men, or that they would be less good at performing higher-paying jobs than men. Similarly, we assume that there is no biological or natural explanation for which African-Americans would be more likely to recommit a crime if released early from prison than Caucasians. If there is a significant difference between these two distributions, we can attribute this difference to different forms of discrimination, as well as various forms of societal pressures, exercised over time.

By this, we mean that if a group of people is initially disadvantaged, there are mechanisms related to allocative harm by which they will stay disadvantaged over time. In the next section, we discuss two case studies from the social sciences which suggest that this hypothesis is empirically grounded.

2.3 Two case studies

The following two examples qualitatively illustrate the ways in which a marginalised group's average crime rate or credit score might remain the same or worsen due to feedback loops.

2.3.1 Incarceration begets incarceration

One might reasonably assume that sending someone to prison would reduce their probability of committing crimes in the future. Indeed, Bhuller et al. (2016) find that Norwegian prisons are effective at reforming their inmates and significantly reducing their probability of future crime.

However, the studies discussed below suggest that in the United States, prisons are failing at reforming their inmates, and that spending time in American prisons raises the probability of future criminal behaviour.

Nagin et al. (2009) examine 47 different studies, and find that on average it is reasonable to assume that prisons do not make people less likely to commit crimes once they are released, and instead have the opposite effect. Similarly, Cullen et al. (2011) find that "there is little evidence that prisons reduce recidivism and at least some evidence to suggest that they have a criminogenic effect". Bales and Piquero (2012) go further than this, and assert that American prisons have a substantive criminogenic effect, relative to other forms of punishment that do not involve prison. Finally, Jonson (2010) conducts a meta-analysis with 85 studies to quantify these effects, and finds that going to prison increases recidivism rates by 14% relative to forms of punishment that do not involve prison, and that harsh prison conditions (fewer visits allowed from family members, less parole) increase recidivism rates by a further 15%.

These effects can be explained by a variety of factors, listed by Roodman (2017):

- It is harder to find a job after going to prison, as employers are averse to employing convicts, making turning back to crime more likely.
- When someone is imprisoned, their social ties with their acquaintances outside of prison are broken, and are replaced with social ties with criminals, which may encourage criminal behaviour in the future.
- If someone commits a gang- or drug-related crime, then going to prison is probably not going to rehabilitate them, because both gangs and drugs are plentiful in prison.

These studies all suggest that crime can beget more crime, and that one should not ignore possible negative feedback loops in the criminal justice system, especially since there is a strong racial component to sentencing in the United States: African-Americans serve about as much time in prison for a drug offence as Caucasians do for a violent crime (around 5 years) (Mauer and King, 2007); although African-Americans represent around 14% of the American drug users, they represent more than half of inmates in state prisons for drug offences (Mauer, 2007); African-American youth are 40% more likely to be arrested and detained than Caucasian youth (Sneider and Sickmund, 2006); despite only representing 13% of the American population, African-Americans represent around 38% of the American prison population (Wagner and Sawyer, 2018).

2.3.2 Poverty begets poverty

Furthermore, Bowles (2006) finds that similar feedback loops can occur with regards to economic status. Bowles claims that classical economic theories affirm that over generations peoples' fortunes end up regressing to the mean of the distribution of wealth. He argues that these theories fail to account for what he calls "poverty traps", pressures that make people at the low end of the distribution of wealth unlikely to escape it. Indeed, he finds that "the son of a person born to parents in the poorest decile of income earners is 24 times more likely to achieve an income in the lowest decile than in the highest decile when he grows up. The son of parents in the top decile of income earners is 10 times more likely to remain at the top than to fall to the bottom decile".

A relevant model of poverty traps put forth and discussed by Bowles (2006) is that there exist "critical thresholds" of wealth, and that individuals below these thresholds have a significantly lower probability of escaping poverty and acquiring wealth, due to not being able to afford things such as proper housing, enough food, or sufficient education, to escape poverty.

This is particularly relevant in the context of fairness, because just as for incarceration, there are often large racial disparities in the distribution of wealth. In 2016, in the United States, the poverty rate of African-Americans and Hispanics (22% and 19.2%, respectively) was more than double the Caucasian poverty rate of 8.8% (Semega et al., 2017).

These two case studies both suggest that certain decisions made about an individual (granting parole) can cause feedback loops and affect the future decisions made about that same individual.

In the next section, we review previous work on fair classification, and see whether so far this problem has been addressed.

2.4 Previous work on fair classification

We distinguish between three classes of work in the fairness literature.

- Works that focus on synchronic fairness, i.e. the type of fairness that does not take into account the evolution of population distributions over time. This type of work mainly consists in identifying new mathematical formulations for fairness, and on designing algorithms that maximise fairness while remaining accurate (e.g. Zafar et al. (2015); Zemel et al. (2013); Zafar et al. (2017b); Calmon et al. (2017); Zafar et al. (2017a); Hardt et al. (2016)).
- Works that focus on impossibility results that show that given some natural constraints two fairness metrics are incompatible (e.g. Pleiss et al. (2017); Kleinberg et al. (2016); Chouldechova (2017); Berk et al. (2017)).
- The works that focus on diachronic fairness, i.e. the type of fairness whose chief consideration is the evolution of populations over time (e.g. Lum and Isaac (2016); Liu et al. (2018); Ensign et al. (2017)).

2.4.1 Synchronic fairness

A key assumption underlying a large part of synchronic fairness is that the true distributions of marginalised and non-marginalised populations are equal, or very close together, but that the individuals $x_1, \dots, x_n$ in the training set were not sampled uniformly from the world. In other words, the sampling is non-representative, as discussed in Chapter 1, and had the sampling been representative, then their classifiers' predictions would have been fair. In order to redress the effects of non-representative sampling, proponents of synchronic fairness design their models to output predictions that reflect the fact that the underlying distributions of marginalised and non-marginalised populations are identical. They do this by trying to match the statistics of their predictions of individuals from the marginalised group with the statistics of their predictions of individuals from the non-marginalised group.

There are a wide variety of these statistics. Barocas et al. (2018) list 21 of them in their book, too many to fully explain and enumerate. We will briefly define and illustrate those that will be useful to us later on:

- **P%-rule**, used by Zafar et al. (2017a).

  This metric is a measure of disparate impact, i.e. how differently two groups are being treated by a classifier. It is defined as the ratio of the fraction of the marginalised group assigned the positive class to the fraction of the non-marginalised group assigned the positive class.

  Formally, the P%-rule ppr is calculated as:

  $$ppr = \frac{P(y_i^* = 1 \mid s_i = 1)}{P(y_i^* = 1 \mid s_i = 0)}.$$

  A perfect P%-rule of $ppr = 1$ means that $y_i^* \perp s_i$, i.e. the predictions are independent of the sensitive attributes.

  More concretely, suppose a model tries to predict whether or not an individual will consume drugs within the next week. It finds that 80% of Caucasians will not consume drugs within the next week, but only that 20% of African-Americans will not. That model then has a significant disparate impact of $ppr = 20\%/80\% = 0.25$. If the true drug-consumption distributions of Caucasians and African-Americans are roughly equal (which Mauer (2007) argues is the case), then one would expect the model to predict roughly equal predictions for Caucasians and African-Americans, i.e. that ppr would be close to 1.

- **Equalized odds**, introduced by Hardt et al. (2016).

  This metric measures the difference between the rate at which each group receives true/false positives.

  Formally, equalised odds are calculated as:

  $$EqOdds = \frac{P(y_i^* = 1 \mid s_i = 1, y_i)}{P(y_i^* = 1 \mid s_i = 0, y_i)}.$$

  Hardt et al. (2016) argue that this is a richer metric for fairness, because it allows for y_i^* to depend on the sensitive attribute s_i *through* the true label y_i.

  A perfect equalised odds score of $EqOdds = 1$ means that $y_i^* \perp s_i \mid y_i$, i.e. the predicted labels are independent of the sensitive attributes, given the true labels.

- **Equal opportunity**, introduced by Hardt et al. (2016).

  This metric is similar to equalized odds, but is slightly more specific and simpler to implement. It measures the difference between true positive rates between the classifications received by different groups.

Formally, the true positive rates TPR for a group a is calculated as:

$$TPR_a = \frac{\text{number of times } y_a^* = y_a = 1}{\text{number of times } y_a^* = y_a = 1 + \text{number of times } y_a^* = 0 \text{ and } y_a = 1}.$$

Note that the number of times $y_a^* = y_a = 1$ is the number of true positives, and the number of times $y_a^* = 0$ and $y_a = 1$ is the number of false negatives.

And so equal opportunity metric between two groups a and b is defined as:

$$EO = \frac{TPR_a}{TPR_b}.$$

	$y^* = 0$	$y^* = 1$
$y = 0$	25_a / 50_b	30_a / 10_b
$y = 1$	40_a / 10_b	5_a / 30_b

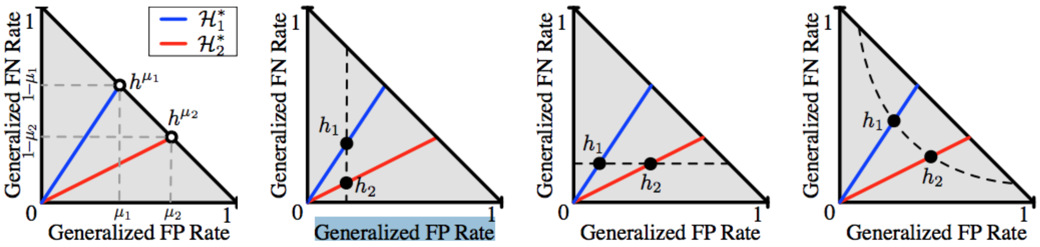
Table 2.1: Table illustrating the calculation of the equal opportunity metric

More concretely, suppose the classifier's task is to predict whether an individual will default on a loan, and that its findings are summarised in Table 2.1, for two groups a and b. Then the true positive rate for a would be quite low, since the classifier had relatively few true positives, and quite a few false negatives: $TPR_a = 5/(5+40) = 0.11$. However, the true positive rate for b would be quite high: $TPR_b = 30/(30+10) = 0.75$.

Note that the fairness metrics outlined above are all satisfied if there is no difference between the true distributions from which marginalised and non-marginalised individuals are drawn, $g_0(\cdot)$ and $g_1(\cdot)$. However, if those two distributions are different, then they no longer hold naturally, and machine learning practitioners need to devise techniques to maximise them while maintaining accuracy. These fairness metrics are also subject to impossibility results: some cannot be satisfied at the same time.

2.4.2 Impossibility results

Kleinberg et al. (2016) demonstrate that a classifier cannot simultaneously be calibrated (i.e. have $P(y_i = 0 \mid x_i) = P(y_i^* = 0 \mid x_i)$), and satisfy equalised odds (i.e. have



(a) Possible cal. classifiers $\mathcal{H}_1^*, \mathcal{H}_2^*$ (blue/red).

(b) Satisfying cal. and equal F.P. rates.

(c) Satisfying cal. and equal F.N. rates.

(d) Satisfying cal. and a general constraint.

Figure 2.1: Plot illustrating the impossibility result worked on by Pleiss et al. (2017), namely that a classifier cannot be both calibrated and satisfy equal opportunity if it classifies two populations whose means are different. Reprinted from Pleiss et al. (2017).

equal true positive and negative rates for both groups), except in special circumstances, such as when the classifier is capable of achieving perfect accuracy, or if the initial distributions of the two groups $g_0(\cdot)$ and $g_1(\cdot)$ are identical.

Pleiss et al. (2017) illustrate this with a visual proof in Figure 2.1. In the figure, the x-axis represents the false-positive rate of a classifier for a given population, and the y-axis represents the false-negative rate of a classifier for a given population. Here we have two populations, blue and red, with different populations means, μ_1 and μ_2, respectively. Pleiss et al. (2017) show that if the populations being classified have different means, and the classifier is calibrated, then the classifier's false-positive rate and false-negative rate must increase at different rates for different populations. This is why, in the diagram, the slope of the blue line is different from the slope of the red line: $\mu_1 \neq \mu_2$. In order for the classifier to be both calibrated and satisfy equalised odds (equality of false-positive and false-negative rates across groups), the classifier needs to find points that belong to both the red line and the blue line, which in this case is only possible if the classifier makes perfect predictions.

To complement this theoretical argument, Chouldechova (2017) uses real data from the aforementioned COMPAS dataset gathered by Angwin et al. (2016) to argue the same point, namely that reconciling calibration and equal opportunity is impossible if the base rates among groups differ. She found that COMPAS's predictions were calibrated, but did not satisfy equal opportunity. She also found that if she modified the predictions in such a way that they achieved equal opportunity, then the predictions were no longer calibrated. She attributed this to the difference in recidivism rates among African-Americans and Caucasians in the dataset[1].

Finally, Barocas et al. (2018) prove that under mild assumptions, equalised odds and the P%-rule cannot be satisfied at once. Specifically, if we assume that the sensitive attribute s is correlated with the true label y (i.e. $s \not\perp y$), and that the classifier's label y^* is correlated with the true label y (i.e. $y^* \not\perp y$), then equalised odds and the P%-rule cannot both be true, i.e. $\neg(y^* \perp s \ \wedge \ y^* \perp s \mid y)$.

The proof is brief enough to be reproduced below. Taking the contrapositive, we want to show that $(y^* \perp s \ \wedge \ y^* \perp s \mid y) \implies \neg(s \not\perp y \ \wedge \ y^* \not\perp y)$, i.e. that
$(y^* \perp s \ \wedge \ y^* \perp s \mid y) \implies (s \perp y \ \vee \ y^* \perp y)$

$P(y^* \mid s) = \sum_y P(y^* \mid s,\, y)P(y \mid s)$ by one version of the law of total probability.
$P(y^*) = \sum_y P(y^* \mid y)P(y)$ by the other version of the law of total probability.
Simplifying the first equation using $y^* \perp s \ \wedge \ y^* \perp s \mid y$ gives:
$P(y^*) = \sum_y P(y^* \mid y)P(y \mid s)$.

Thus:

$\sum_y P(y^* \mid y)P(y) = \sum_y P(y^* \mid y)P(y \mid s)$, and since $y \in \{0,1\}$:
$P(y^* \mid y=0)P(y=0) + P(y^* \mid y=1)P(y=1) = P(y^* \mid y=0)P(y=0 \mid s) +$
$P(y^* \mid y=1)P(y=1 \mid s)$.
Letting $p_0 = P(y=0),\ p_s = P(y=0 \mid s),\ q_i = P(y^* \mid y=i)$, we obtain:
$p_0q_0 + (1-p_0)q_1 = p_sq_0 + (1-p_s)q_1$
$p_0(q_0 - q_1) = p_s(q_0 - q_1)$.

There are two ways in which this final equation is satisfied:
If $q_0 - q_1 = 0$, i.e. if $P(y^* \mid y=0) = P(y^* \mid y=1)$, i.e. if $y^* \perp y$,
or,
if $p_0 = p_s$, i.e. if $P(y=0) = P(y=0 \mid s)$, i.e. if $y \perp s$.
That is, $(s \perp y \ \vee \ y^* \perp y)$. *QED.*

[1] Note that this does not entail that African-Americans truly have a higher probability of recidivism than Caucasians. It could instead be due to non-representative sampling, as discussed in section 2.1.

2.4.3 Diachronic Fairness

Diachronic fairness is a type of fairness that explicitly takes into account and tries to harness the effects of the classifier on the world over time, and the distribution from which the data is drawn. We describe two examples of diachronic fairness below.

2.4.3.1 Predictive policing

An area in which this diachronic fairness has been studied is predictive policing, which involves police departments using algorithms and the geographic locations and timings of past crimes to try to predict the locations and timings of future crimes. The areas that the police departments then assign their officers to patrol are a function of these predictions (Perry, 2013). According to Lum and Isaac (2016), police departments in the United States and the United Kingdom are using the predictions of these predictive policing algorithms to dispatch officers to certain locations over others.

This is potentially problematic in certain geographic areas. For example, in Oakland, CA, historical police data on the locations of past drug crimes concentrate around historically African-American neighbourhoods (Lum and Isaac, 2016). As a result Lum and Isaac (2016) find that predictive policing algorithms disproportionately believe that African-American neighbourhoods are much likelier to be the scenes of drug crimes than Caucasian neighbourhoods, despite the fact that drug crimes happen at approximately equal rates across races. This is illustrated in Figure 2.2. They hypothesise that this may be due to the fact that if police are dispatched to a particular area in search of crime, they are more likely to find crime in that area than if they hadn't been sent to that area.

Indeed, they further found that if police being dispatched to an area increases the probability of finding crime in that area by 20%, then over time predictive policing algorithms would believe that a disproportionate amount of crime happens in those specific areas, in line with the phenomenon observed in Oakland.

Ensign et al. (2017) build on the work by Lum and Isaac (2016), and formalise it using the theory of urns (a reinforcement learning framework (Pemantle et al., 2007)), and give a justification grounded in theory for the feedback loop. They attempt to mitigate the effects of the feedback loop by assigning a lower weight to crimes found in parts of the city that are heavily policed, and a higher weight to crimes found in parts of the

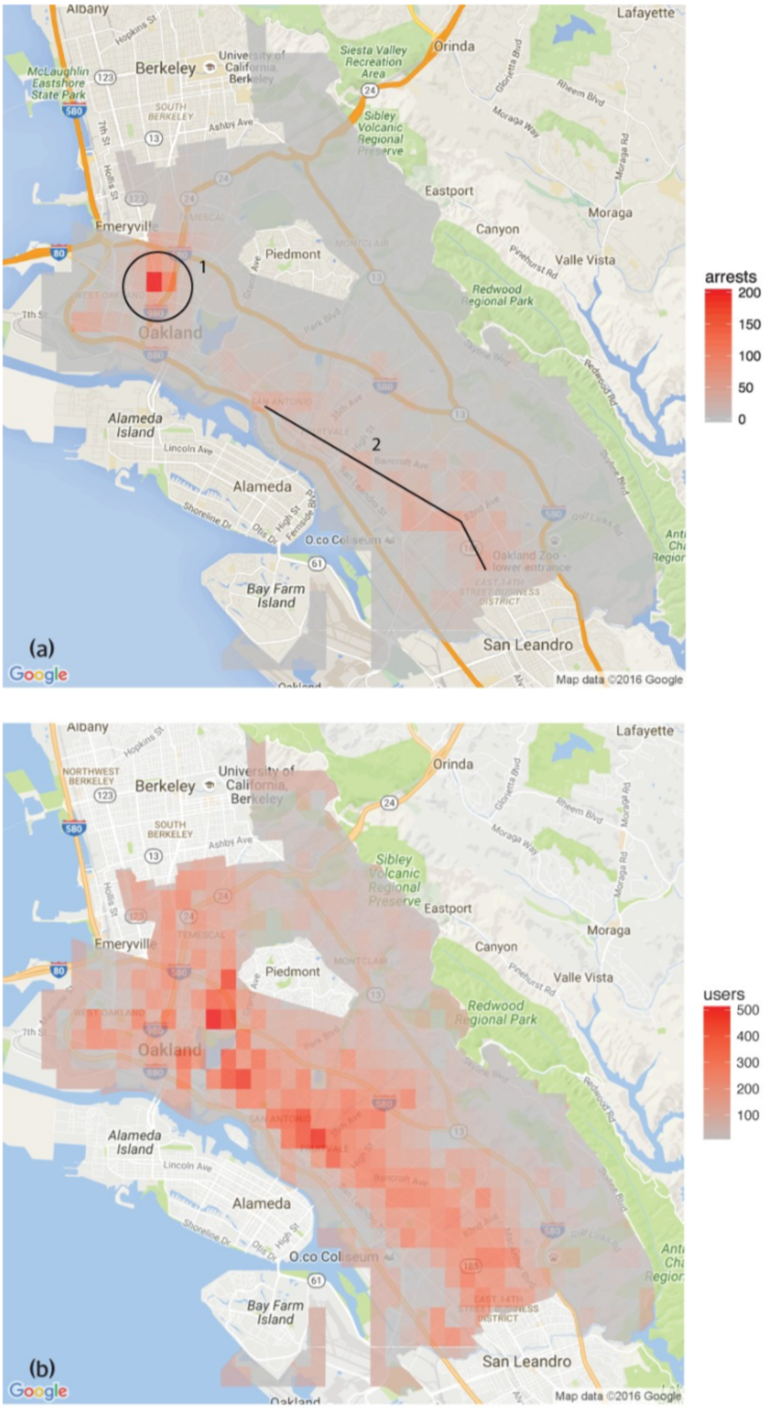


Figure 2.2: (a): number of drug related arrests that the Oakland police department made over the city in 2016. The area with the most arrests (circled) is Hoover-Foster, a neighbourhood historically home to a large African-American community (Crispell, 2015). (b): number of drug users in Oakland, estimated by Lum and Isaac (2016). Note that police target an area that does not have a disproportionate number of drug users. Reprinted from Lum and Isaac (2016).

city that are under-policed.

2.4.3.2 Banking loans

Diachronic fairness has also been studied in the context of banking loans. Liu et al. (2018) construct a framework in which a bank gives loans to marginalised and non-marginalised populations at different rates based on their credit score distributions, and examine the direct effect that this has on the populations' credit score distributions and the bank's profit over one time-step. Our thesis builds on their framework, which is described in more detail in Chapter 3. In short, an individual's credit score decreases significantly if they are given a loan and do not pay back, and increases a little if they are given a loan and pay back (the higher the credit score the more likely an individual is to be granted a loan). Similarly, the bank loses money if it grants a loan to an individual that then defaults on it, and gains money if it grants a loan to an individual who pays it back. Using this framework, they explore the effects of imposing different fairness constraints on the means of the marginalised populations. The fairness constraints they use are Equal Opportunity (which they call `EqOpt`) and the P%-rule (which they call `DemParity`).

They provide a graphical way of explaining their framework, using what they call an "outcome curve", illustrated in Figure 2.3. The outcome curve describes how the mean of the marginalised population's credit score $\Delta\mu$ changes as a function of the bank's acceptance rate β (a number between 0 and 1 that represents what fraction of loan applications from the marginalised population the bank will grant loans to).

Within this framework, they find that imposing various fairness criteria (namely Equal opportunity and the P%-rule, as described in 2.4.1) on the bank's classification algorithm can result in three possible outcomes for the marginalised population's mean credit score: it can increase, decrease, or stay the same. These results also hold in comparison to the marginalised population's credit score when the bank optimises purely for profit, without any fairness constraints. In other words, they find that applying fairness constraints can lead to better, worse, or identical results to situations without fairness constraints. Visually, this means that $\beta^{\texttt{EqOpt}}$ and $\beta^{\texttt{DemParity}}$ can fall in either the "relative harm" spotted yellow zone, or the "relative improvement" blue zone, or the "active harm" striated red zone.

The fact that fairness constraints can lead to an acceptance rate in the "active harm"

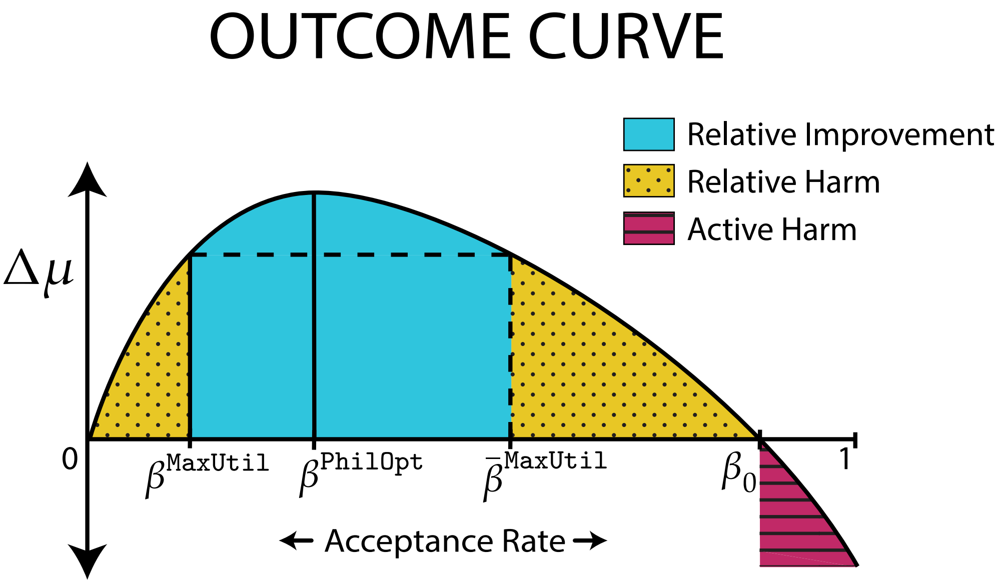


Figure 2.3: This figure illustrates the different possible outcomes for the marginalised population's mean credit score $\Delta\mu$ as a function of the bank's acceptance rate β. The points of the x-axis of note are:

$\beta^{\mathtt{MaxUtil}}$, which represents the acceptance rate for which the bank makes most money, without any fairness constraints.

$\beta^{\mathtt{PhilOpt}}$, which represents the acceptance rate of the highest point of the outcome curve, the acceptance rate for which the mean credit score of the marginalised population most increases.

$\bar{\beta}^{\mathtt{MaxUtil}}$, which represents the acceptance rate up until which $\Delta\mu$ is above its value at $\beta^{\mathtt{MaxUtil}}$.

β_0, which represents the point at which $\Delta\mu = 0$ ($\beta_0 \neq 0$).

As for the colours, the dotted yellow section represents the suboptimal acceptance rates where the marginalised population's mean credit score would increase if the bank simply tried to maximise its profit with no fairness constraint. In contrast, the blue section represents acceptance rates where the marginalised population's mean credit score would *decrease* if the bank reverted to maximising its profits with no fairness constraint. Finally, the red striped section represents acceptance rates where the marginalised population's mean credit score decreases. Reprinted from Liu et al. (2018).

zone –the zone in which the marginalised population's mean credit score decreases when the bank imposes a fairness constraint– is counterintuitive, as one would expect the fairness constraint to lead to better outcomes for the marginalised population. The rest of the thesis explores this phenomenon, and tries to qualify it more precisely.

2.5 Summary

In this chapter, we formalised the distinction between synchronic and diachronic fairness. We introduced Crawford's distinction between representational and allocative harm, and argued that allocative harm was particularly relevant to diachronic fairness, as those types of harm can theoretically have a large impact on the world, and thus on the classifier's future predictions (Crawford, 2017). We described two case studies of allocative harm occurring in the worlds of criminal justice and financial well-being, and found that poverty could beget future poverty, and incarceration could beget future incarceration, suggesting that allocative harm is not just a theoretical construct, and is thus worth studying further. We then reviewed past work on synchronic fairness, focusing on defining metrics for fairness, as well as work on how some of the metrics were incompatible, and cannot be satisfied at the same time. Finally, we introduced three works on synchronic fairness, including the paper by Liu et al. (2018), on whose credit score and banking framework this thesis relies on heavily.

Chapter 3

Methods

This Chapter details the specifics of my experiments simulating the effects of different types of fair classifiers on populations over time. We outline the general framework that we use, namely a bank deciding how to lend to individuals in a population given their credit score. We then describe the way we generate marginalised and non-marginalised populations, and introduce the optimisation problem that we tackle, namely how the bank maximises profit while minimising unfairness. We also introduce the metrics by which we will evaluate the performance of different fairness constraints.

3.1 Simulation methodology

In order to study the effects of imposing different kinds of fairness constraints to a classifier, and study the effects of those fairness constraints over a single time-step, we create simulated populations, and adopt the framework outlined by Liu et al. (2018), described below.

3.1.1 Framework

There are two populations, call them the blue and orange populations, that applied for a loan, and they have different initial credit score distributions, g_b and g_o respectively, with mean μ_b and μ_o, respectively. The blue population is marginalised, the orange is non-marginalised, and so $\mu_b < \mu_o$. The credit scores go from 1 to 100. The credit score of each individual is proportional to that individual's probability of paying back the loan. The bank loses 3£ if the applicant does not pay back the loan, and gains 1£ if they do. The applicants gain 1 point of credit score if they are granted the loan and pay back, and lose 2 points of credit score if they are granted the loan and default on it. The bank decides who gets a loan and who does not based on a credit score cutoff threshold. The threshold is dependent on the applicant's group. If an individual from the blue group has a credit score below the bank's blue cutoff of c_b, that individual does not get a loan. If their credit score is greater or equal to c_b, they get a loan. Similarly, if an individual from the orange group has a credit score below the bank's orange cutoff of c_o, that individual does not get a loan. If their credit score is greater or equal to c_o, they get a loan.

3.1.2 Generating the populations

We choose to represent each population using a one-hundred dimensional vector $v = (v_1, \ldots, v_{100})$, where v_1 is the number of people with credit score 1, v_2 is the number of people with credit score 2, etc. We generate v by using a one-dimensional Brownian motion random walk. We first set $v_1 = 0$. Then over 100 iterations we populate the vector such that:

$$v_{i+1} = v_i + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0,4).$$

We then round all elements of the vector to integers, and shift it up to ensure that the smallest element is 0. Using this method, we generate two distributions, compare the mean of each, and dub the one with the lowest mean the marginalised population v_b, and the one with the highest mean the non-marginalised population v_o. The mean μ is computed as:

$$\mu = \frac{\sum_i v_i \times i}{\sum_i v_i},$$

where i is a credit score and v_i is the number of people with that credit score.

Finally, we multiply each element of each vector by a constant, such that the vector sums to 1,000, to equalise the populations. This creates a pair of distributions, blue and orange, such that the mean of the blue distribution by construction is always smaller than the mean of the orange distribution. Six examples of these simulated distributions can be found in Figure 3.1.

The Brownian motion random walk allows us to generate many different types of distributions. This means that we were not constrained to choose between assuming that score distributions were normal or bimodal or log-normal or uniform, making our results more general[1].

3.1.3 Creating the default/pay-back labels

We assume that the credit scores of both populations are directly proportional to the probability of each individual in the population defaulting. This allows us to generate the subsets of each population that default on their loans, and those that pay back, by giving each individual a probability of default p equal to their credit score c divided by one hundred: $p = c/100$.

Examples of the distributions of these groups can be found in Figure 3.2.

3.1.4 Maximising profit with and without fairness constraints

After creating 1,000 pairs of blue and orange group distributions, we use an off-the-shelf optimiser to find the bank different credit score thresholds c_b and c_o for the blue and orange groups in two scenarios. In the fair scenario, we try to find c_b and c_o such that the bank maximises profit, all while satisfying the equal opportunity constraint described in Chapter 2, i.e. that FPR_b and FPR_o have to be within 1% of each other. In the other scenario, the optimiser simply tries to find c_b and c_o such that the bank maximises profit, with no fairness considerations. Suppose that imposing a cutoff of c_b for the blue population and c_o for the orange population leads to the bank's profit being pb_b and pb_o for the blue and orange populations respectively.

Then formally, when only maximising profit, the bank will find thresholds c_b and c_o such that $pb_b + pb_o$ is maximised, and when imposing fairness constraints, the bank will find thresholds that maximise $pb_b + pb_o$ while satisfying $EqOdd = 1 \pm 0.01$ or $ppr = 1 \pm 0.01$.

We used scipy's (Jones et al., 2001) SQLSP (Sequential Quadratic Programming) optimiser, the default option for scipy's minimize function, because it is fast (one iteration takes six seconds on average), and we needed to run many thousands of iterations for

[1] It is worth noting that FICO credit scores –one of the biggest credit score company in the U.S.– do not appear to follow a normal or bimodal or log-normal distribution (Bubb and Kaufman, 2014), nor would we expect other scores to.

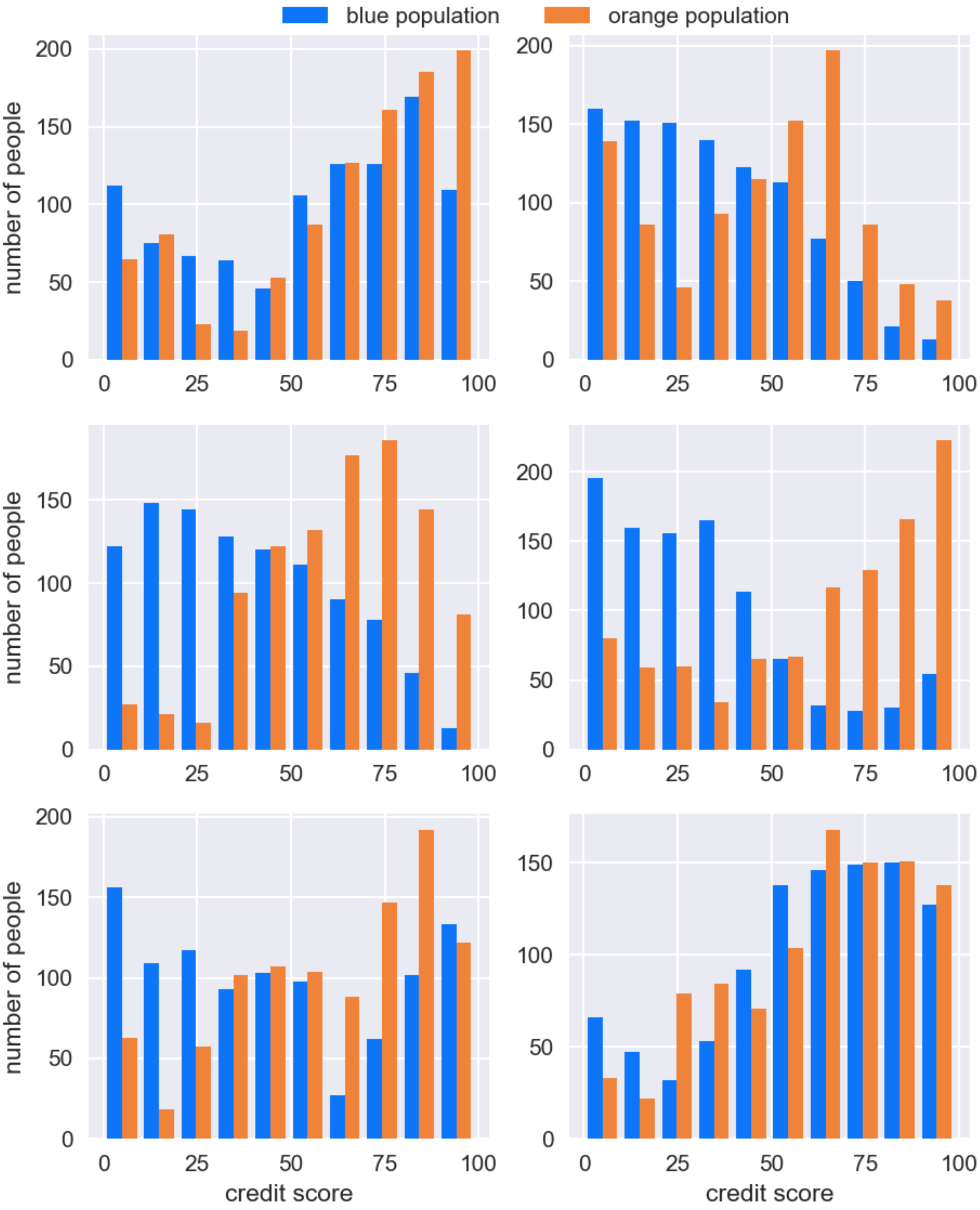


Figure 3.1: Examples of six random populations that we generated using a one-dimensional Brownian motion random walk. Note that the mean of the blue population is always lower than the mean of the orange population.

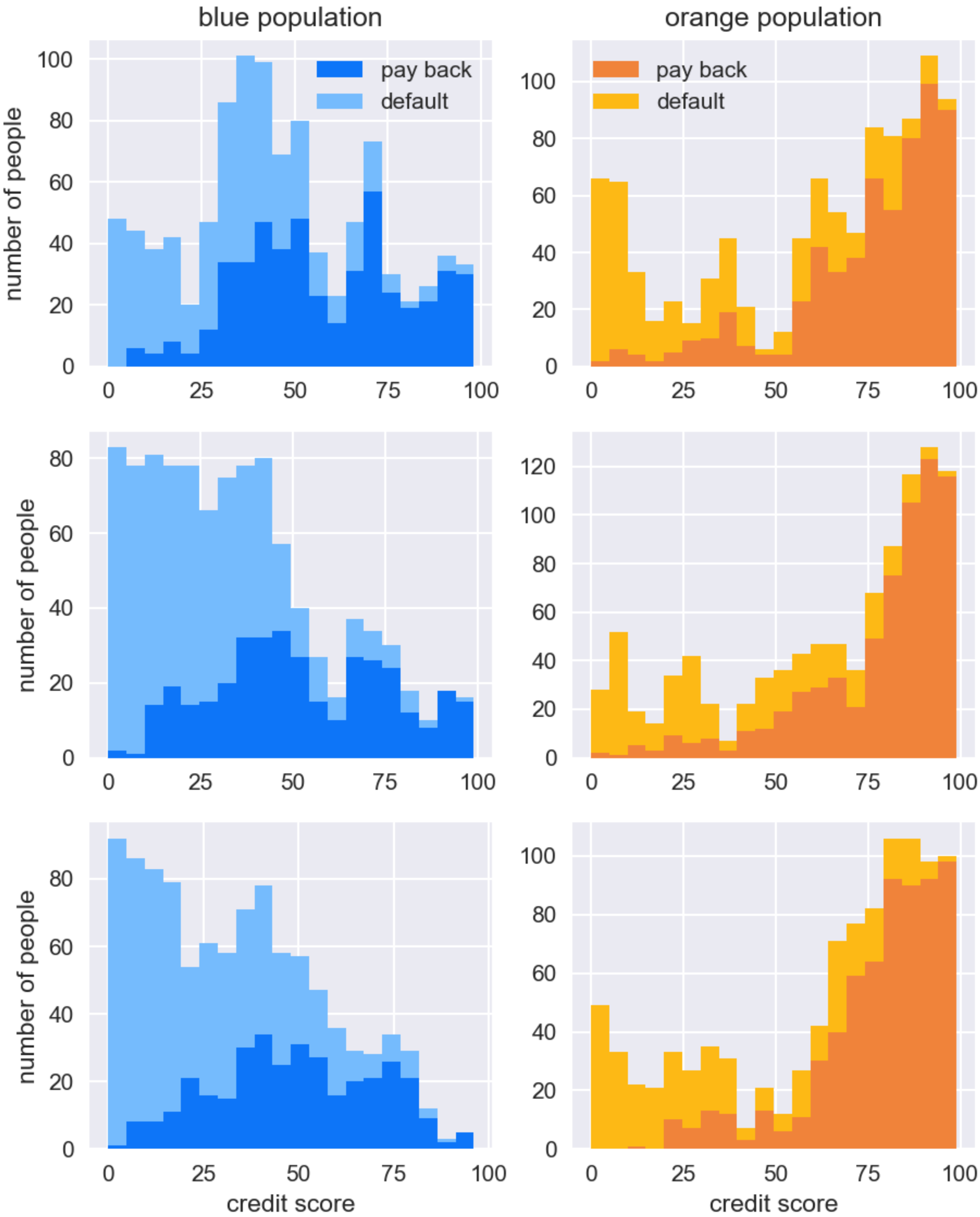


Figure 3.2: Examples of three pairs of blue and orange populations, and their respective rates of paying back and defaulting, respectively. Note that the default rate is proportional to the credit score, and that the mean of the blue population is always lower than that of the orange one.

each experiment (1,000 iterations for each of our experiments with a single time-step, and 5,000 for each of our experiments over multiple time-steps).

Having described the framework of the simulation, we now describe specific experiments aimed at identifying the effects of synchronic fairness constraints –such as equalised odds and the P%-rule– on marginalised populations over time.

3.2 One-step experiment

We first study the effects of an optimisation run over a single time-step, as in Liu et al. (2018), to see the preliminary effects of fairness constraints over a short time. In order to do so, we generate blue and orange populations, and find cutoffs c_b and c_o that maximise the bank's profit with and without equal opportunity. We then update the distributions according to these cutoffs: if an individual was given a loan but defaulted, they lose 2 credit points; if an individual was given a loan and paid back, they gain 1 credit point. We then compute the means of the new distributions, and store them. We repeat this process 1,000 times, in order to get a representative sample, and repeat it 1,000 more times with the P%-rule fairness constraint instead of the equal opportunity fairness constraint.

3.2.1 Questions

Our underlying objective is to understand whether measures of synchronic fairness (here equalised odds and the P%-rule) are effective at increasing fairness in a diachronic setting. In order to do so, we ask the questions:

1. What proportion of times does imposing fairness constraints result in a worse outcome for the marginalised population than imposing no fairness constraint? Similarly, what proportion of times does imposing fairness constraints help, or make no difference?
2. What is the magnitude of the effect of imposing fairness constraints on the marginalised population? In other words, is it the case that when imposing fairness constraints leads to a reduction in mean credit scores, those reductions are relatively small, but when they lead to increases in credit score, those increases are substantial?

3. What is the relationship between the magnitude of the effect of imposing fairness constraints and the initial distance between the two populations? One might think that imposing fairness constraints would have a larger effect if the marginalised and non-marginalised populations are particularly far apart –i.e. if the situation is particularly unfair–, and would have a smaller effect if the populations are closer together.

In order to answer these questions, we define metrics that formalise these notions.

3.2.2 Evaluation metrics

To answer 1. and 2., following Liu et al. (2018), we define $\Delta\mu_b$ to be the difference between the change in mean of the marginalised blue population after an iteration with and without a fairness constraint imposed. Recall that the mean μ of a distribution of credit scores defined by $v = (v_1, \ldots, v_{100})$ is:

$$\mu = \frac{\sum_i v_i \times i}{\sum_i v_i}.$$

Define the change in mean of the marginalised blue population after an iteration with fairness as $\delta\mu_{b,f}$, and the change in mean of the marginalised blue population after an iteration without fairness as $\delta\mu_{b,u}$. Then $\Delta\mu_b = \delta\mu_{b,u} - \delta\mu_{b,f}$.

Furthermore, to answer 3., define the initial means of the blue and orange populations (prior to any decision made by the bank) as $\mu_b^{(0)}$ and $\mu_o^{(0)}$, respectively. Then define the initial distance between the two distributions as $\Delta\mu^{(0)} = \mu_o^{(0)} - \mu_b^{(0)}$. Note that $\Delta\mu^{(0)}$ will always be positive, since $\mu_o^{(0)} > \mu_b^{(0)}$ by construction.

3.3 10-step experiment

We then study the effects of an optimisation run over 10 time-steps, changing the bank's cutoffs c_b and c_o at every iteration, in order to see the effects of fairness constraints over longer time periods. The method is similar to that of the single-step experiment, except that this time instead of updating the populations a single time after an optimisation, we update them 9 more times, at each iteration storing the full population distributions.

3.3.1 Questions

Here, we are focused on understanding the effects of applying synchronic fairness techniques in a diachronic setting. In order to do so, we ask the questions:

1. What proportion of times does imposing fairness constraints result in a worse outcome for the marginalised population than imposing no fairness constraint, over time? In conjunction with this question, we are interested in the difference between the answers to this question, and the answer to question 1. above in section 3.2.1. It is conceivable that over a single iteration, fairness constraints lead to a worse outcome, but that over time these worse outcomes are erased, or exacerbated.

2. As in the single-step case, what is the magnitude of the effect of imposing fairness constraints on the marginalised population over time?

3. What types of pairs of orange and blue distributions lead to the largest reduction in fairness?

4. What effect does imposing fairness constraints have on other properties of the population distributions over time? In particular, how does it affect the standard deviation of the marginalised blue population?

3.3.2 Evaluation metrics

To answer 1., 2., and 3., if a fairness constraint was applied to the optimiser at time $t-1$, then we define the mean of the blue and orange distributions at time t to be $\mu_{b,f}^{(t)}$ and $\mu_{o,f}^{(t)}$, respectively. If no fairness constraints were applied, then we define them as $\mu_{b,u}^{(t)}$ and $\mu_{o,u}^{(t)}$.

Similarly, to answer 4., if a fairness constraint was applied at time $t-1$, then we define the standard deviation of the blue and orange distributions at time t as $\sigma_{b,f}^{(t)}$ and $\sigma_{o,f}^{(t)}$, and as $\sigma_{b,u}^{(t)}$ and $\sigma_{o,u}^{(t)}$ if no fairness constraints were applied.

Chapter 4

Results

In this chapter, we present our findings, namely:

- Fairness constraints such as equal opportunity and the P%-rule can do more harm than good in a single time-step, and lead to a decline in the marginalised population's credit score relative to a situation without fairness constraints.
- This effect is exacerbated in the long run, over multiple time-steps.
- There is no meaningful effect of the initial distance between the means of the two populations.
- However, there are certain patterns of populations that are particularly prone to making fairness constraints lead to worse outcomes for the marginalised population.
- Imposing fairness constraints leads to a greater variability in the population change, i.e. a larger population standard deviation.

4.1 Simulation over one time-step

In this section, we focus on running a single iteration of the bank's optimisation process, and analysing the results. We answer the questions:

- How does the mean population credit score change after the bank applies the P%-rule and equal opportunity fairness constraints to its optimisation procedure?

- How does the initial distance between the blue and orange distributions affect the change in distribution?

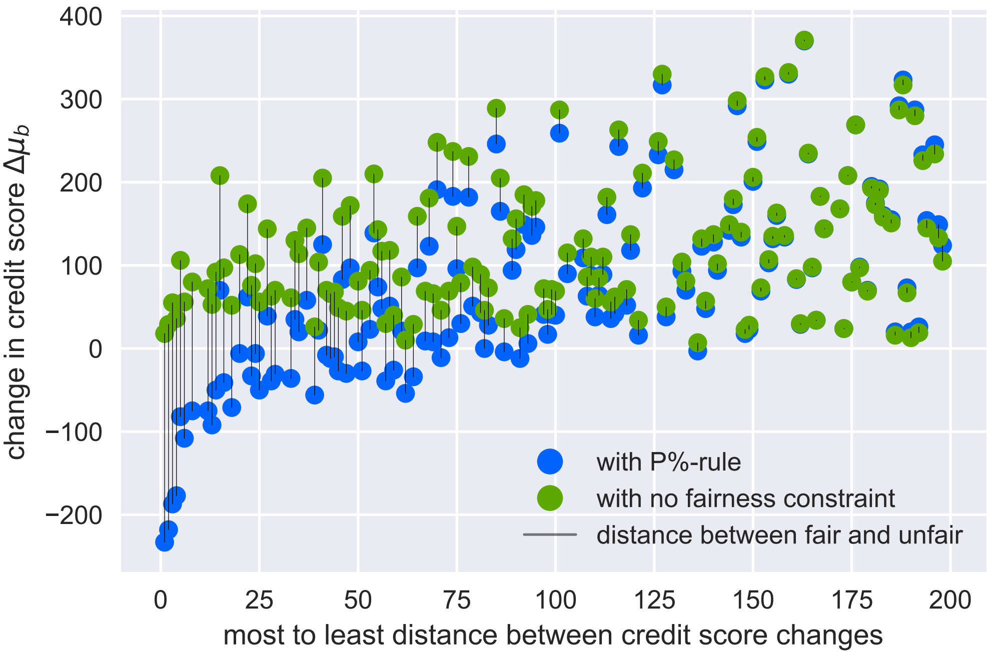
4.1.1 Imposing a fairness constraint harms the marginalised population more often than it helps it

After running the optimisation detailed in section 3.2, we plot the change in credit scores for the fair and unfair methods, ordered with respect to the distance between the credit score change for the fair and unfair methods, for 200 distributions of the disadvantaged blue population. The 200 distributions were subsampled uniformly without replacement from the 1,000 original distributions, to make the plot more legible. Our findings are summarised in Figure 4.1.

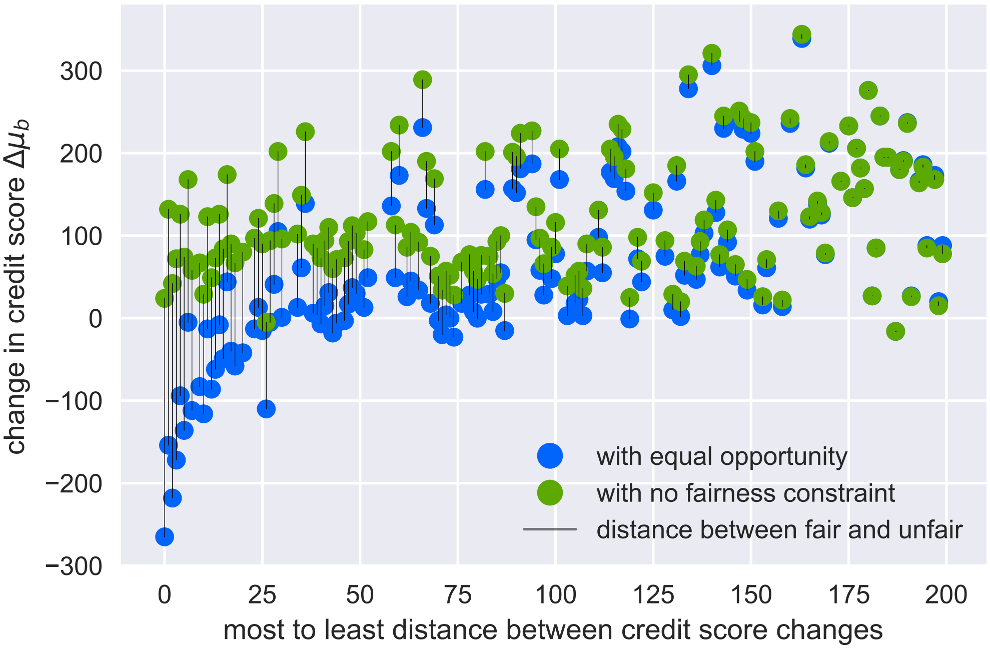
We find that in a narrow sense, Liu et al.'s (2018) claim that an optimisation method that enforces equal opportunity can lead to all possible outcomes for the disadvantaged population is true. Recall that $\Delta\mu_b$ is the difference between the change in credit score of the blue population after an iteration with and without fairness constraints imposed, i.e. $\Delta\mu_b = \delta\mu_{b,u} - \delta\mu_{b,f}$.

Then, comparing the distance between fair and unfair methods, we find that $\Delta\mu_b > 0$, or $\Delta\mu_b < 0$, or $\Delta\mu_b \approx 0$ are all possible outcomes of the optimisation. That is, we find that imposing fairness constraints can lead to an increase, a decrease, or stagnation of the marginalised group's credit score.

This can be seen graphically. The blue points (representing the credit change obtained after imposing the equal opportunity fairness constraint) on the left side of the plot are lower than their corresponding green points (representing the credit change obtained after not imposing fairness constraints). Other blue points are higher than their corresponding green points on the right side of the plot. Finally, some blue are almost indistinguishable from their corresponding green points, towards the middle right side of the plot. This holds for both the P%-rule constraint (cf Figure 4.1a) and the equal opportunity constraint (cf Figure 4.1b). However, a closer analysis finds that Liu et al.'s (2018) claim that an optimisation method that enforces equal opportunity can lead to all possible outcomes for the disadvantaged population needs to be qualified more carefully. For example, we find that $\Delta\mu_b < -10$, i.e. imposing a fairness constraint is better than not imposing one for the marginalised group, occurs about 10% of the



(a) Change in credit score $\Delta\mu_b$ after imposing P%-rule fairness constraint.



(b) Change in credit score $\Delta\mu_b$ after imposing equal opportunity fairness constraint.

Figure 4.1: Ranking of the difference between change in credit score for the marginalised population after imposing/not imposing (a) P%-rule and (b) equal opportunity fairness constraints, from least to most. The leftmost points represent the distributions where imposing the fairness constraint performed worst, and the rightmost points represent the distributions where imposing the fairness constraint performed best.

time. However, $\Delta\mu_b > 10$, i.e. imposing a fairness constraint is detrimental to the marginalised group, happens more, around 35% of the time. Finally, $\Delta\mu_b = 0 \pm 10$, i.e. imposing or not imposing the fairness constraint makes only a minor difference to the marginalised group, happens the remaining 55% of the time. Cf Table A.1 for more details. The same numbers hold for the P%-rule fairness constraint.

4.1.2 The magnitude of the harm is greater than the magnitude of the help

Furthermore, we also find that the magnitude of the difference between the change in credit score of the blue population with fairness is substantially different from that without fairness. Indeed, when $\Delta\mu_b < 0$, i.e. imposing a fairness constraint is better than not imposing one, $\mid \Delta\mu_b \mid$ tends to be small. However, when $\Delta\mu_b > 0$, i.e. imposing a fairness constraint is worse than not imposing one, $\mid \Delta\mu_b \mid$ tends to be large.

This can be seen graphically: the distance between the points on the leftmost side of Figure 4.1 –i.e. the points for which imposing the fairness constraint performed led to a large decrease in credit score for the marginalised population– is significantly larger than the distance between the points on the rightmost side of the figure –the points for which imposing the fairness constraint performed well. The same finding holds for the P%-rule fairness constraint.

So we found that, for a large number of distributions, imposing fairness constraints is detrimental to the marginalised population, and decreases their credit score more often than it increases it. We now investigate the relationship between the initial distance between blue and orange populations and the change in the marginalised population's mean, with and without fairness constraints.

4.1.3 The initial distance between blues and oranges doesn't correlate with the magnitude of the effects of fairness

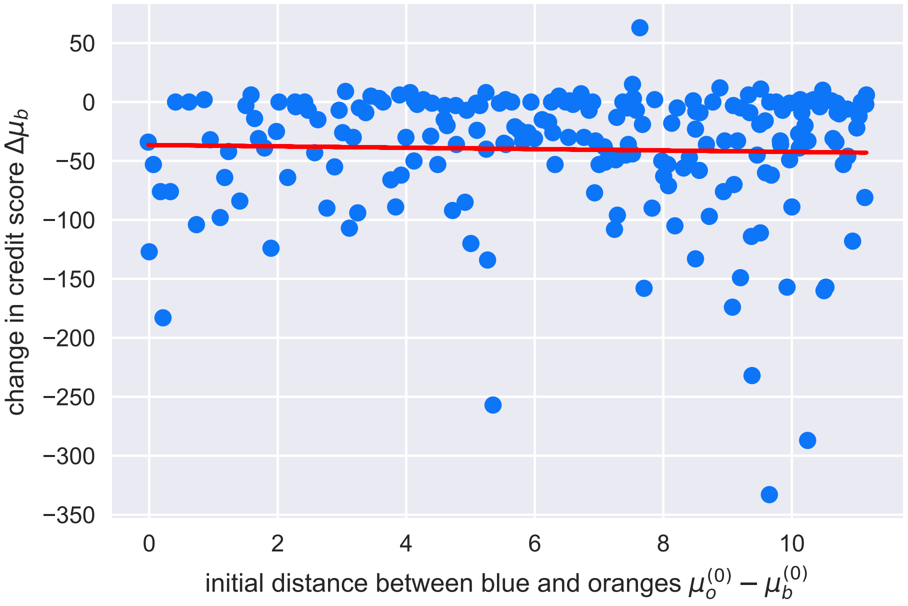
One might hypothesise that the fairness constraints are most effective when there is a large difference between the two initial distributions. Indeed, if the distance between the two initial distributions is large, then the need for fairness grows, and so we would want our fairness constraints to be particularly effective.

In order to establish whether or not the initial distance between $\mu_b^{(0)}$ and $\mu_o^{(0)}$ has an impact on the magnitude of $\Delta\mu_b = \delta\mu_{b,u} - \delta\mu_{b,f}$ (a measure of how poorly fairness constraints perform relative to no fairness constraints), we plot $\Delta\mu_b$ with respect to $\mu_o^{(0)} - \mu_b^{(0)}$, where $\Delta\mu_b$ was obtained with respect to the P%-rule and equal opportunity fairness constraints. In order to make the plot less cluttered, we subsampled 200 points from the original 1,000 points, uniformly at random and without replacement. We used the full 1,000 points to generate the red trend lines. Our results are summarised in Figure 4.2.

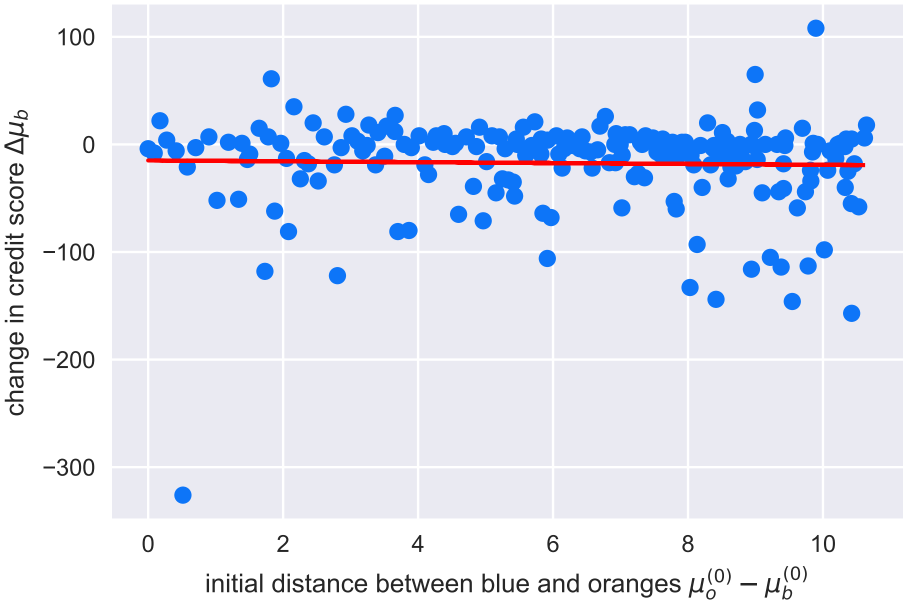
We find that there seems to be no obvious correlation connecting the distance between the two initial distributions of oranges and blues and the size of the negative effects of imposing the fairness constraint. More prosaically, this means that imposing fairness constraints performs no better in cases where the initial distributions between blue and orange populations were particularly far apart (or particularly unfair). The red trend lines do appear to be sloping somewhat downwards, both in the equal opportunity plot Figure 4.2a, and in the P%-rule plot Figure 4.2b. However, their Pearson correlation coefficients are both low, at -0.03 (p-value 0.70) and -0.02 (p-value 0.64) respectively, and so we do not find that the two variables are correlated.

4.1.4 Summary

In summary, we examined the effects of fairness constraints over a single time-step, and found that they can lead to both a sharp decline, or a mild improvement, or no noticeable change in the mean credit score of the marginalised population. We found that the magnitude of the negative effects of fairness constraints are on average larger than their positive effects. Finally, we found no correlation between the initial distance between marginalised and non-marginalised populations and the magnitude of the change in credit score $\Delta\mu_b$. We now turn our attention to whether or not these effects occur in simulations with longer time-steps.



(a) $\Delta\mu_b$ as a function of $\mu_o^{(0)} - \mu_b^{(0)}$ after optimising with the eq. oppt. fairness constraint.



(b) $\Delta\mu_b$ as a function of $\mu_o^{(0)} - \mu_b^{(0)}$ after optimising with the P%-rule fairness constraint.

Figure 4.2: Plots of the difference between the change in credit scores between fair and unfair methods with respect to $\mu_o^{(0)} - \mu_b^{(0)}$. The trend lines appear mostly flat, suggesting that future change in credit score is not a function of the initial distance between distributions, and thus that imposing fairness constraints when the initial means of the blue and orange distributions is no more helpful than imposing fairness constraints where the blue and orange distributions are closer together.

4.2 Simulation over multiple time-steps

In this section, we focus on running 10 steps of the optimisation process, in order to see whether there is a substantial difference between fairness constraints enforced on one run versus multiple runs. We answer the questions:

- What proportion of times does imposing fairness constraints help the marginalised population, and how does this compare to a single-step iteration?
- What is the magnitude of the effects, and how do they compare to a single-step iteration?
- How do multiple iterations affect the standard deviations of the marginalised distributions?

4.2.1 Over multiple time-steps, imposing the P%-rule harms the marginalised population more than it helps

After running the experiment detailed in section 3.3, we plot the mean $\mu_{b,u}^{(\cdot)}$ of the marginalised blue population without fairness with respect to its mean $\mu_{b,f}^{(\cdot)}$ with fairness, over time. We did the same for the non-marginalised orange population. For the sake of legibility, we only plotted 10 trajectories of the 500 we ran, but the full plot is included in Figure A.1 of the Appendix. Our main findings are summarised in Figure 4.3.

We find that most of the time, after 10 time-steps, imposing a fairness constraint is about the same as not imposing one, and that $\mu_{b,f}^{(10)}$ and $\mu_{b,u}^{(10)}$ are within 2 credit score point of each other around 92% of the time. We also find that $\mu_{b,f}^{(10)}$ is significantly larger (more than 2 credit score points larger) than $\mu_{b,u}^{(10)}$ without fairness around 1% of the time. Finally, we find that $\mu_{b,f}^{(10)}$ is significantly smaller (more than 2 credit score points smaller) than $\mu_{b,u}^{(10)}$ around 7% of the time. It is worth noting that a difference of 2 credit score points is quite significant. It is the sort of gap that would occur if the entirety of the blue population were given a loan, and they all defaulted. Further details can be found in Table A.1.

This can be seen graphically in Figure 4.3a, which illustrates the evolution of the means of blue populations over time, with and without fairness constraints.

The y-axis is the mean $\mu_{b,u}^{(\cdot)}$ of the marginalised blue population *without* a fairness constraint. Its x-axis is the mean $\mu_{b,f}^{(\cdot)}$ of the marginalised blue population *with* a fairness constraint. The colour of the points marks the evolution over time: darker points mean more time has elapsed. This means that if the points head northeast over time, up the diagonal $y = x$ line, then the mean μ_b is increasing over time for iterations with and without fairness constraints, and so the optimiser is equally fair with and without fairness constraints. However, if the points head northwest, then *not* having a fairness constraint is improving the mean credit score of the blue population, and *having* a fairness constraint is decreasing the mean credit score. This is because when points go northwest, they are going up the y-axis, increasing $\mu_{b,u}^{(\cdot)}$ without fairness constraints, and going down the x-axis, decreasing $\mu_{b,f}^{(\cdot)}$ with fairness constraints.

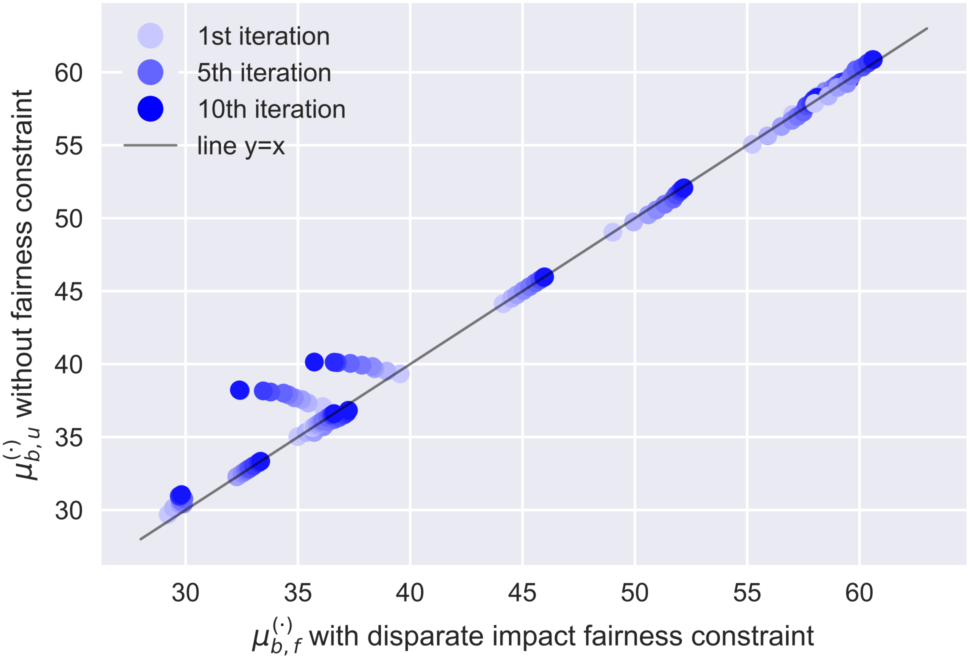
In Figure 4.3a (and also in Figure A.1), we see that while most trajectories increase along the diagonal line, some of them diverge from this line, and head northwest quite strongly, suggesting that in these cases applying a fairness constraint to the optimisation was counterproductive.

Furthermore, we found that for the most part the non-marginalised orange population fared better than the marginalised population, with or without fairness constraints, but did a little better without constraints. Indeed, $\mu_{o,f}^{(\cdot)}$ with fairness and $\mu_{o,u}^{(\cdot)}$ without fairness are within 2 credit score point of each other around 97.5% of the time; $\mu_{o,f}^{(\cdot)}$ is significantly larger (more than 2 credit score point larger) than $\mu_{o,u}^{(\cdot)}$ around 0.5% of the time; $\mu_{o,f}^{(\cdot)}$ is significantly smaller (more than 2 credit score point smaller) than $mu_{o,u}^{(\cdot)}$ around 2% of the time.

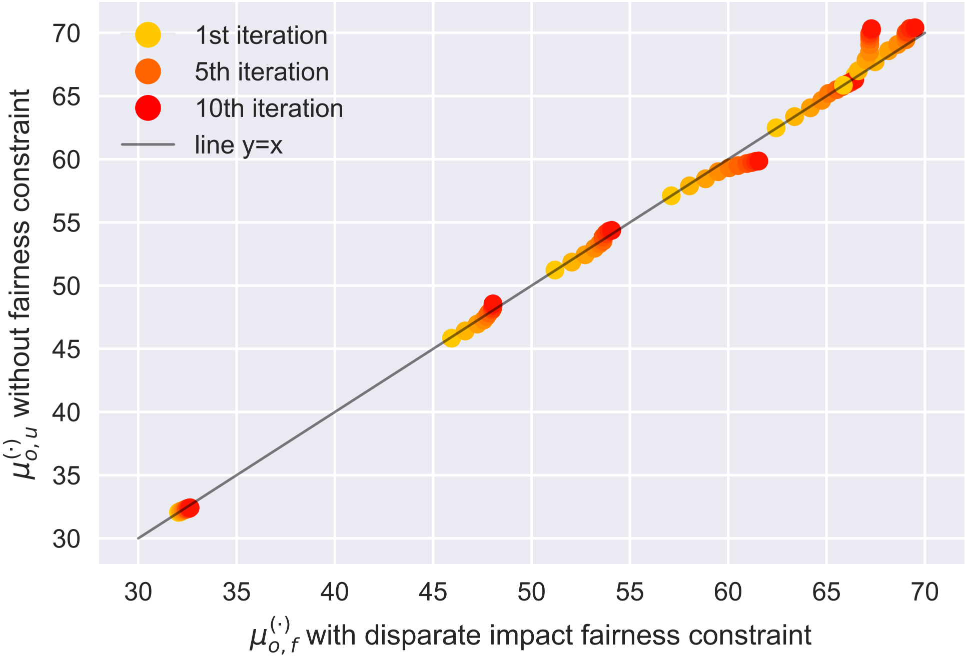
This can also be seen graphically in Figure 4.3b and Figure A.1, where for the most part it appears that the orange trajectories head northeast up the $y = x$ line, and rarely head northwest, at least compared to the marginalised blue population.

4.2.2 Distributions in which most of the non-marginalised population have high credit scores lead to the P%-rule performing worse

7% of marginalised populations suffered a significant drop in mean credit score when fairness constraints were applied to the optimisation process, relative to an optimisation process without fairness constraints. We wish to see if we can find properties that these



(a) Mean of the blue population without fairness with respect to with fairness, over 10 iterations



(b) Mean of the orange population without fairness with respect to with fairness, over 10 iterations

Figure 4.3: Evolution of the (a) marginalised and (b) non-marginalised populations' means over time, with and without fairness constraints. For the most part, the fairness constraints don't seem to have an effect. However, when they do, they have a negative impact on the marginalised blue population.

populations all share.

In order to do so, we plotted the six populations whose mean credit score fell most (i.e. where $\mu_{b,u}^{(\cdot)} - \mu_{b,f}^{(\cdot)}$ was largest) in Figure 4.4, for the P%-rule fairness constraint.

A characteristic that these populations appear to share is that they all have a high concentration of the non-marginalised orange populations in the highest credit score brackets. We suspect that this leads the optimiser to find that it is most profitable for the bank to overlend to the blue population, most of which are in the low credit score range where they're likely to not pay back, because the profit that the bank gains from loaning to a high fraction of the orange population makes up for the bank's losses with the blue one.

Indeed, suppose that 10,000 members of the blue population were uniformly distributed between credit scores 1 and 100 (so there would be 100 blues with credit score 1, 100 with credit score 2, etc.), and the 10,000 members of the orange population were uniformly distributed between credit scores 81 and 100 (so there would be 500 oranges with credit score 81, 500 oranges with credit score 82, etc.)[1]. Suppose again that individuals default proportionally to their credit score, and so 100% of people with credit score 100 would pay back, 99% of people with credit score 99 would pay back, etc. This means that there would be 100 blues with credit score 100 who pay back, $100 \times 0.99 = 99$ blues with credit score 99 who pay back, etc.

Running the optimiser without constraints in this situation returns a cutoff of $c_b = 66.7$ and $c_o = 80$, leading to a total *increase* in the blue population's credit score of $\delta\mu_{b,u} \approx 1,700$ points. However, running the optimiser with a disparate impact P%-rule constraint returns a cutoff of $c_b = 1$ and $c_o = 80$, leading to a total *decrease* in the blue population's credit score of $\delta\mu_{b,f} \approx -4,650$ points.

As a result, the difference $\Delta\mu_b$ between enforcing a P%-rule fairness constraint and no fairness constraint is substantial, with $\Delta\mu_b \approx -6,350$, the equivalent of each member of the marginalised population losing more than half a credit score point after a single iteration. Compare this $\Delta\mu_b$ to those in Figure 4.1, where the largest drops in blue credit score $\Delta\mu_b \approx -250$, which in a population of 1,000 individuals leads to an average decrease of ≈ 0.25 credit score points for an individual. This suggests that having a high concentration of oranges in the top credit score range leads to the P%-rule performing poorly.

[1] We use 10,000 instead of our usual 1,000 to avoid decimals in the number of people who default.

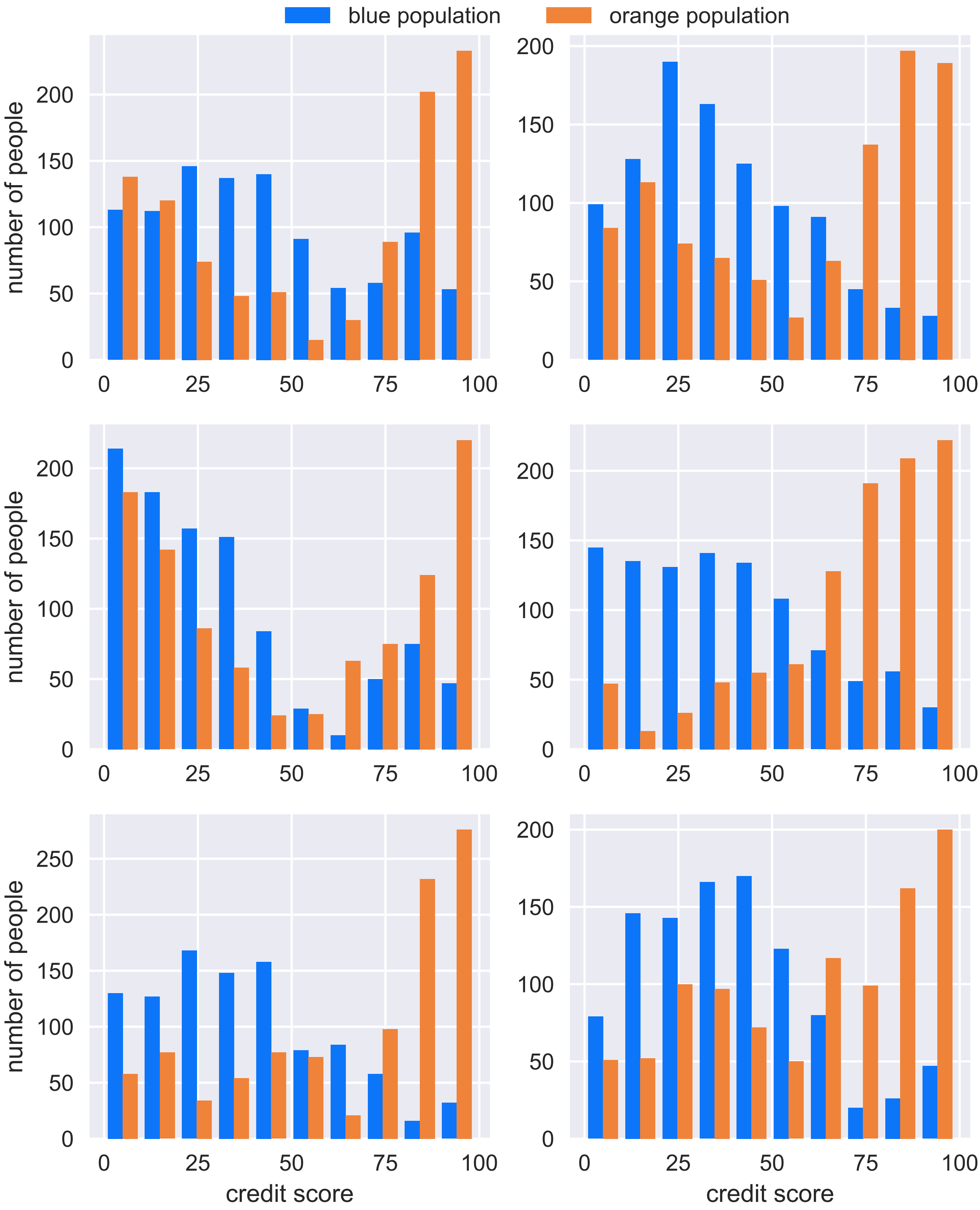


Figure 4.4: Examples of six pairs of distributions which led to the biggest drop in the marginalised population's mean credit score when imposed a fairness constraint, relative to no fairness constraint.

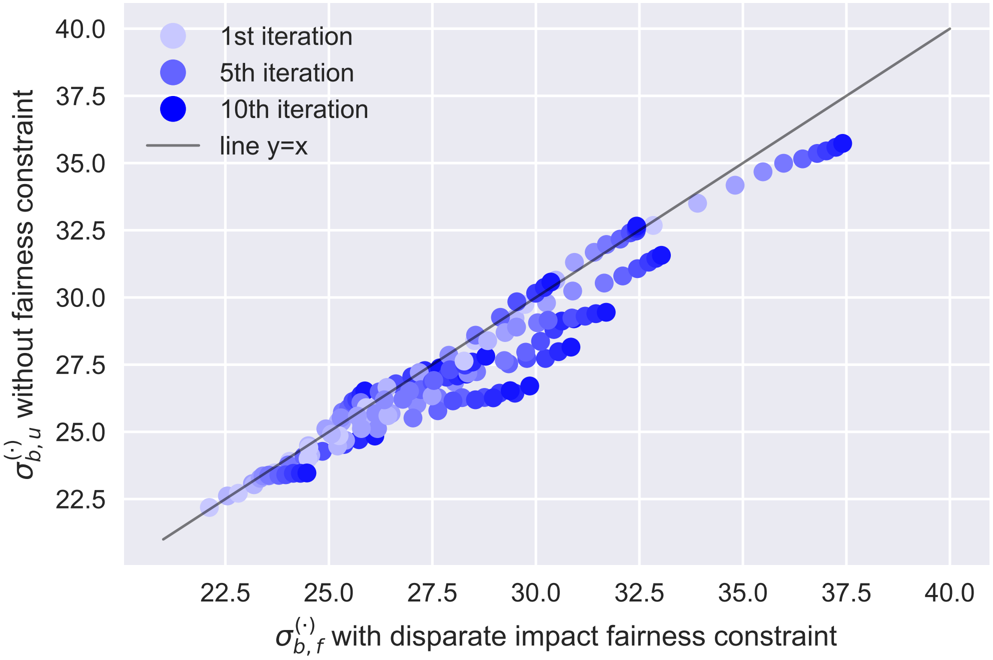
4.2.3 Fairness constraints increase the standard deviation of the marginalised population over time

Although we have primarily focused on the impact of fairness constraints on the mean of the populations, it is also worth investigating the effects of the constraints on other relevant statistics of the population. Here, we investigate the evolution of the standard deviations of the blue and orange populations, $\sigma_{b,\cdot}^{(\cdot)}$ and $\sigma_{o,\cdot}^{(\cdot)}$ respectively, over 10 iterations. Our results are plotted in Figure 4.5. We chose to only plot 10 pairs of trajectories, rather than the full 500, in order to make the plot more clear. The plot with all trajectories plotted can be found in the Appendix, in Figure A.2.

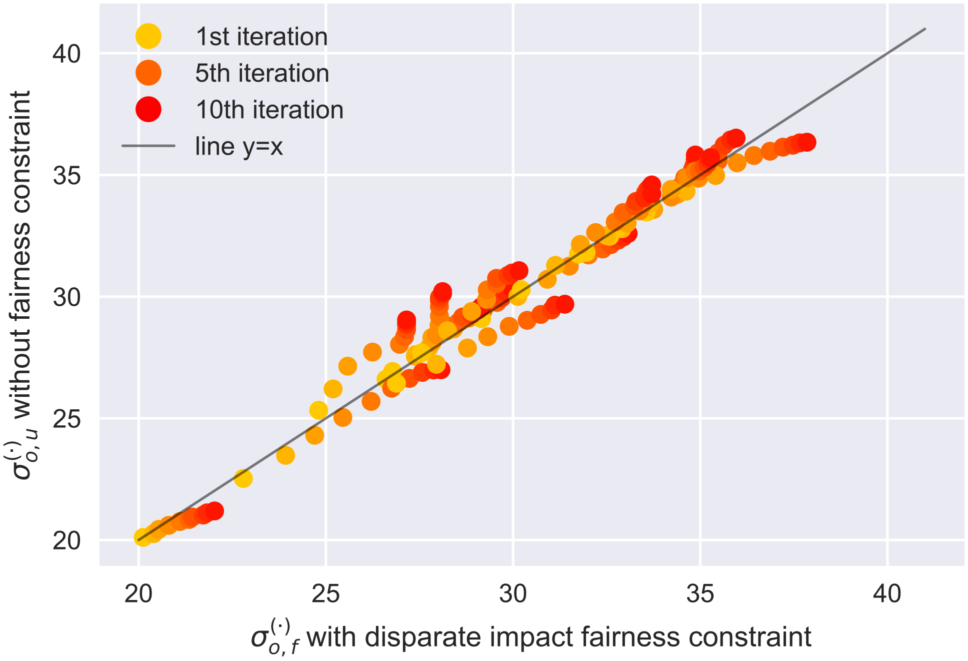
We find that for the most part, $\sigma_{b,f}^{(\cdot)}$ grows significantly faster than $\sigma_{b,u}^{(\cdot)}$ when the optimisation features a fairness constraint than when it does not. The standard deviation of the orange population $\sigma_{o,\cdot}^{(\cdot)}$, however, does not seem to have a propensity to lean one way or another.

We can see this graphically in Figure 4.5: the slope of the blue trajectories is usually smaller than 1 (which is the slope of the diagonal line $y = x$), which means that $\sigma_{b,\cdot}^{(\cdot)}$ is growing faster with the fairness constraint than without. The slope of the orange trajectories, on the other hand, seem about equally likely to be above or below 1 (i.e. above or below the diagonal line).

If the standard deviation of the marginalised distribution $\sigma_{b,\cdot}^{(\cdot)}$ grows faster *with* fairness than *without* fairness constraints, then imposing fairness constraints during the optimisation process leads to individuals in the blue population getting further away from the mean of the blue population at a faster clip than they would if no fairness constraints were imposed. This might account for why the fairness constraint causes the blue population's mean to decrease dramatically in 7% of simulated populations: the fairness constraint leads to many more individuals going further from the mean (in the negative direction), pushing the mean down, and repeating this cycle 10 times. The fact that the fairness constraint increases the standard deviation $\sigma_{b,\cdot}^{(\cdot)}$ allows the population's mean to decrease faster.



(a) Standard deviation of the blue population without fairness with respect to with fairness, over 10 iterations, for 10 trajectories.



(b) Standard deviation of the orange population without fairness with respect to with fairness, over 10 iterations, for 10 trajectories.

Figure 4.5: (a) Marginalised and (b) non-marginalised populations' standard deviations over time, with and without fairness constraints, for 10 trajectories.

Chapter 5

Conclusion

5.1 Summary

We investigated two case studies in the world of criminal justice and financial well-being, and found that poverty could beget future poverty, and incarceration could beget future incarceration.

This led us to draw and formalise a distinction between synchronic and diachronic fairness: the former is concerned with fairness in an instant, the latter with fairness over time.

We simulated the dynamics of marginalised and non-marginalised populations over time in a hypothetical bank loan situation, and applied two synchronic fairness constraints to the bank's loans. We compared the effects that those constraints had to similar situations without fairness constraints.

Our results raise doubts that synchronic fairness metrics succeed at causing fair outcomes in the long run, and instead suggest that imposing fairness constraints has the capacity to exacerbate differences between populations, and increase unfairness.

5.2 Limitations and future work

A significant limitation of our thesis was that in all of our experiments the size of the populations were equal, with 1,000 individual in each group. One could easily imagine fairness constraints having a different effect if the populations were significantly imbalanced, and the marginalised population were significantly larger or smaller than the non-marginalised one. There are marginalised populations that are equal in size to their respective non-marginalised populations (e.g. there are about as many women as men), so studying populations of equal size is not without merit. However, there are other cases in which there are significantly less people in the marginalised population than non-marginalised population, and vice versa. To draw examples from history, in 1933, during the rise of Nazism, Jews represented a very small fraction of the German population, at approximately 0.75% (United States Holocaust Memorial Museum, 2018); in India, in 1931, the marginalised lower castes Dalit and Shudra constituted 70% of the population (Vinod, 2010). This could be addressed by combining the literature around fair classification with the literature on imbalanced data (He and Garcia, 2009; Provost, 2000).

Furthermore, our experiments mostly ignored other effects that occur in parallel to banking loans and that could have an effect on a population's average credit score, what Liu et al. (2018) call "background economic variables". This makes our experiments somewhat unrealistic. To remedy this, we could increase the complexity of the simulation, for example by incorporating the phenomenon of regression to the mean (Barnett et al., 2004) to the simulation, where the blue marginalised population was naturally pushed towards the orange non-marginalised population, and seen what effect the fairness constraints would have had in this environment. We could also have gone further, and incorporated "poverty traps" into the dynamics of the simulation, as discussed in Chapter 2.

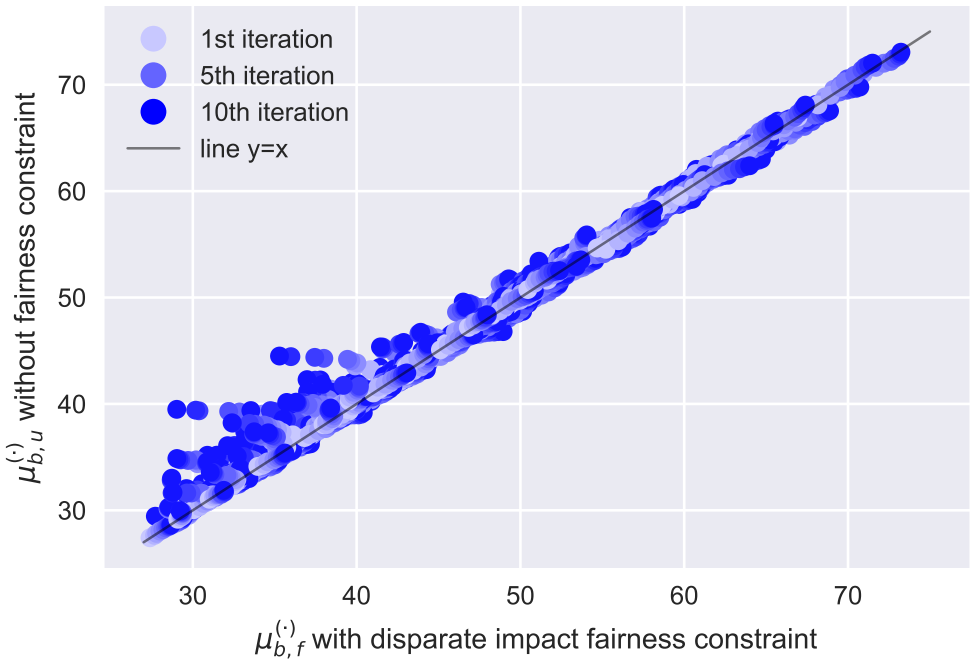
In conclusion, when building a machine learning system, practitioners ought to pay as much attention to achieving fairness in the short run as achieving fairness in the long run, because as we have demonstrated, those two goals do not necessarily go hand in hand.

Appendix A

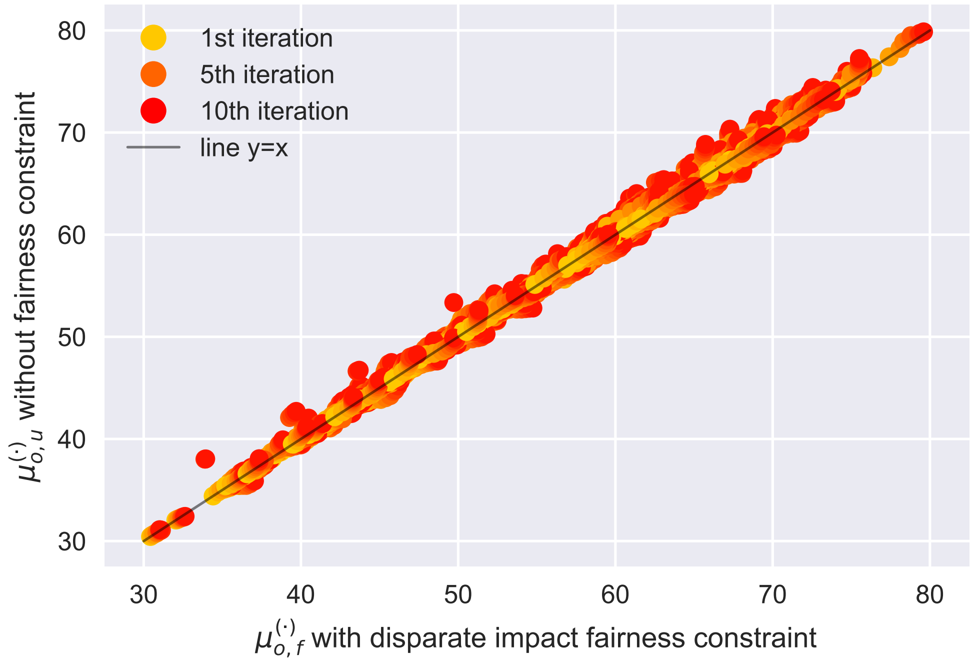
Additional plots and and tables

Distance between fair and unfair	Expression	Fraction
distance of 1 for blue pops (over 10 iterations)	$\mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} > 1$ $\mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} < -1$ $\mid \mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} \mid \leqslant 1$	4% 17% 79%
distance of 1 for orange pops (over 10 iterations)	$\mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} > 1$ $\mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} < -1$ $\mid \mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} \mid \leqslant 1$	4% 16% 80%
distance of 2 for blue pops (over 10 iterations)	$\mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} > 2$ $\mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} < -2$ $\mid \mu_{b,f}^{(10)} - \mu_{b,u}^{(10)} \mid \leqslant 2$	1% 7% 92%
distance of 2 for orange pops (over 10 iterations)	$\mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} > 2$ $\mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} < -2$ $\mid \mu_{o,f}^{(10)} - \mu_{o,u}^{(10)} \mid \leqslant 2$	0.5% 2% 97.5%

Table A.1: Table summarising the fraction of trajectories for which μ_b with fairness did better, about the same, and less well than μ_b without fairness. We define “about the same” as μ_b with fairness and μ_b without fairness as being either within 1 or 2 credit points of each other.

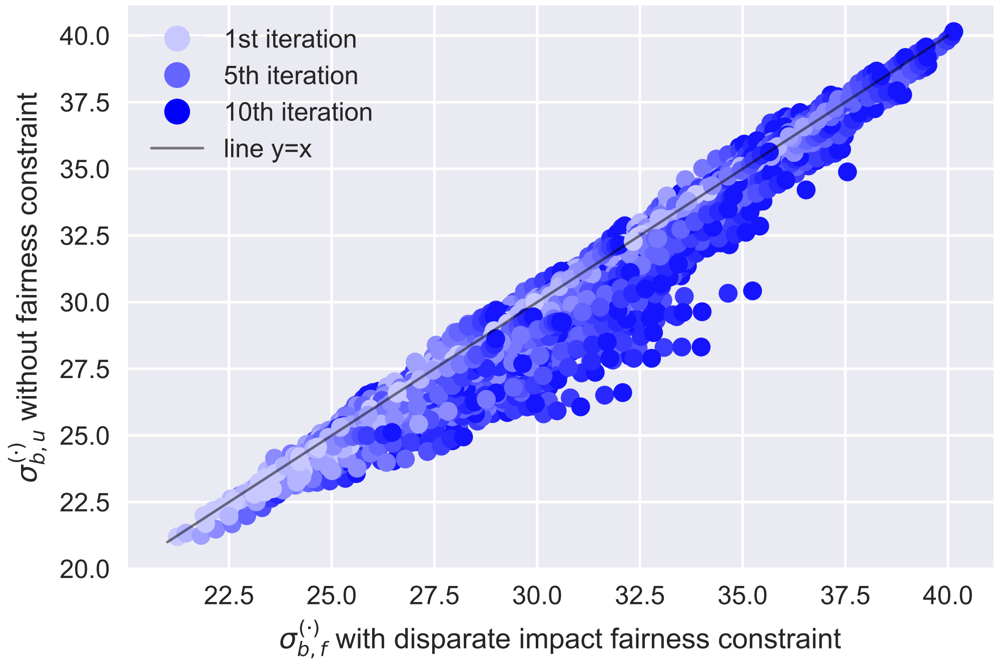


(a) Plot of all 500 $\mu_{b,\cdot}^{(t)}$ with fairness versus without fairness. Towards the bottom left of the plot, a fair number of trajectories seem to be heading northwest, suggesting that the fairness constraint is decreasing $\mu_{b,\cdot}^{(t)}$ over time.

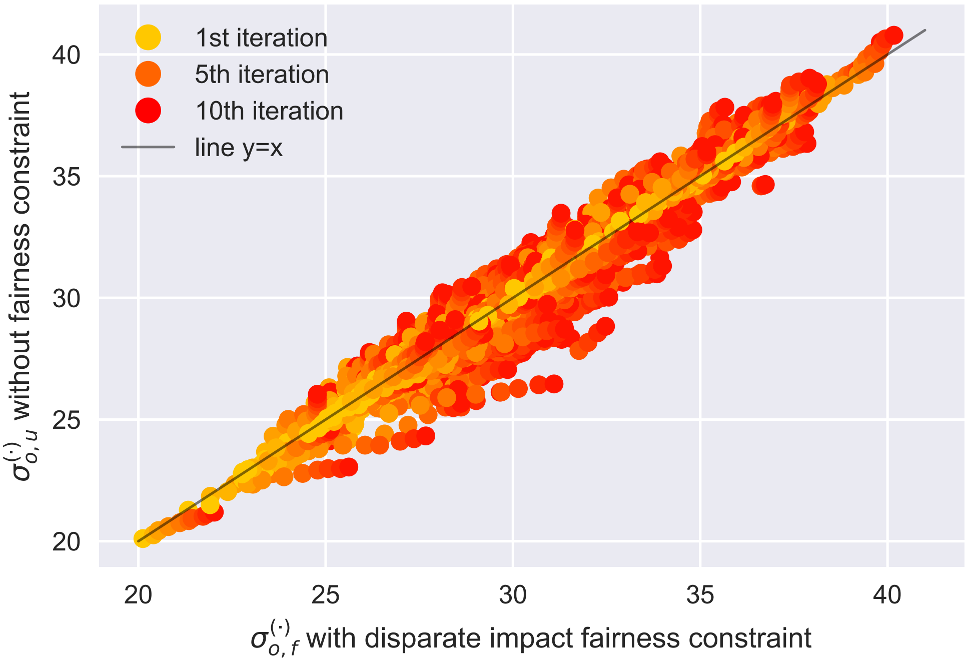


(b) Plot of all 500 $\mu_{o,\cdot}^{(t)}$ with fairness versus without fairness. For the most part, the trajectories seem to be on the diagonal, meaning that fairness constraints neither hurt nor harm.

Figure A.1: Plot of all 500 (a) $\mu_{b,\cdot}^{(t)}$ and (b) $\mu_{o,\cdot}^{(t)}$ with fairness versus without fairness.



(a) Standard deviation of the blue population without fairness wrt with fairness, over 10 iterations, for 500 trajectories. There seems to be more trajectories heading east than west.



(b) Standard deviation of the orange population without fairness wrt with fairness, over 10 iterations, for 500 trajectories. The distribution of trajectories seems to be equal in both directions.

Figure A.2: (a) Marginalised and (b) non-marginalised populations' standard deviations over time, with and without fairness constraints, for 500 trajectories. The plot suggests that imposing the P%-rule leads to more movement in the blue population.

Bibliography

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. [Online; posted 23-May-2016].

Bales, W. D. and Piquero, A. R. (2012). Assessing the impact of imprisonment on recidivism. *Journal of Experimental Criminology*, 8(1):71–101.

Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, 34(1):215–220.

Barocas, S., Hardt, M., and Narayanan, A. (2018). *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*.

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2016). Incarceration, recidivism and employment. Technical report, National Bureau of Economic Research.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Bowles, S. (2006). Institutional poverty traps. *Poverty traps*, pages 116–38.

Bubb, R. and Kaufman, A. (2014). Securitization and moral hazard: Evidence from credit score cutoff rules. *Journal of Monetary Economics*, 63:1–18.

calipsa.io (2018). About Us. `http://calipsa.io/about/`. Online; accessed 31 July 2018.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3995–4004.

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

Crawford (2017). The trouble with bias. NIPS Keynote: `https://www.youtube.com/watch?v=fMym_BKWQzk`. Online; accessed 31 July 2018.

Crispell (2015). “past is prologue” in oakland’s macarthur bart neighborhoods. `http://www.urbandisplacement.org/blog/past-prologue-oaklands-macarthur-bart-neighborhoods`. Online; accessed 15 August 2018.

Cullen, F. T., Jonson, C. L., and Nagin, D. S. (2011). Prisons do not reduce recidivism: The high cost of ignoring science. *The Prison Journal*, 91(3_suppl):48S–65S.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

FICO (2018). Can Machine Learning Build a Better FICO Score? `http://www.fico.com/en/blogs/risk-compliance/can-machine-learning-build-a-better-fico-score/`. Online; accessed 31 July 2018.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

ideal.com (2018). Automated Resume Screening Software Using AI. `https://ideal.com/product/screening/`. Online; accessed 31 July 2018.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed July-August 2018].

Jonson, C. L. (2010). *The impact of imprisonment on reoffending: A meta-analysis*. PhD thesis, University of Cincinnati.

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

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Mauer, M. (2007). Testimony of Marc Mauer. Technical report, House Judiciary Subcommittee on Crime, Terrorism, and Homeland Security.

Mauer, M. and King, R. S. (2007). A 25-Year Quagmire: The War on Drugs and Its Impact on American Society. Technical report, The Sentencing Project.

Mottola, G. R. (2013). In our best interest: Women, financial literacy, and credit card behavior. *Numeracy*, 6(2):4.

Nagin, D. S., Cullen, F. T., and Jonson, C. L. (2009). Imprisonment and reoffending. *Crime and justice*, 38(1):115–200.

Pemantle, R. et al. (2007). A survey of random processes with reinforcement. *Probability surveys*, 4:1–79.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5684–5693.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, pages 1–3.

Roodman, D. (2017). The impacts of incarceration on crime. Technical report, The Open Philanthropy Project.

Semega, J. L., Fontenot, K. R., and Kollar, M. A. (2017). Income and poverty in the united states: 2016. *Current Population Reports*, pages 10–11.

Sneider, H. and Sickmund, M. (2006). Juvenile Offenders and Victims: 2006 National Report. Technical report, National Center for Juvenile Statistics.

United States Holocaust Memorial Museum (2018). Germany: Jewish population in 1933. Holocaust Encyclopedia: `https://www.ushmm.org/wlc/en/article.php?ModuleId=10005276`. Online; accessed 20 August 2018.

Vinod (2010). Counting castes. Caravan: `http://www.caravanmagazine.in/perspectives/counting-castes`. Online; accessed 20 August 2018.

Wagner, P. and Sawyer, W. (2018). Mass Incarceration: The Whole Pie 2018. Technical report, Prison Policy Initiative.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017b). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 228–238.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2015). Learning fair classifiers. *arXiv preprint arXiv:1507.05259*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.