ENSURING FAIRNESS DESPITE DIFFERENCES IN ENVIRONMENT

Karan Deep Singh

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

YORK UNIVERSITY

TORONTO, ONTARIO

APRIL 2021

Abstract

Several fairness definitions have been proposed in the machine learning literature to rectify the issue of demographic groups being treated differently. Given the substantial research in the field, this work aims to provide an entry-level overview of the common definitions and metrics that are essential for a novice reader in the field. In addition, we propose a theorem, where we look at different population distributions and conditions under which our claim holds, that is the disadvantaged individual is expected to be more talented than similar performing advantaged individual. Finally, this work summarizes the six research works and discusses whether the result of our theorem is consistent in each of research work's model settings, culminating in a discussion of how all the authors view the world in terms of a group's talent distribution.

Acknowledgement

Throughout the writing of the thesis, I have received a great deal of support and assistance. I would first like to thank my supervisors, Professor Jeff Edmonds and Professor Ruth Urner, whose expertise was invaluable in formulating the research questions and methodology. Your expertise in the field of machine learning, theory, and mathematics really helped me to excel with the research. I would also like to thank my supervisory committee, Professor Aijun An and Professor Neal Madras for their valuable time to review my research work and for providing insightful tips.

Table of Contents

Abstract ii

Acknowledgements iii

Table of Contents iii

List of Tables ix

List of Figures x

1 Introduction **1**

1.1 Problem and Motivation 1

1.2 Research Objectives and Questions 3

1.3 Contributions 4

2 Common Literature & Background **6**

2.1 Related Work 6

2.2 Defining Sensitive Attributes 7

2.3 Fairness Metrics: Group v/s Individual Fairness 8

2.4 Group Fairness Policies and Impossibility Theorem 9

2.4.1 Demographic Parity 10

2.4.2 Equal Opportunity 10

2.4.3 Predictive Parity 10

2.4.4 Equalized Odds 11

2.4.5 Disparate Impact 11

2.4.6 Impossibility Theorem 12

2.5 Fairness Interventions Approaches 12

2.5.1 Pre-Processing 12

2.5.2 In-Processing/Constrained Optimization 13

2.5.3 Post-Processing 13

3 Theorem 1 **15**

3.1 Introduction 15

3.1.1 Model 16

3.1.2 Motivation 18

3.2 Uniform Talent and Environment Distribution 19

3.2.1 Merging Distributions $\langle T, E_A \rangle$ and $\langle T, E_B \rangle$ 19

3.2.2 Extreme x values and function r 21

3.3 Gaussian Distributions 23

3.4 Graceful Talent and Narrow Environment($r(x) = 2$) 25

3.5 Non-Graceful Talent or Wide Environment($r(x) = 0$) 28

3.5.1 Intuition Behind Log-Concave Distributions and Examples 31

3.5.2 Examples of Log-Concave Distributions 32

3.5.3 Non-Log-Concave Distributions 34

3.6 Appendix 36

3.6.1 Proof for Gaussian Distributions 37

3.6.2 Proof for Graceful Talent and Non-Extreme X ($r(x) = 2$) 39

3.6.3 Proof for Non-Uniform Talent or Extreme X Values ($r(x) = 0$) 44

3.7 Conclusion 47

4 Related Research Work Review **48**

4.1 Introduction 48

4.2 Research Review 1: Downstream Effects Of Affirmative Action 51

4.2.1 Model 53

4.2.2 Fairness Definitions 54

4.2.3 Main Results 54

4.2.4 Comparison to our Theorem 1 56

4.2.5 Worldview Comparison 56

4.3 Research Review 2: The Disparate Effects of Strategic Manipulation 57

4.3.1 Introduction 57

4.3.2 Model and Notion 57
4.3.3 Result 1: Equilibrium Analysis 59
4.3.4 Result 2: Learner Subsidy Strategy 60
4.3.5 Comparison with our work 60
4.3.6 Worldview Comparison 61
4.4 Research Review 3: Simplicity Creates Inequity 62
4.4.1 Model 62
4.4.2 Results 63
4.4.3 General Theorem 66
4.4.4 Comparison with our Theorem 1 66
4.4.5 Worldview Comparison 67
4.5 Research Review 4: From Fair Decision Making To Social Equality 68
4.5.1 Introduction 68
4.5.2 Model 69
4.5.3 Dynamics 70
4.5.4 Assumptions & Definitions 70
4.5.5 Results 72
4.5.6 Conclusion 74
4.5.7 Comparison with our final results 74
4.5.8 Worldview Comparison 74

4.6 Research Review 5: Delayed Impact of Fair Machine Learning 75

4.6.1 Contributions 75

4.6.2 Model 76

4.6.3 Outcome Curve 77

4.6.4 Results 78

4.6.5 Comparison with our final results 81

4.6.6 Worldview Comparison 82

4.7 Research Review 6: Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? 83

4.7.1 Introduction 83

4.7.2 Model 84

4.7.3 Bias in Training Data: 85

4.7.4 Results 85

4.7.5 Comparison with our final results 86

4.7.6 Conclusion 87

4.7.7 Worldview Comparison 87

5 Conclusion and Future Work **89**

5.1 Conclusion 89

5.2 Future Work 90

Bibliography **91**

List of Tables

4.1 Table showing the features and f-values(Table taken from EC'19 talk ([22])) 64

4.2 Simplification: Not considering $x^{\langle 2 \rangle}$ and γ (Table taken from EC'19 talk ([22])) . . 65

4.3 Simplification: Not considering only $x^{\langle 2 \rangle}$ (Table taken from EC'19 talk ([22])) . . . 65

4.4 Not Pareto Optimal (Table taken from EC'19 talk ([22])) 66

List of Figures

Figure 3.1 Group A and B's Distributions when Environment Range is Narrower than Talent's 20

Figure 3.2 Representing X'_A and X_B on the same image 20

Figure 3.3 Environment's range wider than talents. 21

Figure 3.4 Extreme Regions: Environment Range is Narrower vs Wider than Talent's . 22

Figure 3.5 Graceful Talent and Non-Extreme X 28

Figure 3.6 Though the values are not officially extreme because the range of the talent distribution is artificially large, the talent distribution is effectively uniform in a very narrow range, making all values extreme. In the (a) figure, this gives an example of a non-graceful T that does not have the properties stated in Theorem 4. In the (b) figure, because the B person is completely in the infinitesimal part of T, the probability of this occurring is infinitesimal. However, conditioned on this event happens, his expected T value is K bigger than A's. 29

Figure 3.7 An example of Log curve 31

Figure 3.8 Exponentially Increasing 32

Figure 3.9 Exponentially Decreasing 32

Figure 3.10 Combination Curve 33

Figure 3.11 Linearly Increasing Environment Distribution 33

Figure 3.12 Simple Concave Environment Distribution 34

Figure 3.13 Concave Ending in Uniform Functions 34

Figure 3.14 Single Step Fall Environment and Uniform Talent 35

Figure 3.15 Single Step Rise Environment and Uniform Talent 35

Figure 3.16 A Environment Distribution 36

Figure 4.1 Hu et. al.[18] Figure 1, Group cost functions for a one-dimensional feature x. τ_A and τ_B signify unmanipulated true thresholds. The threshold σ_A and σ_B perfectly classify group A and B candidates; A learner selects an equilibrium threshold $\sigma^* \in [\sigma_B, \sigma_A]$, committing false positives on group A (red bracket) and false negatives on group B (blue bracket). 59

Figure 4.2 Hu et. al.[18] (figure from FATML 2019 talk) Distribution for D_A and D_B with true thresholds τ_A and τ_B 61

Figure 4.3 Points showing unconstrained equilibriums(Figure 3 from the paper Srebro et. al. [26]) 73

Figure 4.4 Equilibrium points for AA^- and AA^+ respectively (Figure 3 from the paper Srebro et. al. [26]) 73

Figure 4.5 Outcome Curve (Figure 1 from Liu et. al. [25]) 78

Chapter 1

Introduction

Machine bias is prevalent in several machine learning applications where the classifier mimics the partial behaviour in the data used while training. With many recent studies proving vulnerability of these classifiers to the same biases as those of humans, such as the ProPublica Article[7] demonstrating the bias of Compass (an AI Software tool) against black individuals in predicting recidivism, the research in fairness has increased significantly[25] in the last decade.

Most research works[19, 18, 25, 1] in the ML literature assume that the disadvantaged group generally performs worse on standardized tests in the world. While such assumptions are pragmatic, these models portray a disparate talent distribution amongst groups. In this work, we assume that all the groups have a similar distribution of talents and it is the external factors that affect different groups differently.

1.1 Problem and Motivation

Given the substantial research approaches in the field of Machine Learning Fairness, the area has grown complex and is often hard to understand for a novice reader. This work seeks to provide

a comprehensive review (Chapter 2) of the state of the art theoretical fairness research works. This chapter would introduce a beginner to the terminology and common methods in the Fairness literature in machine learning (ML). We discuss generic concepts such as finding sensitive attributes and several approaches of formal measures of fairness (such as individual and group fairness). Furthermore, we discuss the fairness measures, which are the most common ways of assessing fairness of a ML classifiers, such as Demographic Parity and Equal Opportunity. We conclude this section with a brief discussion about the fairness intervention approaches found in the ML Literature. Such approaches can be broadly divided in 3 sub-sections i.e. pre-processing, in-processing and post-processing.

The second problem that we address in this work is to propose a generic theoretical model (Section 3.1.1) which instead of assuming a discriminatory view of the world, captures the sources of disadvantage for particular groups. This model is novel as compared to previous works[19, 18, 25, 1] as we consider the support environment available to an individual as one of the components in our model and we hypothesize that the environment consideration would help fix the bias discrepancy.

Considering the model, we start with the research problem where we show that if two individuals have a similar performance score on a test (such as on a Job Interview or SATs), then the disadvantaged individual is expected to have a higher talent score. In order to provide a comprehensive analysis on this theorem, we consider several different possible combinations of Talent and Environment distributions (such as Gaussian, Uniform, and other more complex distributions) and show that under which scenarios can we guarantee that our argument holds. We also discuss some counterexamples of possible distributions under which the theorem's claim will not hold. We start with an informal discussion and Uniform Talent and Environment distributions in Section 3.2.1 and then transition into Gaussian distribution in Section 3.3. Following that, we discuss a property which we call "log-concave" for a distribution curve for which our claim holds. Then, we discuss a some examples and counter-examples for the "log-concave" distributions.

Finally, from Chapter 4 onwards we provide a detailed review of seven research works

[19, 18, 26, 25, 1, 12] which we believe are the most relevant to our work. For each of the research works, we have outlined a brief summary of the model, the assumptions and the population distribution. We then discuss the main results for the research works and then analyse that whether our theorem's result would be consistent in the respective model given its assumptions. We culminate with a discussion about the worldview comparison of each of the works compared to our model's worldview. As we consider that all individual groups are born equal and the talent distribution is the same across groups, we discuss whether we find this well reflected in the models of these other research works.

1.2 Research Objectives and Questions

Apart from the literature review in machine learning fairness, the goal of this research is to answer the following research question– under which possible combinations of Talent and Environment distributions (such as Gaussians, Uniform, and other more complex distributions) and under which scenarios can we guarantee that our theorem's argument holds. Of course, the argument is that if two individuals have a similar performance score on a test (such as on a Job Interview or SATs), then the disadvantaged individual is expected to have a higher talent score.

We consider different cases of Talent and Environment distributions starting with Uniform and Gaussian distributions. We then divide the problem more generically into two parts, first considering what we refer to as graceful Talent distributions and any arbitrary Environment Distribution that we refer to as non-extreme score values. The second part considers the complement of the first i.e. with Non-graceful Talent or Extreme score values.

1.3 Contributions

With the recent growth on research in mitigating bias and promoting fairness in ML-based classifiers, the area is has become complex and hard to penetrate for newcomers to the domain. Hence, the first contribution (Chapter 2) of thesis seeks to provide an overview of the different schools of thought and approaches to mitigating (social) biases and increase fairness in the ML literature.

Our second contribution is the bias model we propose. Our motivation behind proposing a new model was that many research works appeared to reflect discriminatory worldview where they considered that the disadvantaged groups are inherently less talented. In this work, we aim to provide a model for equal distribution of talent among groups by taking the support environment of the individuals of different demographic groups as an input to our model, in order to address past discrimination. We believe that the talent is equitably distributed amongst demographic groups, however due to unequal access to resources and support for specific group's individuals, they generally tend to perform worse on standardised tests or interviews. We believe that considering the environment aspect in the model will help yielding a more just and accurate outcome by the classifier. The model discussed in Section 3.1.1 will demonstrate the consideration of environment variable to the scores we observe for an individual. We hypothesize that the performance scores that individuals get for example SAT scores for the students or an interview assessment test, is not simply a representation of their talents but also of the support environments available around them. The environment could be their education, household income or the country they are born in.

Our third contribution, (Chapter 3) which is the first theorem result shows that if two individuals have a similar performance on a screening test, the individual from the disadvantaged group is expected to be more talented than the advantage group's individual since in general the advantaged group individual has better support environment. We have considered several different possibilities of the distribution types for both Talent/Environment and illustrated both graphically and formally that for which distribution our main claim would hold.

Finally, this work considers the six most relevant research works (Chapter 4) which focus on achieving Group Fairness notions and have detailed summary of the model, assumptions and main conclusions of each of the research works. In addition, we compare the models with our work and verify whether our theorem's assumptions and conclusion are consistent with the modelling presented in the work. Finally, we analyze the different worldviews of research works, i.e. how the model views the population distribution of the world. We see that while there are a few models[1, 26], with which the worldviews closely align with our model's belief– that talent is the evenly distributed for both the groups– there are others which view the world from a discriminatory standpoint i.e. the assumption is that the disadvantaged group's talents or inherent capabilities are in itself biased.

Chapter 2

Common Literature & Background

2.1 Related Work

Our proposed model was inspired by Kannan et al.[19], which considers a two staged model of screening decision— first, the students are admitted to the college on the basis of high school grades (which are a noisy signal of the student's talent) and second the admitted students are hired by the employer based on college grades(again a noisy signal of talents). Our model on the other hand is single staged, where the performance scores are a noisy and biased estimate of the underlying talent. We introduce the bias and noise in our model by considering the environment as one of the components in determining the performance of the individuals.

In order to model disadvantage or bias in the system, several recent research works [25, 19, 18] have considered separate distributions of scores for the two groups, with the disadvantaged group having a stochastically lower distribution as compared to the advantaged. We have followed a similar approach in our model setup, although we consider that the Talent distributions of the two groups is exactly the same.

An interesting distinction was proposed by Friedler et al. [12]. Their discussion introduced

the concept of Construct Space, which is the attribute that is truly relevant for prediction task and the Observed Space, which is the distribution space of performance scores that are accessible to the decision maker. One could also compare our model with this framework, where talent and environment distributions would be a part of the Construct Space while the performance scores of the Observed space.

Several recent research works also compare fairness measures such as demographic parity (DP) and equal opportunity (EO). DP has been considered in numerous recent fairness papers [4, 27] and was proposed in Dwork et al.[11]. A recent work by Hardt et al. (2016)[16] introduced the concept of equality of opportunity. We formally define these fairness measures in the Section 2.4. While we primarily looked at papers which considered group fairness notions, one of the papers we came across during our research was Roth et.al.[9], which considered individual fairness notion i.e. that all similar individuals should be treated similarly.

2.2 Defining Sensitive Attributes

Almost all fairness policies require the knowledge of sensitive attributes in the data-set to minimize bias against the unprivileged groups. One might argue that simply removing sensitive attributes could help resolve the bias in a classifier, however many studies such as Liu et al. [25] show that unconstrained learning harms the disadvantaged. Therefore, before continuing with the study of fairness in ML, a discussion on how to decide these protected attributes is essential.

Common examples of sensitive attributes are gender, age and race. However there are not so common sensitive attributes which could encompass any feature of the data that involves or concerns people. There could be features which are strongly correlated to protected variables, and not considering such features as sensitive could make the model discriminate against the underprivileged group. While legally governments generally define the sensitive attributes such as race, gender and age [31], yet, there is still the question of variables that are not strictly sensitive,

but have a relationship with one or more sensitive variables. One of the examples of such an attribute is the address of an individual which could be used to ascertain the group membership of an individual with high accuracy [6].

A few approaches try to anonymize data by finding correlation between explicitly sensitive data and other features such as graph and network-based[32] methods for discovering proxies, which are features correlated to sensitive attributes. One of the more common approach is to use causal methods to find correlation[6].

Finding a positive correlation among sensitive and any other attribute does not guarantee that the attribute is a proxy of the sensitive attribute. Therefore, several recent works[14, 29] focus on finding causal relationship among the sensitive and non-sensitive variables. The main objective behind using causal methods is to uncover relationships in the data and find dependencies. Therefore, causal methods considered to be the most efficient methods to identifying proxies of sensitive variables[15]. However determining the causal relationship between variables is an inherently difficult problem. Existing methods rely on strong assumptions and there is no agreed upon definition of causality.

2.3 Fairness Metrics: Group v/s Individual Fairness

Fairness Metrics could either be a group based for instance ensuring equality across men and women, or it could be individual based having the idea that all similar individual should be treated similarly. Group fairness notions try to equalize the two demographic groups, such as demographic parity where the selection rates across the groups are equal. Often in group fairness to ensure equality, the members of the disadvantaged group are given an advantage (affirmative action)[19] which comes at the cost of individual fairness being impossible.

Individual fairness, as its name suggests, focuses on individuals rather than the entire groups. It was first proposed in Fairness Through Awareness by Dwork et al.[11] in 2012, which

is one of the most important foundational papers in the field. The notion of individual fairness emphasizes that all similar individuals should be treated similarly i.e. rather than focusing on group, we tend to care more about the individuals. Besides, individual fairness is more fine-grained than any group-notion fairness: it imposes restriction on the treatment for each pair of individuals.

Several studies show that Individual and Group fairness are irreconcilable[3] and cannot co-exist apart from non-degenerate group score distributions. Non-degenerate score distributions implies that the features or scores of the two groups have the same distribution, and if this is the case then we can achieve individual and group fairness together. However, non-degenerate cases are rare in real world situations and there exists several studies which have shown this tension between individual and group fairness[30, 8, 21]. In our survey, we have we have primarily focused on Group Fairness notions, while we also looked at one research work with individual fairness [9].

2.4 Group Fairness Policies and Impossibility Theorem

The most common way to assess fairness is to compare the outcome of the classifier for the two groups and if there is a discrepancy, we find ways to fix it. The crux of most fairness research lies in how to compare the output of the model's classifier and compare its result for the two groups.

There are numerous different fairness policy definitions[28] which have been proposed over time about how to compare the classifier's output for the two groups to ensure fairness. This also gives rise to the impossibility theorem which states that although most of the fairness criterion are achievable individually [10], these fairness criterion are not achievable simultaneously as shown by Kleinberg et. al[23]. This section will next discuss the fairness measures which are common across the literature we review. Consider that a positive outcome by a classifier represents something good in the society like a loan. Although this list is not exhaustive, the below 4 fairness measures are the most relevant to our work and the thesis will discuss each in detail.

In statistical learning theory, we model the data generation from a probability distribution

on the Cartesian product of domain set $\mathcal{X}$ and label set $\mathcal{Y}$ i.e. $P = \mathcal{X} \times \mathcal{Y}$. Considering this terminology, we consider the definitions in the following sections.

2.4.1 Demographic Parity

One of the earliest definitions of fairness, this metric defines fairness discussed in Dwork et al. ([11]), states that the proportion of each segment of a protected class (such as race) should receive the positive outcome with equal probabilities.

$$\mathbb{P}_{x\sim P}(h(x) = 1|x \in A) = \mathbb{P}_{x\sim P}(h(x) = 1|x \in B) \tag{2.1}$$

where $\{A, B\}$ represent the group membership and $h(x)$ represents a binary classifier's output function such that $h : \mathcal{X} \to \mathcal{Y}$. For the case when h is a randomized classifier, then all the probabilities in this sections are also over the random bits of the predictor.

2.4.2 Equal Opportunity

Proposed by Hardt et. al.(2016) [16] equal opportunity requires that the true positive rate in Group B is the same as the true positive rate in Group A i.e. Equal True Positive rates across Groups. It was first proposed in the fairness literature by Hardt et al. [16].

$$\mathbb{P}_{x,y\sim P}(h(x) = 1|y = 1, x \in A) = \mathbb{P}_{x,y\sim P}(h(x) = 1|y = 1, x \in B) \tag{2.2}$$

2.4.3 Predictive Parity

This metric introduced in Zafar et. al [34] ensures that the calibration of the model is not dependent on the sensitive attribute value. Thus, the probability of correctness of a prediction is the same for all values of the sensitive attribute. This prevents models from being biased towards making

incorrect predictions for any sensitive group. If the below two equations hold, for a classifier h, then it satisfies Predictive Parity.

$$\mathbb{P}_{x,y\sim P}(y = 1|h(x) = 1, x \in A) = \mathbb{P}_{x,y\sim P}(y = 1|h(x) = 1, x \in B) \tag{2.3}$$

$$\mathbb{P}_{x,y\sim P}(y = 1|h(x) = 0, x \in A) = \mathbb{P}_{x,y\sim P}(y = 1|h(x) = 0, x \in B) \tag{2.4}$$

2.4.4 Equalized Odds

Equalized Odds [16] is a similar notion, also introduced in Hardt et. al.[25]. In addition to requiring equal true positive rates across groups, equalized odds also requires that the false positive rates are equal across both groups. Equivalently, we can define equalized odds as $h \perp A|Y$, meaning that h is independent of the sensitive attribute, conditioned on the true label Y .

$$\mathbb{P}_{x,y\sim P}(h(x) = 1|y = 1, x \in A) = \mathbb{P}_{x,y\sim P}(h(x) = 1|y = 1, x \in B) \tag{2.5}$$

$$\mathbb{P}_{x,y\sim P}(h(x) = 1|y = 0, x \in A) = \mathbb{P}_{x,y\sim P}(h(x) = 1|y = 0, x \in B) \tag{2.6}$$

2.4.5 Disparate Impact

Disparate impact was first proposed in the fairness literature by [18] Similar to Demographic Parity, it is the ratio of positive classification rate of two groups.

$$\frac{\mathbb{P}_{x\sim P}(h(x) = 1|x \in A)}{\mathbb{P}_{x\sim P}(h(x) = 1|x \in B)} \tag{2.7}$$

2.4.6 Impossibility Theorem

The Impossibility Theorem introduced in Karthik[20] states that no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time except in non-degenerate cases i.e. when the groups have the same distributions. In non-degenerate cases, any two of the three criteria discussed in 2.4.1, 2.4.2 and 2.4.3 are mutually exclusive. Consider that G is the group membership and Y is the "true" label distribution then consider the below:

1. Demographic Parity VS Predictive Parity: If G is dependent on Y, then either Demographic Parity holds or Predictive Rate Parity but not both.
2. Demographic Parity v/s Equalized Odds: If G is dependent of Y and $\hat{Y}$ is dependent of Y, then either Demographic Parity holds or Equalized Odds but not both.
3. Equalized Odds VS Predictive Rate Parity: Assume all events in the joint distribution of $(G, \hat{Y}, Y)$ have positive probability. If G is dependent of Y, either Equalized Odds holds or Predictive Rate Parity but not both.

2.5 Fairness Interventions Approaches

The Fairness in Machine Learning Research could broadly be classified into three separate ways of applying intervention to a classifier:

2.5.1 Pre-Processing

This approach targets fixing the bias in the Data itself. One of the papers (Jiang et al[17]) we looked at, followed the pre-processing approach where they claimed that the disadvantaged group's members are often underrepresented in the datasets and hence "re-weighted" the sample from that group.

Pre-processing is motivated by the assumption that it is often the data itself which is biased, and the distributions of specific sensitive or protected variables are biased or discriminatory. Therefore, pre-processing techniques generally fix this bias within data with respect to the sensitive attributes. After that, the classifier is trained on this "repaired" data set. Pre-processing is argued to be the most flexible part of the data science pipeline, as it makes no assumptions with respect to the choice of subsequently applied modeling technique.

An example of pre-processing technique is data reweighing where we change the training data distribution to correct for the bias process and then train on the new distribution. For instance, suppose that the individuals from the disadvantaged group are under-represented in the training data so we can intervene by up-weighting the observed fraction of positives in the training data from Group B to match the fraction of positives from the advantaged group's training data.

2.5.2 In-Processing/Constrained Optimization

The constrained optimization approach to apply fairness intervention is the technique used by us in our findings. In-processing considers that modeling techniques often become biased by dominant features, other distributional effects, or try to find a balance between multiple model objectives, for example having a model which is both accurate and fair. In-processing approaches tackle this by often incorporating one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness.

2.5.3 Post-Processing

Post-processing approaches recognize that the actual output of an ML model may be unfair to one or more protected variables and/or subgroup(s) within the protected variable. Thus, post-processing approaches tend to apply transformations to model output to improve prediction fairness. Post-processing is one of the most flexible approaches as it only needs access to the predictions and

sensitive attribute information, without requiring access to the actual algorithms and ML models. This makes them applicable for black-box scenarios where not the entire ML pipeline is exposed.

Chapter 3

Theorem 1

3.1 Introduction

Our model considers that every demographic group is born with equal inherent capabilities, regardless of race, gender or the color of their skin. No demographic group is intrinsically different in the population distribution of their inherent Talent. However, due to societal imbalance of opportunities and lesser access to education, the disadvantaged group tends to perform worse on screening decisions or standardized tests such as Job Interviews or SATs. This eventually distorts the distribution of their performance scores on these tests, which becomes biased against the disadvantaged group.

To anticipate the difference in support available to different groups, we propose a model that achieves fair screening decisions with the consideration of the support environment available to an individual. Presuming that the performance score distribution of the disadvantaged group is lower than their actual talent scores (due to fewer opportunities available to them), we provide theoretical arguments of the cases when it is beneficial for the employer to hire individuals of the disadvantaged group.

This theorem shows that if two individuals have a similar performance score on a test (such as on a Job Interview or SATs), then the disadvantaged individual is expected to have a higher talent score. In order to provide a comprehensive analysis on this theorem, we consider several different possible combinations of Talent and Environment distributions (such as Gaussians, Uniform, and other more complex distributions) and prove under which scenarios can we guarantee that our argument holds.

The Section 3.1.1 will outline the model we propose and the motivation behind the problem we are trying to solve. Then we start with the simplest case of Uniform Talent and Environment distributions in Section 3.2.1. Following that, we consider Gaussian distributions for both in Section 3.3. We then divide the problem more generically into two parts. The first part considers what we refer to as *graceful* Talent distributions and any Environment Distribution that refer to as *non-extreme* score values. The second part considers the complement of the first i.e. with Non-graceful Talent or Extreme score values. Finally, Section 3.6 outlines the proofs of the theorems we outline in this section.

3.1.1 Model

We continue with the model of three distribution spaces that describe the target attribute of a prediction model from Friedler et al. [1]. The *Construct Space(CS)* represents the value of the attribute that is truly relevant for the prediction task, such as Talent of a student. This value is usually not measurable, so prediction models in a supervised learning problem are instead trained with a related measurable label, whose values are sampled from the *Observed Space(OS)*. Finally, the *Decision Space(DS)* describes the output of the model. We consider two possible group membership for an individual, that is either A or B and the membership is represented by $G \in \{A, B\}$.

Construct Space: The construct space consists of two distributions:

Talent Distribution Our main goal is to determine this Talent of an individual where T

is the random variable chosen from the same talent distribution. Since we do not have direct access to this space, we want to approximate it using the *Observed Space* discussed later.

Environment Distribution: Unlike the previous models [25, 19], we also take into account the Environment component, which is a measure of how conducive things are around an individual to promote her success. Let E represent the random variable chosen from the Environment distribution. We believe that the performance of an individual depends on the environment around her and therefore considering the environment could help estimate the talent with more accuracy. To model disadvantage, we assume that the environments scores of the disadvantaged have a distribution which is not as good as the advantaged, and the subsequent sections will discuss how specifically the environment distributions are shifted for the two groups.

The random variable E chosen from the environment distribution will not always be available during the training phase and we plan to keep it in the Construct Space. If in case, we have access to the Environment for each individual, then we could also consider the environment in the Observed Space, which is defined below.

Observed Space(Training Distribution): The observed space contains the feature vectors correlated to the construct space, for example SAT score, or high school grades to measure the talent in construct space.

Score Distribution: In order to approximate the Talent T for an individual, we consider that the employer has access to the performance score X. We presume that the scores X are influenced by not only the Talent T of an individual but also the environment E. Hence we consider that the random variable X of scores distribution is the sum of the Talent Distribution T and Environment Distribution E.

$$X = T + E \tag{3.1}$$

In general, the environment scores are distributed in such a way that they harm the disadvantaged group, for example in the case of Gaussian distributions, the mean for Group A is higher than Group B. This difference in distributions will be defined in each of the scenarios in the coming sections. From now on we consider that T_A and T_B represent the random variables for Group A and Group B sampled from the same talent distribution. Similarly, E_A and E_B represent the random variables for the environment distribution. And finally, X_A and X_B represent the random variables for the score distribution.

Decision Space (DS): Finally, we consider a trained classifier function $h : (X) \mapsto \{1, 0\}$, which could be the output from a machine learning model. Given an input performance vector X, the function h gives a binary screening decision such as whether to hire a candidate or not.

3.1.2 Motivation

Considering the model and the assumption that Environment for Group A is better, we will now analyze different distributions starting from Uniform, Gaussians, to some more complex ones and discuss that for which distributions we can have our theorem's claim i.e. given two individuals with the same Performance score x, the disadvantaged individual between the two is expected to be more talented as she had to undergo more difficulties to reach the same score x. Equation 3.2 represents the formal equivalent of our main claim:

$$Exp[\, t \,|\, X{=}x \,\&\, G{=}B] > Exp[\, t \,|\, X{=}x \,\&\, G{=}A] \tag{3.2}$$

In addition, if c represents my talent threshold for which we want to hire individuals, then:

$$\forall c,\ Pr[\, t \geq c |\, X{=}x \,\&\, G{=}B] > Pr[\, t \geq c |\, X{=}x \,\&\, G{=}A] \tag{3.3}$$

We argue that an employer wants to know the group membership of an individual since she

would rather hire someone from the disadvantaged group and in-fact improve the talent expectancy. In addition, we will consider the difference between the group's expected talents i.e.

$$Exp[\,t\,|X=x\,\&\,G=B] - Exp[\,t\,|\,X{=}x\,\&\,G{=}A] \tag{3.4}$$

for Uniform (section 3.2.2), Gaussian (Section 3.3) and other 3.5.1 Distributions of talents and environments. We will show that in the best case the expected difference is positive, and in the worst case it could even fall to negative. We will see that two important issues are whether the range of the Environments E is lower than the Talents T and that the environment has a property that of log-concave distribution, which include Uniform, Gaussian, and concave distributions (Section 3.5.2).

3.2 Uniform Talent and Environment Distribution

Being the easiest, we will start with all the distributions being uniform. This will allow us to explain the key issues that will arise in general.

3.2.1 Merging Distributions $\langle T, E_A\rangle$ and $\langle T, E_B\rangle$

Our first step is to understand the probability spaces for the two groups and to merge them into one for comparison. Person A and B both receive their talent T from the same uniform distribution $\mathcal{U}\,(\,t_{min},\,t_{max}\,)$. The B person receives their environment score E_B from $\mathcal{U}(\,e_{min},\,e_{max}\,)$ while the A person receives E_A from the shifted distribution $\mathcal{U}(\,e_{min}+K,\,e_{max}+K\,)$ where $K>0$. Their performance scores are computed as the sum $X_g = T_g + E_g$. Our assumption is that these two people received the same performance score x.

Our goal is to compare their talents, i.e. $Exp[T_A|X_A=x]$ vs $Exp[T_B|X_B=x]$. We have represented the full probability space as the $\langle T, E_A\rangle$ vs $\langle T, E_B\rangle$ rectangles in Figure 3.1. Here

talent is on the y-axis and environment on the x. Each tilted green line represents the narrowed probability space when conditioned on the performance score being fixed to $X_A = X_B = x_i$, namely $\bigcup_e T_B = x - e$ & $E_B = e$. Note the equation of each line solves $X_g = T_g + E_g$ giving $T_g = x_i - E_g$. Note how the y-intercept, $T_A = x_i - (e_{min} + K)$ vs $T_B = x_i - e_{min}$, is $x_2 - x_1$ higher for x_2 than x_1 and K lower for group A than group B. Because T, and E_g are uniform, so is the distribution within each of these green lines. Because we ultimately only care about the talent values, we project these green lines onto the y-axis giving the distribution $[T|X_g = x_i]$. From their ranges, we can deduce that the expected talent for x_2 is greater by this difference $x_2 - x_1$ in performance, namely $Exp[T_g|X_g = x_2] - Exp[T_g|X_g = x_1] = x_2 - x_1$ (but this is not always possible).

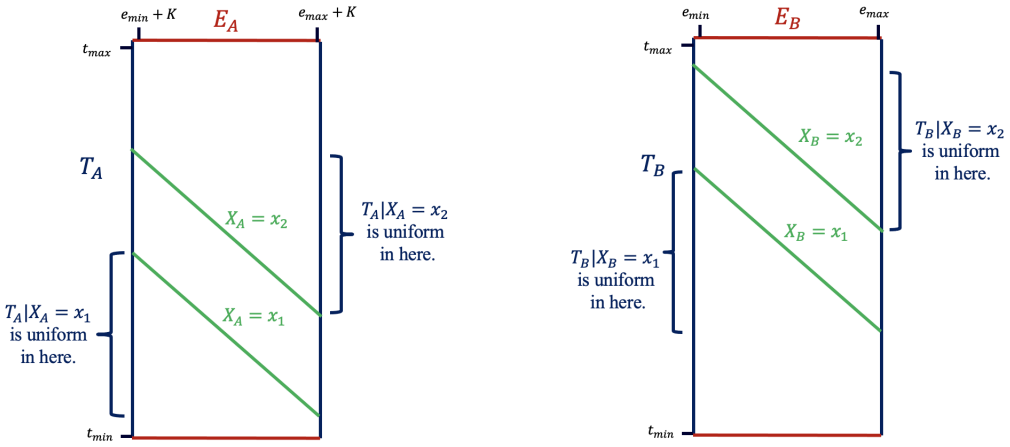

Figure 3.1: Group A and B's Distributions when Environment Range is Narrower than Talent's

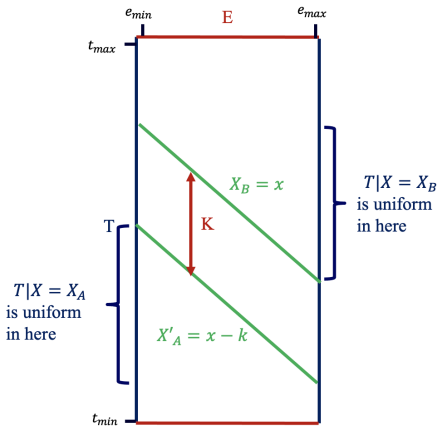

Figure 3.2: Representing X'_A and X_B on the same image

For comparison, let us now merge the two groups probability spaces $\langle T, E_A \rangle$ and $\langle T, E_B \rangle$ into one. In order to be able to plot them both on the same x-axis, independently draw an environment score E'_A and E_B from the same distribution $E = \mathcal{U}(e_{min}, e_{max})$. We advantage the A person by computing $E_A = E'_A + K$ and $X_A = T_A + E_A$. Instead lets compute $X'_A = T_A + E'_A$ and $X_A = X'_A + K$. The earlier condition $X_A = X_B = x$ is equivalent to $X_B = x$ and $X'_A = x - K$. As before, A's y-intercept, $T_A = x - (e_{min} + K)$ is K lower than B's $T_B = x - e_{min}$. Projecting these green lines onto the y-axis gives the required result that $Exp[T_B|X_B = x] - Exp[T_A|X_A = x] = K$ in the best possible scenarios, but we will see that this difference is not always possible to achieve.

In the next section, we will look at the Talent and Environment Distributions are Uniform such that the range of the Environments is larger than the Talents i.e. $(e_{max} - e_{min}) > (t_{max} - t_{min})$. Similar to the case previous case of Narrow Environment, we have a figure representing the score X'_A and X_B on the same figure 3.3. However unlike the previous figure, here we show two possible values of X which have the same talent values, which is discussed in detail in the next sections.

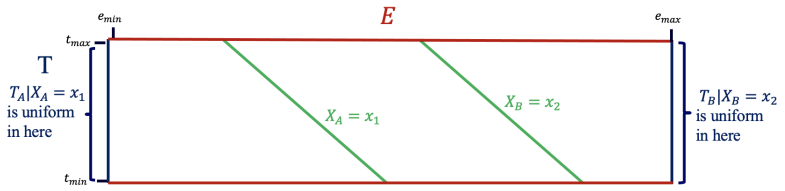

Figure 3.3: Environment's range wider than talents.

3.2.2 Extreme x values and function r

In previous section, we considered in more detail the cases where the range of the environment is narrower than that of talent. Here we will contrast this to the case where it is wider and hence the noise of the environment makes it harder to estimate the person's talent. In addition, we will also look at when the x-values are in the extremes(corners) and define a function $r : X \to \{0, 1, 2\}$ which we will need for categorizing the region types of x values.

Denote the talent's range by $[t_{min}, t_{max}]$ and the environment's by $[e_{min,g}, e_{max,g}]$. Condition on the fact that the performance score $X_g = T_g + E_g$ is fixed to some value x. Rearranging and considering the environment range gives that $T_g = x - E_g \in [x - e_{max,g}, x - e_{min,g}]$. If x is an *extreme* low value, then this low range $x - e_{max,g}$ becomes smaller than the talent's low range t_{min} and hence the bound t_{min} kicks in. Similarly, if x is an *extreme* high value, then the high range $x - e_{min,g}$ is trumped by t_{max}. We define $r(x)$ to be the number of endpoint for which this does not happen, i.e. the number of blue $y - axis$ lines that the green line intersects with. Figure 3.4 gives an example of each of the six cases.

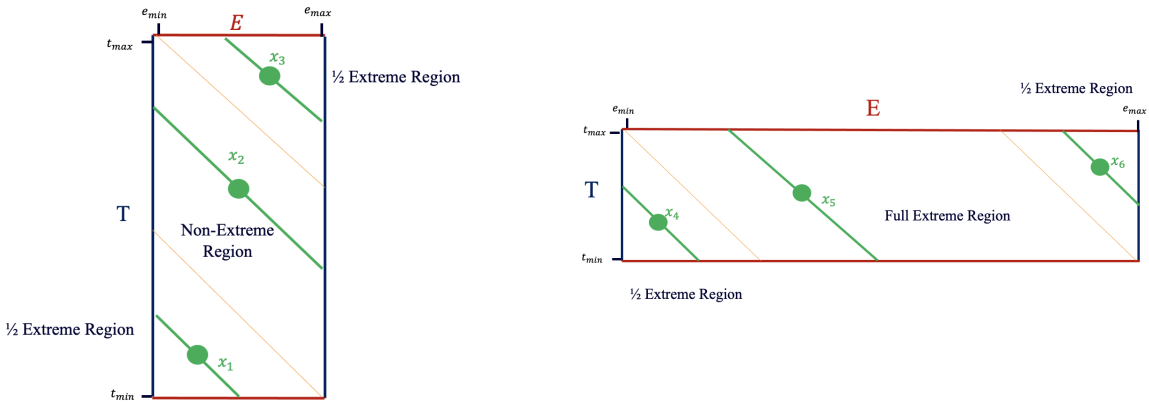

Figure 3.4: Extreme Regions: Environment Range is Narrower vs Wider than Talent's

In the non-extreme $r(x) = 2$ case, the $X_g = x_2$ conditioned talent range is $T_g \in [x - e_{max,g},\ x - e_{min,g}]$. In the bottom half-extreme $r(x) = 1$ cases, the $X_g = x_1$ or x_4 conditioned talent range is $T_g \in [t_{min}, x - e_{min,g}]$. In the top half-extreme $r(x) = 1$ cases, the $X_g = x_3$ or x_6 conditioned talent range is $T_g \in [x - e_{max,g}, t_{max}]$. Finally, in the totally-extreme $r(x) = 0$ case, the $X_g = x_5$ conditioned talent range is $T_g \in [t_{min}, t_{max}]$. In each case, the "green dot" locates the expected value of T_g within the stated range, i.e. half of the sum of its bottom and top limit. Note that $r(x)$ also denotes the number of these limits that contains an x term. Hence, if you increase x by δx, then $Exp(T_g)$ increases by $r(x) \cdot \frac{1}{2} \cdot \delta x$. Figure 3.3 shows how our conditioning is effectively that $X_B = x$ and $X'_A = x - K$. Because group A's effective x value is lowered by K from B's, $Exp(T_A)$ decreases by $r(x) \cdot \frac{1}{2} \cdot K$. This gives the result

Theorem 1. *We randomly choose the birth talent T_A and T_B of a privileged (group A) and a disadvantaged (group B) person from the same uniform distribution T.*

We randomly choose the environment score E_B of the disadvantaged person from a uniform distribution E and privileged E_A from $E+K$, i.e. the same privileged by constant K.

We set their performance to be $X_g = T_g + E_g$ the sum of their talent and performance.

Then we condition on these people having the same performance $X_A = X_B = x$.

Let $r(x) \in \{0, 1, 2\}$ measure how extreme *x is.*

Then the B people are distributionalally more or equally talented than the A people, namely

$$Exp[T_B | X_B = x] - Exp[T_A | X_A = x] = \tfrac{1}{2} r(x) K \tag{3.5}$$

This chapter has a section on non-uniform non-extreme ($r(x) = 2$) cases and another on non-uniform extreme ($r(x) = 0$) cases. Before that, however, our next task is to consider Gaussian distributions. Though Gausians have infinite ranges, the mass is concentrated within a few standard deviations. As such, we will show that its effective $r(x)$ value is between 0 and 2 depending on the variances of T, E_A, and E_B.

3.3 Gaussian Distributions

In this section, we will redo the proof with Gaussian instead of uniform distributions. By the sum of expectations, we know $Exp(X) = Exp(T) + Exp(E)$. One might be tempted to turn this around and assume $Exp(T|X = x) = x - Exp(E)$, but we already saw in the uniform case that this is not true. What is fun, however, is that we can used Bayes' rule to show that the conditional distribution $[T|X = T + E = x]$ is also Gaussian.

Theorem 2. *Let both groups get the same talent distribution $\mathcal{N}(\mu_t, \sigma_t^2)$ and T be the random*

variable drawn from this distribution. Group A gets environment distribution $\mathcal{N}(\mu_{E_A}, \sigma^2_{E_A})$ *and group B gets* $\mathcal{N}(\mu_{E_B}, \sigma^2_{E_B})$ *and let* E_A *and* E_B *represent the respective random variables. The expected environment of the privileged group is higher, namely* $\mu_{E_A} - \mu_{E_B} = K$*. Their performance, defined by* $X_g = T_g + E_g$*, is conditioned to be the same. Then considering that* T_A, T_B, E_A *and* E_B *are all independent, it follows that the B person's conditional talent is expected to be higher, namely*

$$\begin{aligned}
[T|X=x] &= \mathcal{N}\left(\frac{(x-\mu_E)\sigma_T^2 + \mu_T\sigma_E^2}{\sigma_T^2+\sigma_E^2}, \frac{\sigma_T^2\sigma_E^2}{\sigma_T^2+\sigma_E^2}\right) \\
[T_B - T_A | X_A = X_B = x] &= \mathcal{N}\left(K\frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}, \frac{2\sigma_T^2\sigma_E^2}{\sigma_T^2+\sigma_E^2}\right) \\
&\quad or \quad \mathcal{N}\left(K\frac{2\sigma_T^2}{2\sigma_T^2+\sigma_{E_A}^2+\sigma_{E_B}^2}, \frac{\sigma_T^2(\sigma_{E_A}^2+\sigma_{E_B}^2)}{2\sigma_T^2+\sigma_{E_A}^2+\sigma_{E_B}^2}\right)
\end{aligned}$$

In Section 3.2.2, we explained that if the environment range is wider than that of talent, then the noise of the environment makes it impossible to estimate the person's talent. We defined $r(x) \in \{0, 1, 2\}$ to measure the extent to which this happens and we proved that the expected difference in talents in each of these cases is $r(x)\frac{K}{2}$. Amusingly we can get the same result by setting $r = 2\frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$. Unlike the previous section, the difference in expected talent values, and hence $r(x)$, does not depend on the value of x.

It is worth noting the two extremes. When the environment range is much smaller than that of talent, i.e. when $\sigma_E^2 \ll \sigma_T^2$, then $r \to 2$ and the difference in expected talent values is maximized, namely $Exp(T_B - T_A | X_B = X_A) \to K$. In contrast, when the environment range is much larger than that of talent, i.e. when $\sigma_E^2 \gg \sigma_T^2$, then $r \to 0$ and the difference in expected talent values is minimized, namely $Exp(T_B - T_A | X_B = X_A) \to 0$. It is also interesting that if the variance $\sigma_E^2 \to 0$, then E becomes a constant along with X and hence $T = X - E$ also becomes a constant. This is why $\frac{\sigma_T^2\sigma_E^2}{\sigma_T^2+\sigma_E^2} \to 0$

The proof is in the Appendix in Section 3.6.

3.4 Graceful Talent and Narrow Environment$(r(x) = 2)$

Our goal is to define a function $F_x(t_A) = t_B$ so that $Pr(T_B \geq t_B | X_B = x)$ and $Pr(T_A \geq t_A | X_A = x)$ are close.

Story: To motivate this, consider the following story. Your job is to choose who to accept for some job/university. Being a mediumly desired job, everyone who applies happens to have performance level exactly x. Your goal of course is to accept people whose talent is as high as possible. This paper explains why you should favor people from the disadvantaged B group over those from the privileged A group. The first step show the cases where the following holds

$$Exp(T_B | X_B = x) - Exp(T_A | X_A = x) = Exp(E_A) - Exp(E_B).$$

Line Them Up by Conditional Talent: But we can say more as follows. Choose N people from group A and N from B randomly conditioned on their performances being x. Sort each group by talent into two parallel lines. For each percentile $p \in [0, 1]$, get the $p \cdot N^{th}$ person in each line to shake hands. Let t_A and t_B denote their respective talent. This can be expressed as

$$Pr(T_B \geq t_B | X_B = x) = Pr(T_A \geq t_A | X_A = x)$$

This might be useful if you suspect that those people whose talent is higher than percentile p within the privileged group A and higher than the same percentile p within the disadvantaged group B will likely accept a better offer somewhere else. Or maybe p is the risk level you are willing to take. Either way our goal is to compare these two talent levels by defining the function $t_B = F_x(t_A)$ mapping between them and by proving that $t_B > t_A$. Ideally, we even obtain $(t_B - t_A) \approx (Exp(E_A) - Exp(E_B))$.

Shift Line: In some worst cases which we will discuss in the coming sections, to get the desired result i.e. $(t_B - t_A) \approx (Exp(E_A) - Exp(E_B))$, the $p \cdot N^{th}$ B person not shake the hand of the person directly across from them but will need to shift over to less talented A people, and

shake hands with the $(p+p_\Delta)\cdot N^{th}$ A person. Here is p_Δ is a parameter that is zero when the talent distribution T is uniform and is bounded when it is *graceful*.

Extreme Values ($r(x) = 2$): We consider that the talent and environment distributions are confined to a bounded interval similar to the Uniform distribution case.

In order to prove that when they have the same performance value $X_A = X_B = x$, a disadvantaged person is expected be more talented, we must estimate the talent $t \in T$ of a person from her performance score $x \in X$. The noise making this estimating hard is the person's environment $e_g \in E_g$. In an *extreme* case, the range within which these environment values lie is wider than that for the talent. In this case, this noise overwhelms our signal and all the information about the talent is lost. In this section, we give quite a comprehensive version of the theorem under the condition that the performance measure x is extreme, i.e. $r(x) = 2$. More formally, this means that the talent range $[X-e^{max}, X-e^{min}]$ imposed by the environment is a subset of the range $[t^{min}, t^{max}]$ imposed by the talent.

Def^n $\langle s_1, s_2\rangle$-graceful: The section requires that the talent distribution T is what we call $\langle s_1, s_2\rangle$-*graceful*. We know that something similar is needed because Figure 3.6.a gives an example of a non-graceful talent distribution for which the section's result does not hold because the probability weights on the edges of the talent ranges is very small (infinitesimally small). Let the range of talent values be denoted by $[t_{min}, t_{max}]$. Because the area under its density function $P_T(t)$ is one, the average value of $P_T(t)$ is $\frac{1}{t_{max}-t_{min}}$. Our graceful condition requires that its minimum value at least an s_1 factor of this, namely $Min_t P_T(t) \geq \frac{1}{s_1(t_{max}-t_{min})}$. If its density function $P_T(t_g)$ was a line from zero to a maximum probability, then this maximum value would be $\frac{2}{t_{max}-t_{min}}$ and the slope would be $\frac{2}{(t_{max}-t_{min})^2}$. Our graceful condition requires that this slope is never more than $\frac{s_2}{(t_{max}-t_{min})^2}$. Define $p_\Delta = s_1 s_2 \frac{t_B-t_A}{t_{max}-t_{min}}$. If the talent distribution T is uniform, then $s_2 = 0$. Though we want the group's talent difference $t_B - t_A$ to be as large as possible, we are assuming that it is small relative to the size of the range $[t_{min}, t_{max}]$.

Def^n Privileged Environment: We say that group A is privileged over group B if their envi-

ronment distributions are such that when ever $Pr(E_B \geq e_B) = Pr(E_A \geq e_A)$, we have that $e_B \ll e_A$. Note that this does not say that the highest advantaged B person is less advantaged than the lowest advantaged A person. Again, we sort each group but this time by their environment value. For each percentile $p \in [0, 1]$, get the $p \cdot N^{th}$ person in each line to shake hands. Let e_A and e_B denote their respective environments. Because group A is *privileged* over B, the A person would have a significantly better environment value, giving $e_B \ll e_A$.

Theorem 3. *Here we only consider x_g that are non-extreme performance scores, i.e. $r(x_g) = 2$. If both groups have the same* uniform *talent distribution T, group A is privileged over group B with respect to their environment distributions E_A and E_B, and the measure of performance is the sum $X_g = T_g + E_g$ of the talent and environment, then*

$$Exp(T_B|X_B = x) - Exp(T_A|X_A = x) = Exp(E_A) - Exp(E_B).$$

$$Pr(T_B \geq t_B|X_B = x_g) = Pr(T_A \geq t_A|X_A = x_g) \implies t_B > t_A.$$

Theorem 4. *The talent random variables T_A and T_B for the two groups are drawn independently according to the same arbitrary $\langle s_1, s_2 \rangle$-*graceful *distribution (i.e. almost uniform). The environment random variables E'_A and E_B are drawn independently according to the same arbitrary distribution and then define $E_A = d \cdot E'_A + k$ for constants d and k. Define each person's performance to be $X_g = T_g + E_g$. Consider only the people whose performance is $X_g = x$. Here x needs to be non-extreme, i.e. $r(x_g) = 2$. (This will require talent's range to be wider than the environment's.) Suppose we sort each conditioned group according to their talent and shift these lines so that the A person with talent t_A is shaking hands with the B person with talent $T_B = F(T_A) = T_A + k + (d-1)e_B$. (Written to make it clear that $T_B = T_A + K$ when $d = 1$. Here $e_B = x - T_B$. The amount that these lines need to shift is $p_\Delta = Pr(T_A \geq T_A|X_A = x) - Pr(T_B \geq T_B|X_B = x)$. The result is that this required shift is*

small as long as the talent range is much bigger than the difference in their talents or the graceful parameters make the talent distribution almost uniform, i.e. $p_\Delta \leq s_1 s_2 \frac{t_B - t_A}{t_{max} - t_{min}}$.

See Figure 3.5. The proof is in the Appendix in Section 3.6.

Corollary 5. *Having a more general performance score computed by* $X_g = X(T_g, E_g) = uT_g + vE_g + x_0$ *(at least locally within the range* $t \in [t_A, t_B]$*) has no effect on the result, because one can achieve the same effect, by first scaling the uniform talent distribution and both environment distributions linearly.*

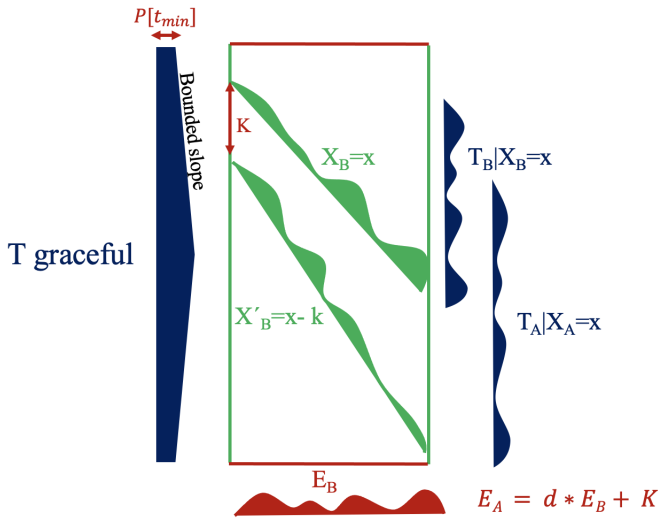

Figure 3.5: Graceful Talent and Non-Extreme X

The next section will consider such remaining cases.

3.5 Non-Graceful Talent or Wide Environment$(r(x) = 0)$

Because this section handles the harder cases, the results are not as good. Before we were able to prove that the conditioned talent of the B group is $O(K)$ bigger than that for A. This is no longer the case. Instead, we sometimes we might be able to prove B is slightly better, but there are

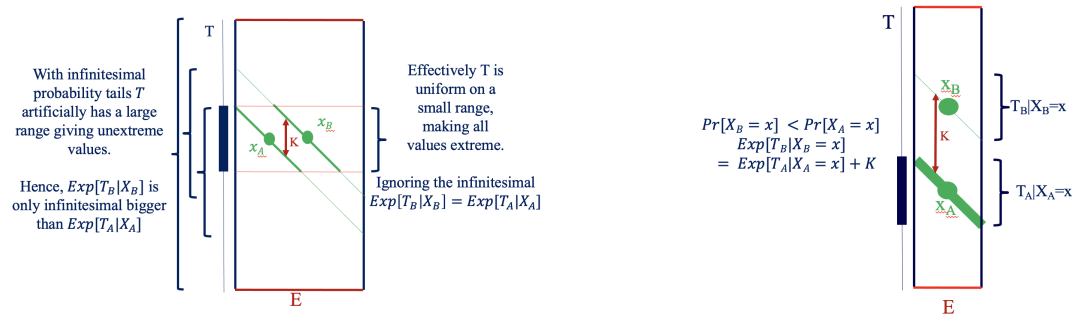

Figure 3.6: Though the values are not officially extreme because the range of the talent distribution is artificially large, the talent distribution is effectively uniform in a very narrow range, making all values extreme. In the (a) figure, this gives an example of a non-graceful T that does not have the properties stated in Theorem 4. In the (b) figure, because the B person is completely in the infinitesimal part of T, the probability of this occurring is infinitesimal. However, conditioned on this event happens, his expected T value is K bigger than A's.

abnormalities, where group A is better. What make the difference between these cases turns out to be whether the Environment distributions is what we call *log-concave* or not.

The previous section handled the cases in which the talent distribution T is uniform and the performance values are not *extreme*. This section does the complement. It allows for arbitrary (worst case) talent distributions T and/or allows extreme values. Amusingly these two concepts have much the same effect, namely that even after conditioning on the same performances $X_A = X_B = x$, the range of talents of the two groups, are effectively the same. Lets review. The talent's range was denoted by $[t_{min}, t_{max}]$ and the environment's by $[e_{min,g}, e_{max,g}]$. Condition on the fact that the performance score $X_g = T_g + E_g$ is fixed to some value x. Rearranging and considering the environment range gives that $T_g = x - E_g \in [x - e_{max,g}, x - e_{min,g}]$. When x is an *extreme* value, the talent's original range is a subset of that induced by the environment. Hence, both groups talents are restricted to the range $[t_{min}, t_{max}]$. See Figure 3.3. Allowing a worst case talent distribution can cause the same effect even when the values are not extreme. See Figure 3.6.a, If the values are not extreme, then the talent's range induced by the environment is a subset of it's original range. This means that person A's conditional talent T_A is in the range $[x - e_{max,A}, x - e_{min,A}]$ and B's $T_B \in [x - e_{max,B}, x - e_{min,B}]$. The reason B's expected conditional talent is higher is because this range is higher. However, this section is allowing for the worst cases talent distributions and as such the distribution puts effectively little weight on the high talent values t that are possible for B

and not for A and effectively little on the low high values that are possible for A and not for B. Instead, all the weight is within some sub-range $[t'_{min}, t'_{max}]$ that is possible for both A and for B. This means that effectively these values are "extreme". Having the same range, means that there are no immediate reasons that $Exp(T_A|X_A=x)$ and $Exp(T_B|X_B=x)$ are different.

Lemma 1. Log-concave *In convex analysis [2], a non-negative function $f : \mathbb{R}^n \to R_+$ is logarithmically concave (or log-concave for short) if its domain is a convex set, and if it satisfies the inequality:*

$$f(\theta x + (1-\theta)y) \geq f(x)^\theta f(y)^{1-\theta}$$

for all $x, y \in$ dom f and $0 < \theta < 1$. If f is strictly positive, this is equivalent to saying that the logarithm of the function, $log.f$, is concave; that is,

$$\log f(\theta x + (1-\theta)y) \geq \theta \log f(x) + (1-\theta) \log f(y)$$

for all $x, y \in$ dom f and $0 < \theta < 1$.

Theorem 6. *Let talent distribution to be arbitrary within the groups shared sub-range $[t_{min}, t_{max}]$. Let group B environment distribution E_B be an arbitrary* log-concave *distribution. Let A's be $E_A = E_B + K$. Let the performance scores be calculated with $X_g = T_g + E_g$. It follows that*

$$\forall \widehat{t}\ Pr(T_B \geq \widehat{t}|X_B=x) \geq Pr(T_A \geq \widehat{t}|X_A=x)$$

$$Exp(T_B|X_B=x) > Exp(T_A|X_A=x)$$

Definition: *An environment distribution E_B is* log-concave *if the following function is non-*

increasing.

$$H(e) \quad = \quad \frac{P_{E_B}(e)}{P_{E_A}(e)} = \frac{Pr(E_B \in [e, e+\delta e])}{Pr(E_A \in [e, e+\delta e])} = \frac{Pr(E_B \in [e, e+\delta e])}{Pr(E_B + K \in [e, e+\delta e])} = \frac{H_B(e)}{H_A(e)}$$

We will see that this includes Uniform, and Gaussians, and any thing this is concave within the range and zero outside the range. It fails for distribution that have two levels or are convex.

3.5.1 Intuition Behind Log-Concave Distributions and Examples

The function $E(e) = e^2 + 1$ *is clearly convex everywhere and as such does not fit into the above definition. As suspected Wolfram Alpha plots* $H(e) = (e^2+1)/((e+1)^2+1)$ *to be decreasing for* $e \in [-1, 1]$*. However, it plots to be increasing for* $e \in [1, \infty]$ *and hence is also log-concave.*

We will now look into further detail the intuition behind the log-concave distributions with some examples. Note $H_B(e)$ is simply the density function for the environment distribution E_B and $H_A(e) = H_B(e-k)$ is that shifted back by k. In order to understand the ratio $\frac{H_B}{H_A}$, consider an example of a concave function in figure 3.7 where the pair of vertical lines represent H_A and H_B respectively. Then a curve is considered log-concave if the ratio $\frac{H_B}{H_A}$ is falling as we proceed from left to right. For example, if we have a pair of lines in the beginning of the curve as demonstrated in the figure, then the ratio is only falling as we move ahead.

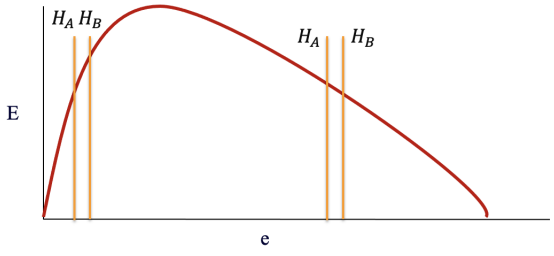

Figure 3.7: An example of Log curve

3.5.2 Examples of Log-Concave Distributions

Example 1: Let's consider a curve which is exponentially increasing as in figure 3.8. Let's consider that the $E(e) = c^e$, where $c > 1$ is a constant and e is our environment variable. Then $H_B = c^{(e-e_B)}$ and $H_A = c^{(e-e_A)}$. Hence the ratio $\frac{H_B}{H_A} = c^{(e_A-e_B)}$, which is a constant. Since this is non-decreasing, this is a log-concave function.

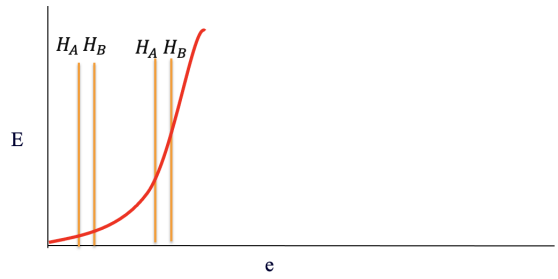

Figure 3.8: Exponentially Increasing

Example 2: Similar could be argued with an exponentially decreasing distribution, where $H_B = c^{(e-e_B)}$, $H_A = c^{(e-e_A)}$ and $0 < c < 1$ is a constant. Hence the ratio $\frac{H_B}{H_A} = c^{(e_A-e_B)}$, is also a constant. Therefore this is a log-concave function.

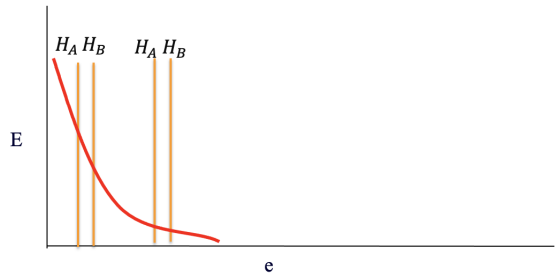

Figure 3.9: Exponentially Decreasing

Example 3: Finally, let's consider the combination of three distributions which we have seen so far. First is the concave curve in figure 3.7 for which we have shown that the lemma **??** would hold. Second, consider the exponentially increasing and finally the exponentially decreasing. The combination of the three figures is demonstrated in the figure 3.10.

We have shown that for each of the individual distributions the lemma would hold, hence

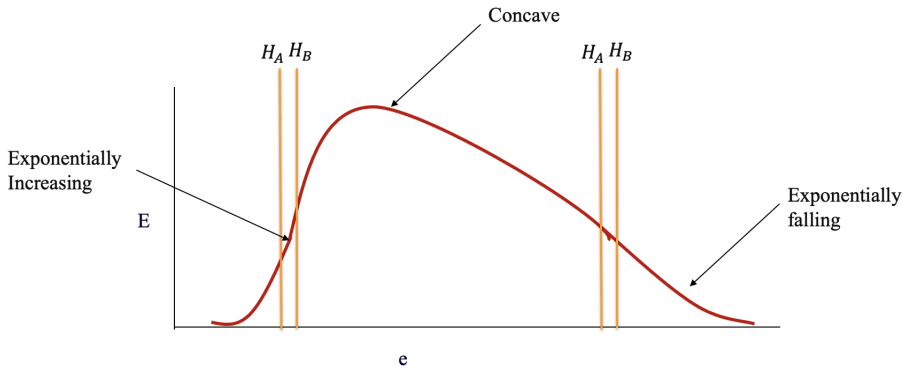

Figure 3.10: Combination Curve

lets focus on the transition between the distributions in the figure. There are two pairs of H_A and H_B highlighted on the figure such that in the first pair, H_A is in the exponentially increasing and H_B is in the concave curve, while in the second pair, H_A is in the concave curve and H_B is in the exponentially decreasing curve. For both of these pairs, the ratio $h = H_B/H_A$ is falling, since for the first pair, H_B will have a lower slope than H_A because the concave function is not increase at a faster rate than H_A. Same is true for the second pair, where H_B will decrease at a faster rate than the concave function.

Example 4: We first consider a linearly increasing Environment distribution while keeping the Talent Distributions uniform (figure 3.11) the thickness of the green lines (scores) represents the probability distribution density. We consider a linearly increasing Environment distribution as a log-concave function as the fraction $h = H_B/H_A$ will be constant as described in Section 3.5.1. Hence the claim follows.

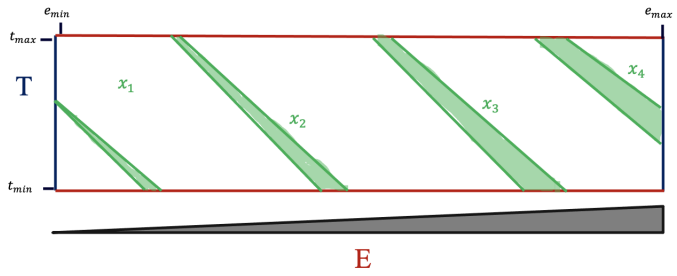

Figure 3.11: Linearly Increasing Environment Distribution

Example 5: Next, we consider a concave function, which increases linearly and then decreases. This is demonstrated in figure 3.12. Our hypothesis also holds in this case as we the fraction $h = H_B/H_A$ will be constant when x_a and x_b are both in the first and second half, and the fraction decreases otherwise.

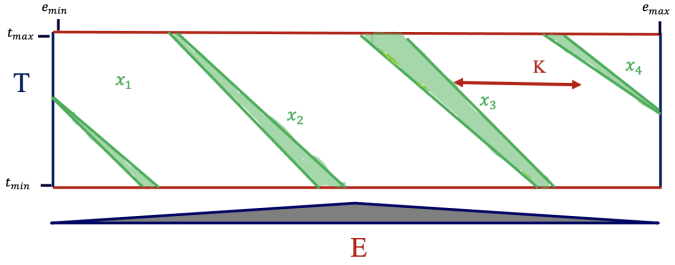

Figure 3.12: Simple Concave Environment Distribution

3.5.3 Non-Log-Concave Distributions

Example 1: Concave with Uniform Edges Let's consider an example of a function which is not log-concave. The figure 3.13 has a concave curve which ends in Uniform edges. Here, we analyse the fraction $\frac{\hat{E}_B(t)}{\hat{E}_A(t)} = \frac{\hat{E}_B(t)}{\hat{E}_B(t+k)}$ and deduce that the lemma doesn't hold. This is because, for the two pairs highlighted in the figure, the fraction will decrease while the lemma 2 holds, only when this fraction increases. In the first pair, $\hat{E}_B$ is still in uniform while $\hat{E}_A$ begins to rise. Similarly, in the second pair E_A enters the uniform while $\hat{E}_B$ begins to fall. This makes the ratio decrease for the two pairs demonstrated in the figure and hence the lemma fails to hold.

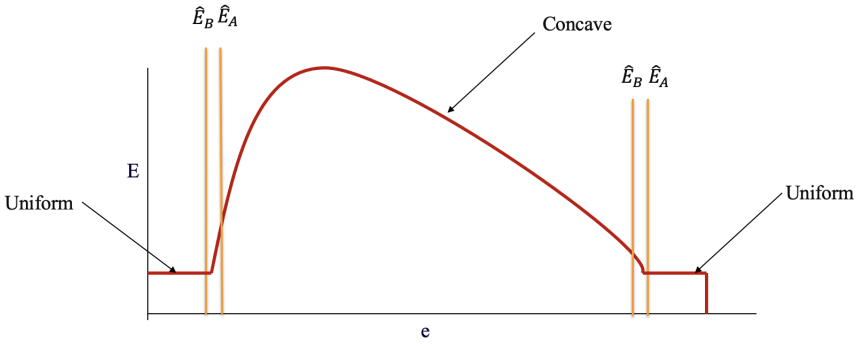

Figure 3.13: Concave Ending in Uniform Functions

Example 2: Step functions We now consider an environment distributions for which our theorem's claim fails, i.e. "Single Step" functions or "Bump" distribution function where the Environment distribution is higher in the beginning and then falls steeply. Given such a distribution and Uniform Talent distribution in figure 3.14 and 3.15, Our hypothesis fails in this case as if we look at the values x_A and x_B, then $Exp[T \mid X = x_B] < Exp[T \mid X = x_A]$. In figure 3.14, Group A's individual has a higher talent expectation as it has more probability mass in the higher talent values due to the single step. Similarly, in figure 3.15, Group B's individual has a lower talent expectation as it has more probability mass in the lower talent values due to the single step in the end of the Environment distribution.

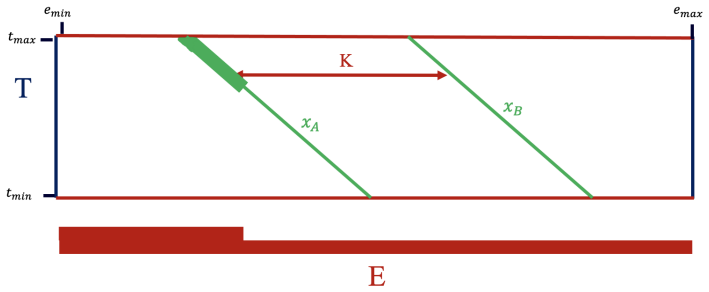

Figure 3.14: Single Step Fall Environment and Uniform Talent

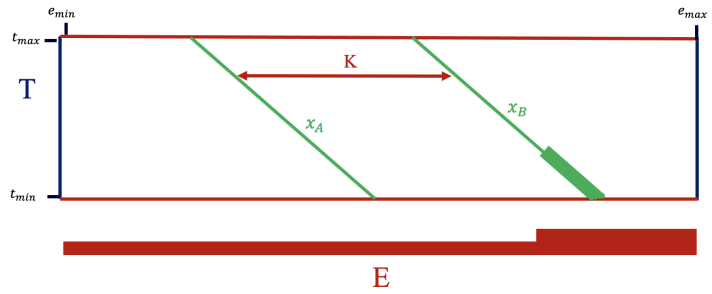

Figure 3.15: Single Step Rise Environment and Uniform Talent

Comparison of Single Step to Exponent Function In this part we compare the single step rising function with the exponentially rising function in figure 3.8. We later demonstrate in Section 3.5.1 that our claim holds for a function which is exponentially rising, while it doesn't for step function. We later delve deeper into the discussion about why the expectation claim holds for exponential but not step.

Example 3: Convex functions: Suppose that Environment distribution is some convex function as in figure 3.16 where the environment distribution is large initially, and then it decreases towards the middle becoming almost uniform and then increases again. In the real world, one could compare this to a distribution when the gap between the rich and poor is very high, such that the population is polarized and either they are mostly rich or mostly poor, while the middle class individuals lower in number.

Then in figure 3.16, if we consider the first pair of the scores, x_A will have a higher probability distribution towards the better talent values as compared to x_B. This is because the slope of curve is large initially and then falls, making the probability of large talents for x_A higher than x_B. This violates our main claim. Similar is the case with the middle two values of x_A and x_B where while x_B is in kind of uniform part of the graph, x_a has still a better distribution of environment for higher talent values. A similar argument could be derived for the last pair of scores. Hence for this distribution, excluding the extreme x values, one could argue that the claim is false for all values of x.

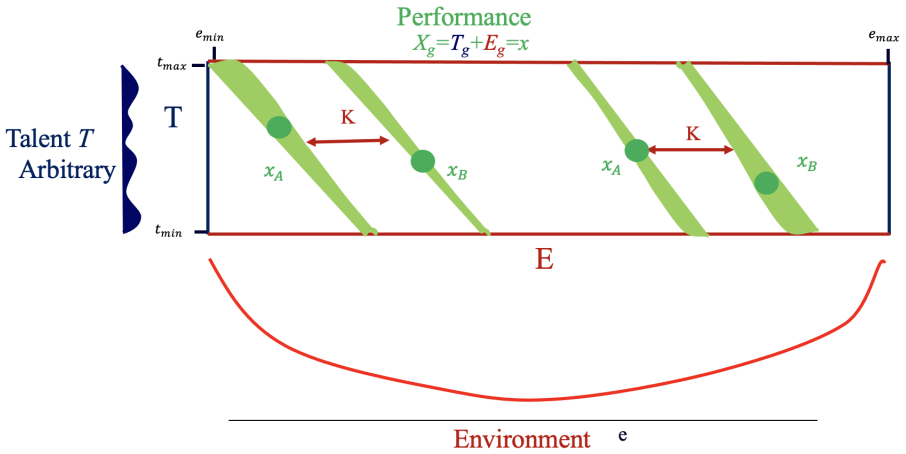

Figure 3.16: A Environment Distribution

3.6 Appendix

We give the longer math proofs in this section.

3.6.1 Proof for Gaussian Distributions

In this section, we will formally discuss the case when the Talent and Environments have Gaussian Distributions such that $E_A \in \mathcal{N}(\mu_{E_A}, \sigma^2_{E_A})$, $E_B \in \mathcal{N}(\mu_{E_B}, \sigma^2_{E_B})$, $T_A \in \mathcal{N}(\mu_{T_A}, \sigma^2_T)$, $T_B \in \mathcal{N}(\mu_{T_B}, \sigma^2_T)$ and $\mu_{T_A} = \mu_{T_B}$ (i.e. $T_A = T_B$, equality in distribution only). Then assuming that the Score Distribution X_A and X_B is the linear sum of Talent and Environment i.e. $X_A = T_A + E_A$ and $X_B = T_B + E_B$, then

$$\mathbb{E}(T_B - T_A | X_B = X_A) = \frac{K}{(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})}$$

where K is $\mathbb{E}(E_A - E_B)$.

Proof: Let K be a constant which is defined as the difference between the expected value of E_B and E_A, we assume that $K > 0$ as $K = \mathbb{E}(E_A) - \mathbb{E}(E_B) > 0$

We are interested in finding the following probability, and with Bayes rule:

$$Pr(T_B - T_A \in [t, t+dt] \,\Big|\, |X_B - X_A| < dx) \tag{3.6}$$

$$= \frac{Pr(|X_B - X_A| < dx \,\Big|\, T_B - T_A \in [t, t+dt]) \times Pr(T_B - T_A \in [t, t+dt])}{Pr(|X_B - X_A| < dx)}$$

To determine the individual probabilities on the RHS of 3.6, lets first consider $Pr(T_B - T_A \in [t, t+dt])$. The distribution of $T_B - T_A = \mathcal{N}\big(0, 2\sigma^2_T\big)$. Therefore:

$$Pr(T_B - T_A \in [t, t+dt]) = \frac{1}{2\sqrt{\pi}\sigma_T} e^{-\frac{t^2}{4\sigma^2_T}} dt \tag{3.7}$$

Similarly to find the probability $Pr(X_B - X_A < dx)$, we compute the distribution of $X_B - X_A = (T_B - T_A) + (E_B - E_A) = \mathcal{N}\big(-K, 2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B}\big)$.

Using the standard probability density function for Normal Distributions, we can infer:

$$Pr(|X_B - X_A| < dx) = \frac{1}{\sqrt{2\pi}\sqrt{2\sigma_T^2 + \sigma_{E_A}^2 + \sigma_{E_B}^2}} e^{-\frac{K^2}{2(2\sigma_T^2 + \sigma_{E_A}^2 + \sigma_{E_B}^2)}} dx \tag{3.8}$$

Finally, to find the third probability $Pr(X_B - X_A < dx \mid T_B - T_A \in [t, t+dt])$, has the distribution of $[|X_B - X_A| < dx \mid T_B - T_A \in [t, t+dt]]$.

$$Exp[X_B - X_A] = Exp[(T_B - T_A)] + Exp[(E_B - E_A)] = t - K$$

To calculate variance of $X_B - X_A \mid Y = y$, we need to take the sum variances of individual environment distributions. Note we do not consider the variance of talent distribution since we condition on talent difference being equal to t.

$$\sigma^2_{X_B - X_A | T_B - T_A = t.dt} = \sigma_{E_A}^2 + \sigma_{E_B}^2$$

Finally, the distribution of $(X_B - X_A \mid T_B - T_A \in [t, t+dt])$ is Normal $\mathcal{N}\big(t - K, \sigma_{E_A}^2 + \sigma_{E_B}^2\big)$.

Using the standard probability density function for Normal Distributions, we can infer that:

$$Pr(|X_B - X_A| < dx \,|\, T_B - T_A \in [t, t+dt]) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{E_A}^2 + \sigma_{E_B}^2}} e^{-\frac{(t-K)^2}{2(\sigma_{E_A}^2 + \sigma_{E_B}^2)}} .dx \tag{3.9}$$

Substituting Equation 3.7, 3.8, 3.9 into 3.6, we conclude:

$$Pr(T_B - T_A \in [t, t+dt] \,\big|\, |X_B - X_A| < dx) = \frac{1}{2\sqrt{\pi}\sigma_A} e^{-\frac{1}{2\sigma_A^2}\left(t + \frac{-2K}{(\sigma_{E_A}^2 + \sigma_{E_B}^2)/\sigma_A^2}\right)^2} dt \tag{3.10}$$

where $\sigma_A^2 = \frac{2\sigma_T^2(\sigma_{E_A}^2 + \sigma_{E_B}^2)}{2\sigma_T^2 + \sigma_{E_A}^2 + \sigma_{E_B}^2}$.

Equation 3.10 could now be compared to the Probability Density Function of a Normal distribution with mean and variance:

$$\mu = \frac{K}{(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})} \text{ and } \sigma^2 = \frac{2\sigma^2_T(\sigma^2_{E_A}+\sigma^2_{E_B})}{2\sigma^2_T+\sigma^2_{E_A}+\sigma^2_{E_B}}$$

Therefore:

$$(T_B - T_A \,\Big|\, |X_B - X_A| < dx) = \mathcal{N}\Big(\frac{K}{(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})}, \frac{2\sigma^2_T(\sigma^2_{E_A}+\sigma^2_{E_B})}{2\sigma^2_T+\sigma^2_{E_A}+\sigma^2_{E_B}}\Big) \tag{3.11}$$

Since K is positive, therefore the expected value(mean) of the above distribution is positive.

3.6.2 Proof for Graceful Talent and Non-Extreme X $(r(x) = 2)$

The goal of this section is to define a function $F_x(t_A) = t_B$ so that $Pr(T_B \geq t_B|X_B = x)$ and $Pr(T_A \geq t_A|X_A = x)$ are close. To do conditional probabilities, we need to be able to multiply and divide non-zero probabilities. However, if T is a continuous random variable, then $Pr(T = t) = 0$ for any specific value of t. The standard way of dealing with this is to define the density function $P_T(t)$ so that $Pr(T \in [t, t+\delta t]) = P_T(t)\delta t$.

Lemma 2. *Let $P_T(t)$ denote the density function of the talent distribution T and $P_{E_g}(e_g)$ for the environment distribution E_g. Then the density function of $Pr(T_g \in [t_g, t_g + \delta t]|X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g)P_{E_g}(x - t_g)$ for some constant c_g.*

Theorem 3 *Here we only consider x_g that are non-extreme performance scores, i.e. $r(x_g) = 2$. If both groups have the same* uniform *talent distribution T, group A is privileged over group B with respect to their environment distributions E_A and E_B, and the measure of performance is the sum $X_g = T_g + E_g$ of the talent and environment, then*

$$Exp(T_B|X_B = x) - Exp(T_A|X_A = x) = Exp(E_A) - Exp(E_B).$$

$$Pr(T_B \geq t_B|X_B = x_g) = Pr(T_A \geq t_A|X_A = x_g) \iff t_B > t_A.$$

Proof of Theorem 3: Lemma 2 gives that the density function of $Pr(T_g \in [t_g, t_g+\delta t]|X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x-t_g)$, where $P_T(t)$ denotes the density function of the talent distribution T, $P_{E_g}(e_g)$ for the environment distribution E_g, and c_g is a constant. When the talent distribution is uniform, $P_T(t)$ is a constant within its range. Hence, the density function of $Pr(T_g \in [t_g, t_g+\delta t]|X_g = x)$ is simply $c'_g P_{E_g}(x-t_g)$. This means that the conditional talent distributions is just a linear transformation of the environment X_g distribution. Hence, by the linearity of expectation, $Exp(T_g|X_g=x) = x - Exp(E_g)$. The following two results follow

$$Exp(T_B|X_B=x) - Exp(T_A|X_A=x) = Exp(E_A) - Exp(E_B).$$

$$Pr(T_B \geq t_B|X_B=x_g) = Pr(T_A \geq t_A|X_A=x_g) \iff Pr(E_B \geq x-t_B) = Pr(E_A \geq x-t_A).$$

This says that if the A person received environment value $e_A = x-t_A$ and the B received $e_B = x-t_A$, then they are at the same percentiles within their respective groups. However, because group A is privileged over B, the A person would have a significantly better environment value, giving $e_B \ll e_A$ and hence $t_B \gg t_A$. ■

Theorem 4 *If $r(x_g) = 2$, the talent distribution T is $\langle s_1, s_2\rangle$-*graceful*, the privileged environment distribution is $E_A = d \cdot E_B + k$ for arbitrary distribution E_B and constants $\langle d, k\rangle$, then*

$$Pr(T_B \geq t_B|X_B=x_g) \geq Pr(T_A \geq t_A|X_A=x_g) - p_\Delta \iff F_x(t_A) = t_B = t_A + k + (d-1)e_B.$$

Here $p_\Delta \leq s_1 s_2 \frac{t_B - t_A}{t_{max} - t_{min}}$ and $t_B = t_A + k$ when $E_A = E_B + k$.

Proof of Theorem 4: Our goal is to define a function $F_x(t_A) = t_B$ so that $Pr(T_B \geq t_B|X_B=x)$ and $Pr(T_A \geq t_A|X_A=x)$ are close. The A and B person's performance scores are computed as the sum $X_g = T_g + E_g$ and these are conditioned to be the same value x. Combining these gives that $F_x(t_A) = t_B = x - e_B = t_A + e_A - e_B$. Because the environment distributions E_A and E_B are

complicated but related we want to ground our function $F_x(t_A) = t_B$ by setting e_A and e_B to be corresponding points in these distributions, namely $Pr(E_A \geq e_A) = Pr(E_B \geq e_B)$. Because we have restricted the privileged distribution to be $E_A = E_B + k$, i.e we randomly choose a value E'_A from the distribution E_B and then set $E_A = E'_A + K$. Setting $E'_A = E_B$ then gives $e_A - e_B = k$ and $F_x(t_A) = t_A + e_A - e_B = t_A + K$. More generally, we could let $E_A = d \cdot E'_A + k$. Setting $E'_A = E_B$ then gives $F_x(t_A) = t_B = t_A + e_A - e_B = t_A + (d \cdot e_B + k) - e_B = t_A + k + (d-1)e_B$.

Having defined $F_x(t_A) = t_B = t_A + k + (d-1)e_B$, our goal now is to compare $Pr(T_B \geq t_B | X_B = x)$ and $Pr(T_A \geq t_A | X_A = x)$. Lemma 2 gives that the density function of $Pr(T_g \in [t_g, t_g + \delta t] | X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x - t_g)$, where $P_T(t)$ denotes the density function of the talent distribution T, $P_{E_g}(e_g)$ for the environment distribution E_g, and c_g is a constant. Locally, this does not tell us much, however, we can integrate to get a global probability

$$Pr(T_g \leq t_g | X_g = x) = \int_{T_g \leq t_g} P_{x,B}(T_g) \delta T_g = \int_{T_g \leq t_g} c_g P_T(T_g) P_{E_g}(x - T_g) \delta T_g$$

The problem is that for an arbitrary environment distribution $P_{E_g}(x - T_g)$, even if the talent distribution $P_T(T_g)$ is linear, we have no idea how to integrate this. The method is to convert each piece of this integral from A to B person, i.e. $P_{E_A}(e_A)\delta e = d^{-1} P_{E_B}(e_B)\delta e$, $P_T(T_A) - P_T(T_B) \leq \Delta$, and $\delta t_A = d\ \delta t_b$. Then we are sorry to say the constant c_g gives us a hard time.

Lets compare the density functions of the two groups environmental distributions. Having $E_A = d \cdot E_B + K$ gives $P_{E_A}(e_A)\delta e = Pr(E_A \in [e_A, e_A + \delta e]) = Pr(d \cdot E_B + k \in [e_A, e_A + \delta e]) = Pr(E_B \in [e_B, e_B + d^{-1} \cdot \delta e]) = d^{-1} P_{E_B}(e_B)\delta e$.

We also need to bound $P_T(T_A) - P_T(T_B) \leq \Delta$. If the talent distribution T is uniform, then this is zero. If its density function $P_T(t_g)$ was a line from zero to a maximum probability, then this maximum value would be $\frac{2}{t_{max} - t_{min}}$ and the slope would be $\frac{2}{(t_{max} - t_{min})^2}$. Our graceful condition requires that this slope is never more than $\frac{s_2}{(t_{max} - t_{min})^2}$. This gives $\Delta \leq \frac{s_2 (T_B - T_A)}{(t_{max} - t_{min})^2}$.

The form $F_x(t_A) = t_B = t_A + k + (d-1)e_B$ is the best form for seeing that $T_B \geq T_A + k$. However, in order to be able to take the derivative, lets go farther, namely $t_B = t_A + k + (d-1)\left[d^{-1}(e_A - k)\right] = t_A + \frac{K}{d} + \frac{d-1}{d}e_A = t_A + \frac{K}{d} + \frac{d-1}{d}(x - t_A) = \frac{t_A + K}{d} + \frac{d-1}{d}x$. From this we get $\frac{\delta t_b}{\delta t_A} = \frac{1}{d}$.

With this we can continue to compare our group conditional probabilities.

$$\begin{aligned}
Pr(T_A \leq t_A | X_A = x) &= \int_{T_A \leq t_A} c_A P_T(T_A) P_{E_A}(x - T_A) \delta T_A \\
&\leq c_A \int_{T_B \leq t_B} [P_T(T_B) + \Delta][d^{-1} P_{E_B}(x - T_B)][d\delta T_B] \\
&\leq c_A \left[\int_{T_B \leq t_B} P_T(t_B) P_{E_B}(x - t_B) \delta T_B + \Delta \int_{T_B} P_{E_B}(x - T_B) \delta T_B \right] \\
&= c_A \left[{c_B}^{-1} \int_{T_B \leq t_B} P_{x,B}(T_B) \delta T_B + \Delta \cdot 1 \right] \\
&= c_A \left[{c_B}^{-1} Pr(T_B \leq t_B | X_B = x) + \Delta \right].
\end{aligned}$$

Suppose $t_A = t_B = \infty$, then the probabilities are one. This gives $1 = c_A \left[{c_B}^{-1} \cdot 1 + \Delta \right]$. Setting $C_A = c$ and solving gives ${c_B}^{-1} = \frac{1 - c\Delta}{c}$ from which we get

$$P_A = c_A \left[{c_B}^{-1} P_B + \Delta \right] = c \left[\frac{1 - c\Delta}{c} P_B + \Delta \right] = (1 + c\Delta) P_B - c\Delta \leq P_B - c\Delta$$

We were disappointed that the constant c_g appears in this result, We are not really sure why it does. It can be viewed in two ways. First, it appears in the density function $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x - t_g)$ to make the area under the function one. Second, $\frac{1}{c_g} = Pr(X_g = x)$. In Figure 3.6.b because the B person is completely in the infinitesimal part of T, the probability of this occurring is infinitesimal. The makes c_B infinitely large. In this case, at least, this is not a problem because conditioned on this event happens, his expected T value is K bigger than A's.

In order to deal with the c in $P_A \leq P_B - c\Delta$, lets get a rough bound on it. Let the range of talent values be denoted by $[t_{min}, t_{max}]$. Because the area under its density function is one, the average value of $P_T(t_g)$ is $\frac{1}{t_{max} - t_{min}}$. Our graceful condition requires that its minimum value at

least an s_1 factor of this. Using this to bound the entire area gives

$$\begin{aligned}
1 &= \int_{T_g} c_g P_T(T_g) P_{E_g}(x-T_g)\delta T_g \\
&\geq \int_{T_g} c_g \frac{1}{s_1(t_{max}-t_{min})} P_{E_g}(x-T_g)\delta T_g \\
&\geq c_g \frac{1}{s_1(t_{max}-t_{min})} \int_{T_g} P_{E_g}(x-T_g)\delta T_g \\
&\geq c_g \frac{1}{s_1(t_{max}-t_{min})} \times 1 \\
c_g &\leq s_1(t_{max}-t_{min})
\end{aligned}$$

From this we can conclude.

$$\begin{aligned}
P_A &\leq \leq P_B - c\Delta \\
&= c_A - s_1(t_{max}-t_{min}) \times \frac{s_2(t_B-t_A)}{(t_{max}-t_{min})^2} \\
&\leq P_B - s_1 s_2 \frac{t_B-t_A}{t_{max}-t_{min}}
\end{aligned}$$

When the talent distribution is uniform, $s_2 = 0$ giving $P_A = P_B$. ■

Lemma 2 *:* Let $P_T(t)$ denote the density function of the talent distribution T and $P_{E_g}(e_g)$ for the environment distribution E_g. Then the density function of $Pr(T_g \in [t_g, t_g+\delta t] | X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x-t_g)$ for some constant c_g.

Proof of Lemma 2: We will drop the subscript $g \in \{A, B\}$, because the statements apply to either group. Because these random variables T and E_g are independent, we can define the cross density function $P(t,e) = P_T(t) \times P_E(e)$ so that $Pr(T \in [t, t+\delta t] \ \& \ E \in [e, e+\delta e]) = \delta t \cdot P_T(t) \times \delta e \cdot P_E(t) = \delta t \delta e \cdot P(t,e)$. We could imagine raising a third dimension coming out of the page on the $\langle T, E \rangle$ rectangle in the figure, so that its height at location $\langle t, e \rangle$ is $P(t,e)$.

Because we have the restriction that x is such that $r(x) = 2$, the talent range $[X - e^{max}, X - e^{min}]$ imposed by the environment is a subset of the range $[t^{min}, t^{max}]$ imposed by the talent. This

means that for all values of t that we care about $P_T(t)$ is still the density function for talent.

Fix a performance value x of which we will require of all group g people that we are considering for acceptance. The performance of a person with talent T and environment E is given by $X = T+E$. Just to check the accuracy of our figure, if $E_A = d \cdot E_B + K$, then the line to which we restrict the $\langle T, E\rangle$ rectangle is $x = T_A + d \cdot E_B + K$ or $T_A = x - d \cdot E_B - K$. Note this lowers the group B line by k and makes its slope $-d$ instead of -1.

Our goal is prove that the probability density function $P_x(t) = Pr(T \in [t, t+\delta t] | X = x)/\delta t$ of the distribution on talents t that arise under this condition is simply $P(t, x-t)$. If X were computed by some more complex function, this would not be the case. Using the standard formula $Pr(T \in [t, t+\delta t]$ & $X = x)/Pr(X = x)$ is awkward because the later is zero. Conditioning on our probability space amounts to narrowing our $\langle T, E\rangle$ rectangle of possibilities to the 1-dimensional line defined by $\{\langle T, E\rangle \,|x = T + E\}$. Lets us define the infinitesimal rectangle of possibilities $S_t = \{T \in [t, t+\delta t]\} \times \{E \in [x-t-\delta t, x-t]\}$. Within this, X is sufficiently close to x, the density function $P(t, e)$ is sufficiently constant. Hence, we will approximate $Pr(T \in [t, t+\delta t]$ & $X = x)$ with $Pr(S_t)$, which is $P(t, x-t) \cdot (\delta t)^2$. Lets return to the awkward fact that $Pr(X = x)$ is zero. Let's define $S_x = \bigcup_t S_t$ to be the union of all of our rectangles within which X is sufficiently close to x. Then we will replacing $Pr(X = x)$ with $Pr(S_x)$. Lets denote this probability with $p_x \cdot \delta t$. We are now able to determine the probability $Pr(T \in [t, t+\delta t] | X = x) = Pr(S_t|S_x) = Pr(S_t)/Pr(S_x) = \left[P(t, x-t) \cdot (\delta t)^2\right] / (p_x \cdot \delta t)$. Our density function $P_x(t)$ is this divided by δt. It follows that the density function of $Pr(T_g \in [t_g, t_g+\delta t] | X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x-t_g)$. Here c_g is the multiplicative constant needed to make the area under the density functions one. ■

3.6.3 Proof for Non-Uniform Talent or Extreme X Values ($r(x) = 0$)

Theorem 6 *Let* talent distribution to be arbitrary within the groups shared sub-range $[t_{min}, t_{max}]$. Let group B environment distribution E_B be an arbitrary *log-concave* distribution. Let A's be

$E_A = E_B + K$. Let the performance scores be calculated with $X_g = T_g + E_g$. It follows that

$$\forall \widehat{t}\ Pr(T_B \geq \widehat{t} | X_B = x) \geq Pr(T_A \geq \widehat{t} | X_A = x)$$

$$Exp(T_B | X_B = x) > Exp(T_A | X_A = x)$$

Proof of Theorem 6: Let $P_T(t)$ denote the density function of the talent distribution T and $P_{E_g}(e)$ for the environment distribution E_g. Lets compare $P_{E_A}(e)$ and $P_{E_B}(e)$. Having $E_A = E_B + K$ gives $P_{E_A}(e)\delta e = Pr(E_A \in [e, e+\delta e]) = Pr(E_B + k \in [e, e+\delta e]) = Pr(E \in [e-k, e+\delta e]) = P_{E_B}(e-k)\delta e$. Define $\widehat{E}_g(t)$ to be $P_{E_g}(x-t)$. Note that $\widehat{E}_A(t) = P_{E_A}(x-t) = P_{E_B}((x-t)-k) = P_{E_B}(x-(t+k)) = \widehat{E}_B(t+k)$. Define $H(t) = \widehat{E}_B(t)/\widehat{E}_A(t)$. The definition of E_B being *log-concave* is that $H(t)$ is increasing.

Lemma 2 states that the density function of $Pr(T_g \in [t, t+\delta t] | X_g = x)$ is $P_{x,g}(t) = c_g P_T(t) P_{E_g}(x-t)$. Fix some threshold $\widehat{t}$.

This gives

$$Pr(T_g \geq \widehat{t} | X_g = x) = \frac{\int_{t \in [t_{min}, \widehat{t}]} P_T(t) \widehat{E}_g(t) \delta t}{\int_{t \in [t_{min}, t_{max}]} P_T(t) \widehat{E}_g(t) \delta t}$$

To simplify notation, let's define a functional $\mathcal{F}$, which for any function $R(t)$ gives the fraction of the area under the curve that is to the right of the $t = \widehat{t}$, namely

$$\mathcal{F}(R) = \frac{\int_{t \in [t_{min}, \widehat{t}]} R(t) \delta t}{\int_{t \in [t_{min}, t_{max}]} R(t) \delta t} \tag{3.12}$$

Hence we can write

$$Pr(T_g \geq \widehat{t} | X_g = x) = \mathcal{F}(P_T \widehat{E}_g)$$

Recall that $H(t) = \widehat{E}_B(t)/\widehat{E}_A(t)$ is increasing. Let c be the constant $H(\widehat{t})$. Hence $\forall t \in [t_{min}, \widehat{t})$ we have $H(t)/c < 1$ and $\forall t \in (\widehat{t}, t_{max}]$ we have $H(t)/c > 1$. Hence, for any function $R(t)$, multiplying $R(t)$ by $H(t)/c$ decreases $R(t)$ for those t before $\widehat{t}$ and increase those after. It follows that the fraction of the area under the curve $R(t)$ that is right of the $t = \widehat{t}$ increases, i.e. $\mathcal{F}(R) < \mathcal{F}(H/c \cdot R)$.

Similarly, $\mathcal{F}(R) = \mathcal{F}(cR)$. The result follows

$$\begin{aligned} Pr(T_A \geq \widehat{t} | X_A = x) &= \mathcal{F}(P_T \widehat{E}_A) = \mathcal{F}(cP_T \widehat{E}_A) < \mathcal{F}(H \cdot P_T \widehat{E}_A) \\ &= \mathcal{F}(\widehat{E}_B / \widehat{E}_A \cdot P_T \widehat{E}_A) \\ &= \mathcal{F}(P_T \widehat{E}_B) = Pr(T_B \geq \widehat{t} | X_B = x) \end{aligned}$$

■

Lemma 3. *If* $\forall \widehat{t}\ Pr(T_B \geq \widehat{t} | X_B = x) > Pr(T_A \geq \widehat{t} | X_A = x)$, *then* $Exp(T_B | X_B = x) > Exp(T_A | X_A = x)$.

Proof of Lemma 3:

$$Exp[T_B | X_B = x] = \int_{t \geq 0} t \cdot Pr(T_B \in [t, t + dt] | X_B = x) \delta t$$

$$= \int_{t \geq 0} \int_{\hat{t} \in [0,t]} Pr(T_B \in [t, t + dt] | X_B = x) \delta \hat{t} \delta t$$

$$= \int_{\hat{t} \geq 0} \int_{t \geq \hat{t}} Pr(T_B \in [t, t + dt] | X_B = x) \delta t \delta \hat{t}$$

$$= \int_{\hat{t} \geq 0} Pr(T_B \geq \widehat{t} | X_B = x) \delta \hat{t}$$

$$> \int_{\hat{t} \geq 0} Pr(T_A \geq \widehat{t} | X_A = x) \delta \hat{t}$$

$$= Exp[T_A | X_A = x]$$

. ■

3.7 Conclusion

In this theorem's analysis, we have considered numerous Talent and Environment distributions and outlined several conditions for which hiring individuals from the disadvantaged group is beneficial for the employer from an accuracy standpoint. In addition, we also bound the difference in the expected values of talents of the advantage and disadvantaged group individuals with the same score for Gaussian and Uniform distributions.

Our modest hope with our work is to demonstrate to the employers the talent advantage they get when hiring individuals of the disadvantaged group, and thus progress towards fairness in decision making.

Chapter 4

Related Research Work Review

4.1 Introduction

During our research, we came across several research works which were similar to our model and the main idea we set out to prove i.e. in Theorem 1 (Section 3). In total, we covered 12 different research works [23, 19, 18, 26, 25, 1, 24, 9, 17, 33, 12, 13]. While we studied each of these works in detail, however in the thesis, we will cover 6 of these works, which are most relevant to us. Other works such as Roth et al.[9] considered individual fairness, which were impertinent to our work and therefore we will not discuss the remaining 6 works in detail.

For each of these research works, we will consider our main claim from Theorem 1 and verify whether it holds in the setting of each of the individual models. That is, in each of the models, we are looking for the phenomenon that given two individuals have the same performance score or feature sets, the disadvantaged is likely to be more talented.

In addition, we compare the worldview of each of the individual models with our worldview. Our prime motivation to draft a new model of considering Talent and Environment in calculating the performance score was the belief that all demographic groups are born equal, and that the

circumstances around the disadvantaged groups are not as conducive as for the advantaged group. Hence the assumption of Talent distribution being the same for all groups supports the idea that the talent among demographic groups is equitably distributed, while the Environment distribution captures the difference in support available to individuals. However, other research works in fairness in machine learning have different theoretical models, within which they model the sources of disparities between groups and solve the problem of bias against the disadvantaged.

The following list will briefly cover why we have chosen to include the following six research works in the thesis.

1. **Downstream Effects Of Affirmative Action:** The Kannan et.al.[19] paper was the initial motivation for us to work in the field of Group Fairness in Machine Learning, which describes a two-staged model where the students are first admitted to college on the basis of their scores. Those students are hired by an employer based on college grades. Given the model, this work studies which Fairness Goals (such as *Equal Opportunity* and *Irrelevance of Group Membership*) could be achieved by the college by updating its admissions rule and grading policy. There are many similarities in this work as compared to ours, the score distributions are Gaussian and they use threshold functions as their hypothesis classes. While this work proposes a two-staged model, our model is single-staged. The Section 4.2 will cover the model, assumptions and main results of this paper in detail. In addition, the section will also illustrate potential similarities with our work.

2. **The Disparate Effects of Strategic Manipulation**: In this work by Immorlica et. al.[18], the model, scores distributions and the results discussed are very similar to ours. The term "strategic manipulation" in this paper[18] refers to the students ability to trick the classifiers such as the SAT test exam by manipulating their feature vectors. The students could achieve this by taking professional test prep services which could help them trick the classifier. And to model disadvantage, this ability to deceive the classifier is not equitably distributed across the group i.e. the disadvantaged have might face a higher cost when trying to manipulate

their features. The main results of this paper show that whenever one group's costs are higher than the other's, the learner's equilibrium strategy (unconstrained learning) exhibits an inequality-reinforcing phenomenon wherein the learner erroneously creates False Positives on the advantaged group, and False negatives of the disadvantaged group. The study also take in account the subsidy intervention (Affirmative Action) and shows cases where the subsidy hurts both the groups.

3. **Simplicity Creates Inequity**: In this work by Kleinberg et. al.[23], there is a discussion on how in order to achieve interpretability, simplicity could in-fact harm the disadvantaged group and increase the bias among groups. More formally, this paper proposes a framework for producing simple prediction functions and shows two results. First, a simplified function is strictly improvable in both equity and accuracy and hence simplification doesn't help to achieve fairness or interpretability. Second, a simple function creates incentive for the employer to use group membership information which is used against the disadvantaged members. Our main inspiration behind choosing this research work in our final thesis was that the results are in contrast with our model. Not only our theorem 1's results are inverted in this paper's model, but also the worldview of population distribution discussed in the work doesn't consider equally distributed talents.

4. **From Fair Decision Making to Social Equality:** This work by Srebro et. al.[26] is novel in terms of its comparison with the long-term influence of applying fairness interventions on the underlying population also known as *dynamics*. The final results of this work show that there are conditions when the Unconstrained Policy achieves the population equality over long term while Demographic Parity causes harm. On the other hand, in more realistic scenarios Unconstrained Policies will not result in equality while DP in these cases could achieve equality. Since this work looks at Group Fairness from dynamics perspective, and shows the effect of Demographic Parity on the downstream, it is co-related with our work. Section 4.5 will cover this model and dynamics discussed in the work and will also show how the Unconstrained learning is always disadvantaged for Group B as the classifier commits

False Positives on Group A and False Negatives on Group B individuals.

5. **Delayed Impact of Fair machine learning:** This work (Liu et. al.[25]) as the name suggests considers the delayed impact or dynamics of applying fairness policies on the population distribution, i.e.: long-term improvement, stagnation, and decline in a variable of interest. This work considers a study of one-step feedback model with common fairness criteria such as demographic parity which in general do not promote improvement over time, and may even cause harm in certain cases where using an unconstrained objective would work better. Then they considers the delayed impact of three standard criteria and also determine their impact on the utility graphs. The Section 4.6 will demonstrate the model and a novel tool called *Outcome Curve*, which is helpful in comparison of delayed impact of different fairness constraints.

6. **Recovering from Biased Data:** This work by Blum et. al.[1] discusses, how to extract the Bayes Optimal Classifier. This work examines the possibility of extracting the Bayes Optimal Classifier when the Fairness Constraint is applied to the ERM. This paper has a similar model as compared to ours. There are two specific types of bias models which were discussed in this paper, under-representation and misrepresentation. On the other hand, our bias model is differnet and more general where we model disadvantage through the environment distribution. In the Section 4.7, there is a detailed discussion of the bias model used by this work as compared to our model. In addition, the section will compare the final results.

4.2 Research Review 1: Downstream Effects Of Affirmative Action

This research work by Roth et. al.[19], was the initial motivation for us to start exploring the Group Fairness policies for our research work. In addition, the model of this paper has a significant overlap with ours. While we consider a one-staged model, this paper discusses a two-staged model

of a hiring process. In the first state, high-school students are admitted to the college or university on the basis of an entrance exam which is signal about their qualifications i.e. (talent/type) with some Gaussian noise. In the second stage, those students who were admitted to college are be hired by an employer as a function of their college grades, which are an independently drawn noisy signal of their types.

The employer who hires at the end of the pipeline is a practical employer (trying to maximize his profit) and calculates a distribution on the types/talents of the students conditional on qualifications from Stage 1 and grades from Stage 2. The final results in this paper considers the conditions under which the two fairness definitions can be met or not. Following is the informal definitions of the two fairness criteria discussed in this work:

1. Equal opportunity: As proposed in Kannan et.al. [19] (Page 7) the definition of equal opportunity is different from the one we discussed earlier, as it considers a two-staged model. The probability that an individual is accepted to college and then ultimately hired by an employer may depend on an individual's type, but conditioned on their type, should not depend on their demographic group.

2. Irrelevance of Group Membership: As proposed in Kannan et.al. [19] (Page 7) IGM takes into account the rationality of employers, where it claims that the employer who is selecting employees from the college population should not make hiring decisions based on group membership.

We will compare in the conclusion Section how the fairness notions IGM and EO overlaps with our Theorem 1 and how IGM is less strict in terms of the inequality as compared to our Theorem 1's definition.

4.2.1 Model

The model considers two population distributions of students represented by $i \in \{1, 2\}$ where $i = 1$ is advantaged and $i = 2$ is disadvantaged. Students have a type drawn from Gaussian distribution with mean μ_i and variance σ_i^2, which in practice we don't have access to. Hence, the Gaussian distribution is represented as $P_i = N(\mu_i, \sigma_i^2)$ and T_i is the random variable. Hence, the Talent distributions itself are discriminatory as this paper considers that the inherent talent between the groups is different. This is in contrast to our model where we regard the Talent Distribution to be same for all the groups.

Continuing with the model, in order to approximate the type of each student, SAT Scores are used: $S_i = T_i + X$ (X being the noise, which follows a normal distribution with mean 0 and variance 1). The college has Admission Rules $A_i(s) : \mathbb{R} \to 0, 1$ which are binary threshold functions and different for each group. A student i with SAT Score s is accepted in the university with the probability $A_i(s)$, such that $A_i(s) = 1$ a student is accepted (i.e. if she is above the β threshold, $S_i \geq \beta_i$) and 0 otherwise.

The second stage to determine the Student type, the paper discusses the use of University Grade G_i: Every student admitted to the university receives a grade $G_i = T_i + Y$ (where Y follows a normal distribution with mean 0 and variance γ^2). Finally, employer makes a hiring decision: The employer knows the priors P_i from which talent is sampled, the admission rules A_1, A_2 used by the school, the grading policy γ, and observes the grades of the students. An individual is hired if the employer's expected utility for accepting a university graduate from population i with grade g is above the threshold C, which is the cost to the employer to hire the individual:

$$E[T_i | G_i = g, A_i = 1] \geq C$$

Hence, a rational employer will hire a student if the above inequality holds. As there are many employers, we consider that the range of possible C values is $C \in [C-, C+]$.

4.2.2 Fairness Definitions

There were two main fairness definitions proposed in this paper:

- Equal Opportunity (EO): This definition of equal opportunity is a little different from the one we discussed earlier, as this considers two stages of admissions. In this case we say that equal opportunity holds of conditional on the talents, the probability of a student being hired by the employer is not dependent on their group membership. i.e. if for all types t $\in$ R,

$$\int_g Pr[G_1 = g \& A_1 = 1 | T_1 = t].\mathbb{1}\{E[T_1|G_1 = g \& A_1 = 1]\} =$$
$$\int_g Pr[G_2 = g \& A_2 = 1 | T_2 = t].\mathbb{1}\{E[T_2|G_2 = g \& A_2 = 1]\}$$

- Irrelevance of Group Membership (IGM): IGM on the other hand holds, if the expected value of talent being greater than the employer threshold holds for both the groups simultaneously conditioned on if they are admitted to the college. i.e. if for all grades g $\in$ R,

$$\mathbb{E}[T_1|G_1 = g \& A_1 = 1] \geq C \iff \mathbb{E}[T_2|G_2 = g \& A_2 = 1] \geq C$$

Next, this work considers different conditions under which we can guarantee the different fairness conditions hold.

4.2.3 Main Results

The below results will discuss the specific Noise distributions and assumptions under which the two fairness goals could be met. However in the worst-case we will see that none of the fairness goals are met.

Case 1: Noiseless Exam Scores

Considers that when hiring a student in the college, the SATs and the high school grades that are used to admit a student are noiseless i.e. the scores perfectly reflect the talent of the individuals.

Of course, this assumption is not practical, however this is the best case scenario for ensuring the two fairness goals are met.

The papers shows that if the model is noiseless then above two fairness goals can be simultaneously achieved only by highly selective colleges (i.e. those with high admissions thresholds greater than $C+$) — and only if they do not report grades to employers.

Claim. Suppose $S_i = T_i$, i.e. a student's score perfectly reveals his type. Then for any hiring interval of hiring costs $[C-, C+] \in R$, the non-zero admissions rule:

$$Ai(s) = 1 \Leftrightarrow s \geq C+ \tag{4.1}$$

for both groups $i \in 1, 2$ satisfies *IGM* and *Equal Opportunity* when paired with any grading policy.

The Single Employer Threshold Case

Let's assume that the grades and SAT scores are noisy and instead of considering a range of thresholds for different employers, this section considers that there is only one employer and has a cost C. If such is the case, then IGM is achievable although we lose Equal Opportunity. More formally, for any grading scheme, and with a single threshold C, the college can separately set different admissions thresholds β_1^* and β_2^* for the two groups respectively such that the posterior expectation for a student type from each group crosses the threshold of C at a grade g^*.

Lemma: For any $C \in \mathbb{R}$, there exists thresholds β_1^* and β_2^* and a grade g^* such that

$$Exp[T1|G1 = g^*, S1 \geq \beta_1^*] = Exp[T2|G2 = g^*, S2 \geq \beta_2^*] \tag{4.2}$$

Multiple Threshold Case

Finally, in this section we consider that there are multiple employers and hence there there are many thresholds of hiring costs $C \in [C-, C+]$ and that the grades/scores are noisy. In such case,

the first claim is that IGM is impossible to achieve. The proof in the paper demonstrates this part.

4.2.4 Comparison to our Theorem 1

We now compare the main claim of our Theorem 1 with the findings in this paper. Recall that Theorem 1 states: Given two individuals with the same scores, the person who belongs to the disadvantaged group is expected to be more talented. i.e. $Exp[t|\, x \,\&\, B] = t_B > t_A = Exp[\, t|\, x \,\&\, A]$ or equivalently $\forall c$, $Pr[t \geq c|\, x \& B] > Pr[t \geq c|\, x \,\&\, A]$.

We say that the definitions of IGM and EO use a similar comparison criteria of fixing the performance scores of the individuals in order to compare i.e. in both of the definitions, this work conditions on having fixed grades and being admitted to college which is comparable to conditioning on having the same scores performance scores in our theorem's setting. However with this paper's model, it cannot be guaranteed that the our claim for Theorem 1 would hold as this work starts with the assumption that the Talent distribution for the disadvantaged group is inherently lower.

4.2.5 Worldview Comparison

The paper Downstream Effects of Affirmative Action [19] was the initial motivation for us to consider the problem of screening decisions where we want to hire a candidate given her scores. There are two limitations of this paper from the bias model's standpoint. First, the Talent Distributions T_i(refereed to as Type) is different for different groups (i represents the group membership). Second, as discussed in Section 4.2, the paper discusses two-staged model where this work assumes that even after going though the university, the type distributions do not change i.e. disadvantaged group remains at a lower talent distributions. Therefore, this research work's model can be viewed as discriminatory against the disadvantaged group. Hence, we conclude that this work's worldview is in contrast to our model.

4.3 Research Review 2: The Disparate Effects of Strategic Manipulation

4.3.1 Introduction

The term "strategic manipulation" in this paper[18] refers to the students ability to trick the classifiers such as the SAT test exam by manipulating their feature vectors. The students could achieve this by taking professional test prep services which could help them trick the classifier. And to model disadvantage, this ability to deceive the classifier is not equitably distributed across the group i.e. the disadvantaged have might face a higher cost when trying to manipulate their features. The main results of this paper show that whenever the cost of manipulation of the disadvantaged group are higher than the advantaged group, any classifier with equilibrium strategy which basically selects same threshold for the two groups will reinforce the inequality between the groups. The learner does that by erroneously creating False Positives on the advantaged group, and False negatives of the disadvantaged group. The study also take in account the subsidy intervention (Affirmative Action) and show cases where the subsidy hurts both the groups.

4.3.2 Model and Notion

To briefly state the model, consider that the candidates have innate set of features $\mathbf{x} \in X = [0,1]^d$ belonging to Group A or B, who respond by manipulating their features to cross the classifier's threshold and get selected.

The manipulation costs are defined according to group such that a candidate from group m who wishes to move from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{y}$ must pay a cost of $c_m(\mathbf{y}) - c_m(\mathbf{x})$ where $y \geq x$. To model disadvantage, the study assumes that

$$c_A(\mathbf{y}) - c_A(\mathbf{x}) \leq c_B(\mathbf{y}) - c_B(\mathbf{x}), \forall \mathbf{y} > \mathbf{x} \tag{4.3}$$

i.e. Group A members pay a lower cost than Group B and the cost functions are monotone increasing with the features. This cost could be thought of as the hardships individuals face while trying afford the tuition fees for SAT test prep services. As its more difficult for the disadvantage group to earn the sum required to pay the tuition than the advantaged, the cost for group B is higher.

Consider that $\mathcal{D}_A$ and $\mathcal{D}_B$ are the distributions over unmanipulated features and to be subject to different true labeling functions h_A and h_B defined as

$$h_A(x) = \begin{cases} 1 & \forall x \text{ such that } \sum_1^d w_{A,i} x_i \geq \tau_A \\ 0 & \forall x \text{ such that } \sum_1^d w_{A,i} x_i < \tau_A \end{cases} \tag{4.4}$$

$$h_B(x) = \begin{cases} 1 & \forall x \text{ such that } \sum_1^d w_{B,i} x_i \geq \tau_B \\ 0 & \forall x \text{ such that } \sum_1^d w_{B,i} x_i < \tau_B \end{cases} \tag{4.5}$$

We assume that $h_A(x) = 1 \implies h_B(x) = 1$ for all $x \in [0, 1]$. For instance, in the SATs, previous works[5] show that the scores are skewed for the demographic groups, with disadvantaged having a lower threshold i.e. $\tau_B < \tau_A$. Just like τ_A and τ_B are the true thresholds on the unmanipulated features, σ_A and σ_B are the thresholds upon manipulated features.

Then the learner issues a classifier f generating binary outputs and each candidate observes f and manipulates her features x to $y \geq x$. Finally the learner incurs a penalty of

$$C_{FP} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m} [h_m(x) = 0, f(y) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m} [h_m(x) = 1, f(y) = 0] \tag{4.6}$$

where C_{FP} and C_{FN} denote the cost of a false positive and a false negative respectively.

4.3.3 Result 1: Equilibrium Analysis

Suppose the cost condition says that group B members face greater costs to manipulation than group A members. Then for an unconstrained or undominated learner, this work proves the following: [18] Given group cost functions c_A and c_B and true label thresholds τ_A and τ_B where $\tau_B \leq \tau_A$, there exists a space of undominated learner threshold strategies $[\sigma_B, \sigma_A] \subset [0, 1]$ where $\sigma_A = c_A^{-1}(c_A(\tau_A)+1)$ and $\sigma_B = c_B^{-1}(c_B(\tau_B)+1)$. That is, for any error penalties C_{FP} and C_{FN}, the learner's equilibrium classifier f is based on a threshold $\sigma \in [\sigma_B, \sigma_A]$ such that for all manipulated features y:

$$f(y) = \begin{cases} 1 & \forall y \geq \sigma \\ 0 & \forall y < \sigma \end{cases} \tag{4.7}$$

To explain this analysis, if the equilibrium classifier were trained on only the samples from Group A, then $\sigma = \sigma_A$ and vice versa. While when the classifier contains samples from both Group A and B, then the equilibrium classifier return a classifier $\sigma^* \in (\sigma^B, \sigma^A)$. This is demonstrated in the figure 4.1 which shows how σ^* commits false positive on group A and false negatives on group B. This is the result of Theorem 1 discussed in this work.

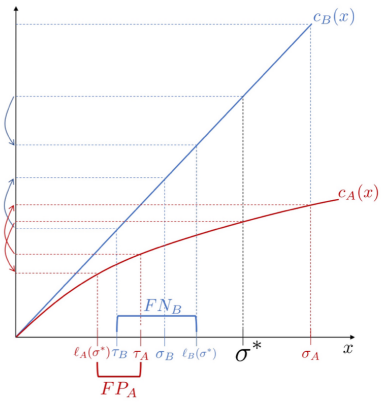

Figure 4.1: Hu et. al.[18] Figure 1, Group cost functions for a one-dimensional feature x. τ_A and τ_B signify unmanipulated true thresholds. The threshold σ_A and σ_B perfectly classify group A and B candidates; A learner selects an equilibrium threshold $\sigma^* \in [\sigma_B, \sigma_A]$, committing false positives on group A (red bracket) and false negatives on group B (blue bracket).

4.3.4 Result 2: Learner Subsidy Strategy

The second result of the work, although not directly related with our conclusion claims that Subsidies can harm both groups. The main idea behind subsidy(or Affirmative Action) is that since the strategic manipulation cost for group B is higher, perhaps the learner could provide a cost subsidy such that the learner pays a fraction $(1-\beta)$ of cost of group B. The equation 4.6 will then become:

$$C_{FP} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m}[h_m(x) = 0, f(y) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m}[h_m(x) = 1, f(y) = 0] + \lambda * cost(f, \beta) \tag{4.8}$$

If above is the learner's error, then Theorem 2 holds:

Theorem 2 (Subsidies can harm both groups). Finally, the second theorem in this work suggests that there exists cases when providing subsidy will be harmful to both the groups, and in fact both the groups would have been better of had they not tried to manipulate their score distribution. Even Group B with a subsidy of Affirmative action will suffer.

4.3.5 Comparison with our work

Comparison with our Theorem 1

The model discussed in this work conforms to our Theorem 13. As discussed in Section 4.3.2, the model discussed in this work assumes that the SAT scores are skewed for the two groups even before manipulation and therefore $\tau_B < \tau_A$. In addition, if we have the true labeling functions h_A and h_B then the assumption in this work is that

$$h_A(x) = 1 \implies h_B(x) = 1 \tag{4.9}$$

Using 4.9 we can interpret that for one particular value of x, if the individual belongs to Group B, he is expected to be more talented.

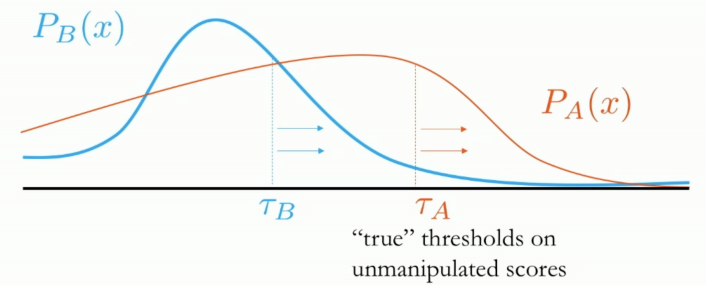

Figure 4.2: Hu et. al.[18] (figure from FATML 2019 talk) Distribution for D_A and D_B with true thresholds τ_A and τ_B

This could also be analyzed graphically as in the figure 4.2, in the region between τ_B and τ_A, if any $x \in A$ then $h(x) = 0$ while if $x \in B$ then $h(x) = 1$. In all the other regions either both are 0 or both are 1. Hence, this follows our idea of Theorem 1 (3).

4.3.6 Worldview Comparison

The assumption regarding the manipulated and unmanipulated score distribution is that both are biased against the disadvantaged group B. However at the same time, the "true" thresholds are also lower for the disadvantaged group B. This is demonstrated in figure 4.1, where $\tau_B < \tau_A$ and though the scores distributions are biased against group B, however this distribution does not represent the talents of the groups. The actual capabilities could be derived using τ_B and τ_A which are such that $\tau_B < \tau_A$ and this signifies that the talents might be relatively the same for two groups. As the figure shows that almost the similar fraction from both the groups are talented, therefore we consider this worldview to be consistent with ours.

4.4 Research Review 3: Simplicity Creates Inequity

This work by Kleinberg et. al.[23] outlines a tension between equity (fairness of predictor) and simplicity (interpretability). Although interpretability through simplicity can assist in ascertaining whether or not a decision is biased or not, simplicity could in turn increase bias.

More formally, this paper proposes a framework for producing simple prediction function and has two main results. First, is that a simplified function is strictly improvable in both equity and accuracy and hence simplification doesn't help to achieve fairness or interpretability. Second, a simple function creates incentive for the employer to use group membership information which is used against the disadvantaged members.

4.4.1 Model

This section will summarize the most relevant parts of the model to compare with our findings. The model overall represents a process of admissions or screening where an applicant is described by a set of (boolean) variables $x = (x^{\langle 1 \rangle}, x^{\langle 2 \rangle} ... x^{\langle k \rangle})$. To denote group membership, consider one more coordinate appended to the feature vector (x, A) or (x, D), where A represents the Advantaged and D represents the disadvantaged.

Assume a function f which perfectly demonstrates the productivity $f(x)$ of an applicant with features x and is independent of group membership i.e. $f(x, A) = f(x, D) = f(x)$ and the objective is to sort by $f - values$ and admit the top r fraction.

To model disadvantage for Group D, the paper consider the *Likelihood ratio condition* that if $f(x) > f(x')$ then:

$$\frac{\mu(x, A)}{\mu(x, D)} \geq \frac{\mu(x', A)}{\mu(x', D)} \tag{4.10}$$

where $\mu(x, \gamma)$ = fraction of population with features x and group membership γ. To state simply, the above disadvantage condition means that better feature vectors are more represented in Group

A.

Simplification: The simplification of a function f say g does not consider one or more of the features from the feature vector.

Given the above disadvantage condition there are two main results discussed in this paper which considers simplification. First, for every admission rule based on simplified variation of function f say g, there exists another function h such that h has a better equity (more fair) and higher accuracy than group D. Section 4.4.2 will cover an example if finding in more detail. Second, if we consider a group agnostic simplification of the function f say g by not considering one of the feature from the feature vectors— the efficiency of the resulting admission rule goes up, and the equity goes down. We conclude that even though group membership is irrelevant to the true value of f, any group-agnostic simplification of f creates an incentive for a decision-maker to use knowledge of group membership- an incentive that wasn't present before simplification, and one that hurts the disadvantaged group D. The Section 4.4.2 will cover one example to illustrate this.

4.4.2 Results

Simplicity transforms disadvantage into bias

When the true function f for ranking the applicants does not depend on an individual's group membership i.e. $f(x, A) = f(x, D) = f(x)$, then any non- trivial simplification of this function creates an incentive to use the group membership information in a way that hurts the disadvantaged group.

To demonstrate this, we now consider an example in Table 4.1. The table shows the feature vector having two binary attributes $x = [x^{\langle 1 \rangle}, x^{\langle 2 \rangle}]$, group membership γ, the utility function f and fraction of population with features x and group membership γ represented by μ.

In the table 4.1, and suppose that the true criterion is the conjunction (dot-product) of

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

Table 4.1: Table showing the features and f-values(Table taken from EC'19 talk ([22]))

$x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$. To support likelihood condition4.10, consider that applicants from group A have $x^{\langle i \rangle} = 1$ with probability 2/3 and applicants from group B have $x^{\langle i \rangle} = 1$ with probability 1/3. Using these probabilities, we can fill all the values in table 4.1. Then for all admission rates with $r \leq 5/18$, we have $f = 1$ (utility) and $\frac{1}{5}$ is the fraction of group D's representation(equity).

Now, let's consider simplifying the function f by dropping $x^{\langle 2 \rangle}$ and using only $x^{\langle 1 \rangle}$ in decision making. There are numerous reasons why shall we ignore $x^{\langle 2 \rangle}$ such as

1. Perhaps $x^{\langle 2 \rangle}$ is expensive to collect.
2. Increase interpretability of the model by reducing cognitive complexiety.
3. Removing a variable that confers disadvantage.
4. Out of sample generalization.

This simplification will hurt group D as shown in table 4.2. For all the selection rates $r \leq \frac{5}{18}$ D's representation increases from $\frac{1}{5}$ in table 4.1 to $\frac{1}{3}$ in table 4.2 but the average f-value has reduced from 1 to $\frac{5}{9}$. Hence, this simplification shows that there are gains in equity but loss in efficiency.

One may argue that since this improves the equity between two groups, simplification may help in ensuring fairness. However, this simplification will transform group D's disadvantage into bias. To elaborate this, consider the table 4.3, where g represents the simplification of true function f. Knowing group membership gives the employer conditional information and accuracy incentive about the average g values, i.e. if we had access to both $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$ as in table 4.1 then

the employer does not need to know group membership because she knows all feature vectors of the applicants. However when we drop $x^{\langle 2 \rangle}$ then as shown in 4.3 the knowledge of group membership will help improve accuracy. This is because selecting from group A will have the expected accuracy of $g = 2/3$ as opposed to $1/3$ in the group D.

Hence, The employer with the access to group membership will now hire individuals from Group A first and this hurts group D.

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	g	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

Table 4.2: Simplification: Not considering $x^{\langle 2 \rangle}$ and γ (Table taken from EC'19 talk ([22]))

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	g	μ
1	any	A	2/3	1/2
1	any	D	1/3	1/2
0	any	A	0	1/2
0	any	D	0	1/2

Table 4.3: Simplification: Not considering only $x^{\langle 2 \rangle}$ (Table taken from EC'19 talk ([22]))

Simplicity is not Pareto Optimal

Theorem shows Simplifying is not Pareto Optimal if you only care about efficiency and equity. Considering that the simplification of the true function f is g such that g is structurally simple for building more interpretable models — then it can be replaced with a another function h possibly more complex that improves on both- performance and equity. In other words, using a simple rule is not necessarily a trade-off between performance and equity, but as a step that necessarily sacrifices both properties relative to other options in the design of a rule.

If we contrast the tables 4.2 and 4.4, then table 4.3 could be re-written as 4.4 by slicing out entries of group D from the first row with $f = 1$ and placing it on the top. Now when the employer hires, it would admit group D first and then group A therefore increasing equity for $r \leq \frac{5}{18}$ and also improving accuracy as for any rate $r \leq \frac{1}{18}$, the accuracy is 1.

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	h	μ
1	1	D	1	1/18
1	any	any	1/2	8/18
0	any	any	0	1/2

Table 4.4: Not Pareto Optimal (Table taken from EC'19 talk ([22]))

4.4.3 General Theorem

The general theorem covers the two results discussed in the Section 4.4.2. The informal version of the general theorem is as stated below[23]:

Theorem: For every Boolean function f with real-valued outputs satisfying the disadvantage condition 4.10 and a genericity assumption (i.e. beyond the condition $f(x, A) = f(x, D)$, there are no "coincidental" equalities in the average values of f), then for every simplification g of f (partitioning the feature vectors into cells by fixing variables):

1. There is always an $f - approximator$ that strictly improves on g in both equity and efficiency.
2. If g does not use group membership, then by adding group membership variable will increase accuracy and reduce equity.

4.4.4 Comparison with our Theorem 1

We now compare the general theorem discussed in this paper with our Theorem 1 (Section 3) which informally states: Given that the groups have same talent distributions and consider two individuals who have the same scores, then hiring the disadvantaged individual is expected to be more accurate. *Simplicity Creates Inequity paper*[23] on the other hand concludes that when we have simplified model and two individuals have the same scores, then hiring the advantaged individual has a higher utility as demonstrated in Section 4.4.2 table 4.2.

To conclude, our finding suggests the opposite of the first result of the *Simplicity Creates Inequity paper* 4.4.2 because the model discussed in this paper is restricted to the cases of only

simple functions as discussed. In addition, this paper assumes that the function f i.e. the true function which gives the score is group agnostic i.e. $f(x, A) = f(x, D)$ which is not the case in our work as we consider that performance scores are biased according to the group.

To conclude, Theorem 1 does not overlap with the result of *Simplicity Creates Inequity paper* [23] however to explain the tension we highlight the two assumptions (Genericity and Likelihood) made in the Klienberg paper[23] which are in contrast to ours. Furthermore, we show that the commonalities in our second finding overlaps with this paper as the unconstrained learning in both works only hurts Group B.

4.4.5 Worldview Comparison

This work by Kleinberg et. al.[23] has the world's assumption that the disadvantaged group's distribution in general has fewer talented people. This paper discusses about two functions in its model, first is the true function f and the second is the simplified function h and shows that using simplified version of a function may increase bias. This assumption is therefore a discriminatory view of the world, considering that even the true function f will result in finding more talented people in the advantaged group which is the likelihood ratio condition that: If $f(x) > f(x')$ then:

$$\frac{\mu(x,A)}{\mu(x,D)} \geq \frac{\mu(x',A)}{\mu(x',D)}$$

In other words, better to worse feature vectors are according to the true distribution function f are favourable to Group A. Therefore, this research work assumes a worldview which can be deemed discriminatory, as the distribution which is generated using the true function considers that the fraction of talented individuals in the disadvantaged is lesser than the disadvantaged group. This belief is not comparable to our model's talent distribution, since we considered that the talent is equitably distributed.

4.5 Research Review 4: From Fair Decision Making To Social Equality

4.5.1 Introduction

Similar to Liu et al.[25], this paper focuses on the delayed impact(referred as dynamics) or the long term influence of applying Fairness interventions on the population. Considering that the notion of balance is eventual equality between the qualifications of the groups, this paper asks that does affirmative action(demographic parity) lead to it? This paper proposes a model which considers the dynamics similar to previous works[25, 16] which propose a utility function representing the profit/loss which is brought by an individual, and conclude whether that improves the group distribution.

The main results of this paper compares two fairness interventions — Unconstrained learning and demographic parity. It shows that unconstrained learning could reach eventual equality between the two groups given the conditions. And when Unconstrained learning doesn't reach equality, applying demographic parity could increase utility. Furthermore, although applying demographic parity may improve utility, there is a danger that the society settles at a worse-case equilibrium when under-accepting and better equilibrium when over-accepting as summarized below:

- Under-acceptance of qualified individuals was shown to guarantee equality at the cost of worse institutional utility and possibly decreasing the population's overall qualification level.
- In over-acceptance case it shows equality cannot be directly guaranteed to hold but when it does, it results, in equilibrium where the population becomes more qualified.

4.5.2 Model

The model is similar to previous works[25, 16, 19] which consider institutional utility in order to access the delayed or downstream impact on population distribution.

The group membership is represented by G where $G = A$ would represent the advantaged group with the fraction g_A and the disadvantaged group $G = B$ with fraction $g_B = 1 - g_A$. Feature Vector $\theta \in \Theta$ contains the information about the applicant's qualification such as θ= [GPA, SAT, Letters of recommendation]. This θ is implicitly mapped through an estimator $F : \theta \to \{0, 1\}$ Although this paper talks about the feature vectors, it considers a function $F : \theta \to \{0, 1\}$ which provides a crisp evaluation of of the qualification $v = 1$ if qualified and $v = 0$ otherwise. Hence the features are not discussed any further in this work.

The Qualification Profile(π) represents the probability distribution for a particular evaluation(V) and group(G) such that the qualification profile of $V = v$ in group $G = j$ is $\pi(V = v|G = j)$. The Institutional Policy (τ is defined by an institution or a policy maker, which maps each individual to a policy of selection $\tau(V = v; G = j) : \{0, 1\} \times \{A, B\} \to \{0, 1\}$.

Institutional Utility $U(\tau)$ is defined considering that $u\{0, 1\} \to R; v \to u(v)$ to be the utility function for *an* individual such that $u(0) \leq 0 \leq u(1)$. Then $U(\tau)$ is given by:

$$U(\tau) = \sum_{j \in \{A,B\}} g_j \sum_{v \in \{0,1\}} u(v).\tau(V = v; G = j).\pi(v|G = j)$$

The Selection Rates(β) for a group j are defined as: $\beta(G = j) = \sum_{v \in \{0,1\}} u(v).\tau(V = v; G = j).\pi(v|G = j)$

We also define selection rate per v value as: $\beta(V = v; G = j) = \tau(V = v; G = j).\pi(V = v|G = j)$

4.5.3 Dynamics

The execution of a selection policy can be thought of as demarcating time t. The main idea of the paper is to looks at the effects of the selection process and how it affects the population. Hence, we would be interested in looking at the difference between the qualification profiles (π) at time t and $t+1$. Consider the following assumption

Assumption 1 (Dynamics): For a given group j , let $\pi_t(1|j) =: \pi_t(1)$ denote the qualification profile of group j for $v = 1$ at time t and let the policies τ_t at that time step induce the selection rates β_t. Then the qualification profiles at time $t+1$ are given by:

$$\pi_{t+1}(1) = \pi_t(1) * f_1\big(\beta_t(0), \beta_t(1)\big) + \pi_t(0) * f_0\big(\beta_t(0), \beta_t(1)\big)$$

Where f_0 and f_1 are two arbitrary continuously differentiable functions from $[0,1] \times [0,1] \to [0,1]$. The pair (f_0, f_1) is referred to as the dynamics. For each group $G = j$, $\pi_t(.|j)$ describes the potential qualification profile. In the above assumption the function f_1 represents the retention at the top i.e. the rate of retention of the sub-population with potential $v = 1$ due to the current policy, and f_0 represents change for the better. Similarly, $1 - f_1$ would represent change for the worse and $1 - f_0$ is the retention at the bottom.

Under given dynamics (f_0, f_1), a policy is said to be equalizing if for all starting $\pi_0(1|A)$ and $\pi_0(1|B)$; we have $\lim_{t\to\infty} |\pi_t(1|A) - \pi_t(1|B)| = 0$.In other words, the population distribution shall look the same after sufficient iterations.

4.5.4 Assumptions & Definitions

To demonstrate the final results, this paper considers a few definitions and assumptions which are as demonstrated below:

Unconstrained maximization would mean that the employer would maximize the utility.

$$\max_\tau U_t(\tau)$$

Also, it is straightforward to see that for the Unconstrained policy, the optimal policies are $\tau_t(1;.) = 1$ and$\tau_t(0;.) = 0$ for all time t.

Affirmative Action: The affirmative action constraint forces the policy to select at an equal rate between the two groups, i.e.:

$$\beta(A) = \beta(B)$$

Within affirmative action, There are two possible cases through which affirmative action impacts the policy, denoted by AA^+ and AA^- which stands for Over-acceptance and Under-acceptance. These represent two drastically different approaches to fairness.

Under-acceptance (AA^-)

$$\text{If } g_j * u(1) + (1 - g_j) * u(0) \leq 0, \text{ then}$$
$$\tau_t(1; j) = \tfrac{\pi_t(1|\neg j)}{\pi t(1|j)}, \tau_t(0; j) = 0 \text{ (under-acceptance)},$$
$$\tau_t(1; \neg j) = 1, \tau_t(0; \neg j) = 0 \ ,$$

AA^- (under acceptance) accepts fewer qualified individuals from the advantaged group so as to equalize the selection rates for qualified individuals between both groups. One could think of AA^- as increasing the standard for the advantaged group and as such reducing total selection rates. Similarly we define over-acceptance:

Over-acceptance(AA^+): One could think of it as reducing the standard for the disadvantaged group.

$$\text{If } g_j * u(1) + (1 - g_j) * u(0) > 0, \text{ then}$$
$$\tau_t(1; j) = 1, \tau_t(0; j) = 0$$
$$\tau_t(1; \neg j) = 1, \tau_t(0; \neg j) = \tfrac{\pi_t(1|j) - \pi_t(1|\neg j)}{1 - \pi t(1|\neg j)} \text{ (over-acceptance)}$$

Finally, also consider assumption 2 before we outline the final results.

Assumption 2: The dynamics under the unconstrained policy (UN) can be written as $f(\pi) := \pi f_1(0,\pi) + (1-\pi)f_0(0,\pi)$ and we assume that f is $L_{UN} - Lipschitz$ with $L_{UN} < 1$, meaning that $\forall \pi, \pi' \in [0,1]$:

$$|f(\pi) - f(\pi')| \leq L_{UN}|\pi - \pi'| \tag{4.11}$$

4.5.5 Results

The main results in this paper could be summarized as follows: affirmative action (demographic parity) is considered as the mean to achieve equality in the qualifications of different groups. Imposing of affirmative action with the under-acceptance strategy of qualified individuals was shown to guarantee equality but at the cost of worse institutional utility and a decrease in the population's overall qualification level. In the second strategy of affirmative action i.e. the over-acceptance of unqualified individuals, however to lead to a policy with different characteristics: equality cannot be directly guaranteed to hold but when it does, it results, in equilibria where the population becomes more qualified.

The first result considers the UN strategy which is similar to our unconstrained strategy. This papers looks at the dynamics over time for the UN, where it informally states that if UN achieves equilibrium, it achieves it in a similar way as that of under-acceptance, while there are cases when equilibrium is not achieved. This is illustrated in **Theorem 1** and figure 4.3

Theorem 1: If equality in population distributions is reached with an unconstrained (UN) policy by way of assumption 2, then it is necessarily reached by an AA^- policy implemented over all time steps, however with no more, and possibly less, utility at each step.

The second result discussed in this paper informally states that whenever AA– equalizes dynamics, it always leads to worse long-term utility than under UN by leading to a population with lower qualification. On the other hand when following AA+, equality is always beneficial. This is

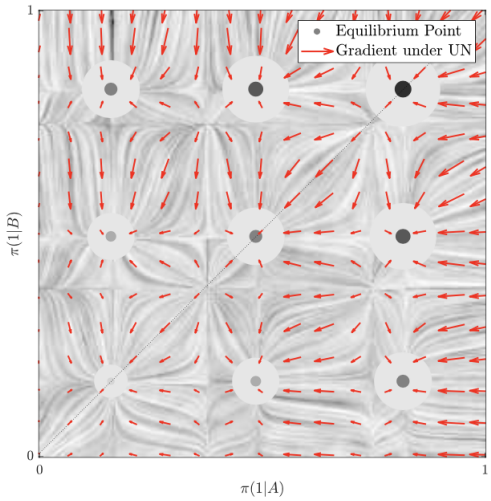

Figure 4.3: Points showing unconstrained equilibriums(Figure 3 from the paper Srebro et. al. [26])

as demonstrated in the figure 4.4.

Theorem 4: If the policy is AA–, then the equalized population generates long-term utility no higher (and possibly lower) than the limiting population under UN. If the policy is AA+ and it leads to social equality, then the equalized population generates long-term utility no lower (and possibly higher) than the limiting population under UN.

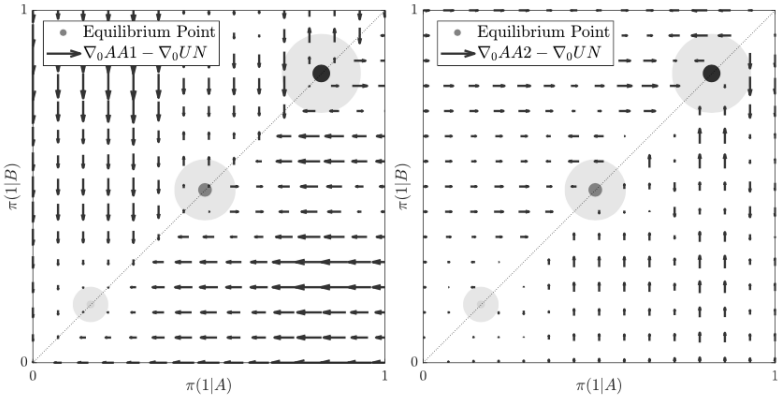

Figure 4.4: Equilibrium points for AA^- and AA^+ respectively (Figure 3 from the paper Srebro et. al. [26])

4.5.6 Conclusion

This paper studies dynamics in affirmative action to equalize the qualifications of different groups. Imposing under acceptance was shown to guarantee equality at the cost of worse institutional utility and possibly decreasing the population's overall qualification level. Over acceptance affirmative action was shown to lead to a policy with different characteristics: equality cannot be directly guaranteed to hold but when it does, it results, in equilibria where the population becomes more qualified.

4.5.7 Comparison with our final results

Our Theorem 1's (3) findings are irreconcilable with this paper's model since this paper does not talk about any bias in the scores. In addition, there is no noise which is considered in the evaluation of labels for the individuals as the work considers a function $F : \theta \to \{0, 1\}$ which provides a crisp evaluation of of the qualification $v = 1$ if qualified and $v = 0$ otherwise. Since our first theorem is based on the bias in the scores, we cannot compare it with this paper.

4.5.8 Worldview Comparison

As discussed in detail the Section 4.5, this work considers the dynamics of applying fairness constraints on the population distribution. Hence, although this work considers that the qualification profiles (π) of different groups are biased such that the disadvantaged group have lower possible score, but applying the fairness constraints eventually achieves equality amongst the groups.

This model considers dynamics and hence is not directly comparable to our model. However, the worldview of this model is still similar to our Talent distribution as after applying the fairness constraint Demographic Parity, this model achieves suggests that there is an equality within the qualification profiles of the individuals. The main result in this paper could be summarized as follows: affirmative action (demographic parity) is considered as the mean to achieve equality in

the qualifications π as $t \to \infty$ of different groups.

To conclude, this work's worldview assumption demonstrates that although initially the qualification profiles are distorted and biased against the disadvantaged group, however over the period of time, applying demographic parity will result in equalizing the population's qualifications. Hence, the assumption aligns with our model's Talent distribution equality claim, as over the period of time the qualification profiles becomes equal.

4.6 Research Review 5: Delayed Impact of Fair Machine Learning

This work amongst only a few others considers the delayed impact or dynamics of the fairness interventions we employ to benefit the disadvantaged group. The paper shows that even in one-step feedback model, common fairness criteria such as demographic parity and equal opportunity do not promote improvement in certain cases while the unconstrained utility maximization does. Another interesting finding of this work is with the consideration of Measurement Error, which broadens the regime in which fairness criteria perform favourably.

The special case of measurement error in this paper is comparable to our model of Talent and Environment as described in section 3.1.1 and we also show that the result with measurement error match with our final results in Theorem 1 (3). The below section will summarize the population distribution and model setup and how it compares to our findings.

4.6.1 Contributions

Given the one-step feedback model[25] and group A represents the disadvantaged group and B the advantaged. The main results of this work could be summarized in the following three points:

1. The two fairness criteria discussed (demographic parity, equal opportunity) can lead the three possible outcomes i.e. improvement, stagnation, and decline in the the change in score distributions in natural parameter regimes. Also, there are a class of settings where equal selection rates cause decline, whereas equal true positive rates do not.

2. This paper also introduces the concept of outcome curve which helps compare Fairness Regimes/interventions in the scores-utility setting discussed in the model (4.6.2).

3. Finally, the paper introduces certain types of measurement errors (e.g., the banks underestimating the repayment ability only for the disadvantaged group) affect the comparison. This is the part where we juxtapose our results with this work and find similarities.

4.6.2 Model

This section briefly discusses the main aspects of the model. The Group A in this work is regarded as the disadvantaged group while B the advantaged, which is the opposite to all the papers we reviewed. Also, g_A and $g_B = 1 - g_A$ represent the respective fractions of the total population.

The respective score distributions are π_A and π_B , $\Delta\mu_j$ represents the change in score distribution for group j, which we represents long-term improvement if $(\Delta\mu_j > 0)$, stagnation if $(\Delta\mu_j = 0)$, and decline if $(\Delta\mu_j < 0)$.

The institution's policy $\tau = (\tau_A, \tau_B)$ are chosen by the institution which corresponds to the probability the institution selects an individual in group j with score $x \in \mathcal{X}$. We assume that the institution is utility-maximizing. There exists a function $u : C \to R$, such that the institution's expected utility for a policy τ is given by:

$$U(\tau) = \sum_{j \in \{A,B\}} g_j \sum_{x \in \chi} \tau(x)\pi_j(x)u(x) \tag{4.12}$$

This work also defines the outcomes in terms of an average effect that a policy τ_j has on

group j. Formally, for a function $\Delta(x) : \mathcal{X} \to R$, average change of the mean score μ_j for group j is represented as:

$$\Delta\mu_j(\tau) = \sum_{x \in X} \pi_j(x)\tau_j(x)\Delta(x) \tag{4.13}$$

Where $\Delta(x)$ is the change in the score values. So better scores represent better life and well-being in general.

Finally, an assumption that the success of an individual is independent of their group membership given the score x. That is, the scores are without noise and can tell the talent of an applicant with certainty. Therefore, this work considers a function $\rho : X \to [0, 1]$ such that individuals of score $x \in X$ succeed with probability $\rho(x)$. This assumption is formally detailed assumptions Section 4.6.4.

4.6.3 Outcome Curve

The paper introduces a graphical tool called the outcome curve to determine if a fairness constraint causes *benefit* or *harm* to the scores distribution after classification.

A policy (τ_A, τ_B) is said to cause a group:

1. active harm to group j if $\Delta\mu_j(\tau_j) < 0$
2. stagnation if $\Delta\mu_j(\tau_j) = 0$
3. and improvement if $\Delta\mu_j(\tau_j) > 0$.

The *MaxUtil* policy make the employer makes the most profit and is chosen in a standard fashion which applies the *same* threshold $\tau^{MaxUtil}$ to both groups agnostic of the distributions π_A and π_B (Change in mean scores are represented as $\Delta\mu_j^{MaxUtil} = \Delta\mu_j(\tau^{MaxUtil})$).

The paper considers that a policy causes relative harm to group j $\Delta\mu_j(\tau_j) < \Delta\mu_j^{MaxUtil}$, active harm if $\Delta\mu_j(\tau_j) < 0$ and relative improvement if $\Delta\mu_j(\tau_j) \geq \Delta\mu_j^{MaxUtil}$. The selection rates

for these thresholds are $\beta_j := \sum_{x \in X} \pi_j(x)\tau_j(x)$, demonstrated in the Figure 4.5.

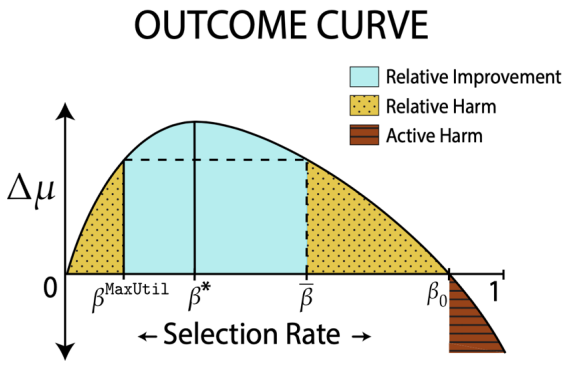

Figure 4.5: Outcome Curve (Figure 1 from Liu et. al. [25])

In the outcome curve figure 4.5, the following rates are of interest. The paper defines $\beta^{MaxUtil}$ as the selection rate for A under MaxUtil Policy. The paper also defines β_0 as the harm threshold, such that $\Delta\mu_A(r_{\pi_A}^{-1}(\beta_0)) = 0$. Similarly β^* is defined as the selection rate such that $\Delta\mu$ is maximized. Finally, $\bar{\beta}$ is defined as the outcome-complement of the MaxUtil selection rate.

The selection rate β^* could also be regarded as the philanthropic optimal threshold as the benifit to the Group A is maximum. Also any selection rate in the Relative Harm or Active harm region are not desirable results as the $\beta^{MaxUtil}$ which is the default behaviour of banks perform better than them.

4.6.4 Results

Using on the outcome curve (fig 4.5), this paper covers a proposition and 5 corollaries which are based upon an assumption. This section will informally outline the assumption and the main corollaries.

Assumptions

Assumption 1: The institution's individual utility is more stringent than the expected score changes, $u(x) > 0 \rightarrow \Delta x > 0$, In other words, in the credit risk setting, if an individual defaults on

a loan than the bank suffers a higher loss ratio than the individual score change ratio i.e.:

$$\frac{u_-}{u_+} < \frac{c_-}{c_+} \tag{4.14}$$

Assumption 2 Monotonocity: Assume that the success of an individual is independent of their group given the score i.e. the score summarizes all relevant information about the success event, so there exists a function $\rho : X \to [0, 1]$ such that individuals of score x succeed with probability $\rho(x)$. Also, higher scores means higher probability of success. i.e. ρ is strictly increasing in x independent of the group.

$$\text{if } x_1 > x_2 \text{ then } \rho(x_1) > \rho(x_2) \tag{4.15}$$

Corollaries

Based on the discussed assumptions, below are the relevant corollaries discussed in this work. For the corollaries, the selection rates in case for Demographic Parity must be equal, but for Equal Opportunity we will need to define a transfer function which is a mapping of selection from Group A to B, $G^{(A \to B)}$, which for every loan rate β in group A gives the loan rate in group B that has the same true positive rate.

To summarize, the corollaries 3.2 states that applying any Fairness Criteria can cause Relative Improvement, while corollaries 3.3 and 3.4 states that there exists scenarios where all fairness criterion could cause active harm. Finally, corollary 3.5 and 3.6 states the conditions where one fairness constraint fails when others don't.

Corollary 3.2

If the assumption 4.14 holds, then corollary 3.2 states that any fairness criteria can cause relative improvement. The below scenarios are for demographic parity and equal opportunity respectively.

1. Under the assumption that $\beta_A^{MaxUtil} < \bar{\beta}$ and $\beta_B^{MaxUtil} > \beta_A^{MaxUtil}$, there exist population proportions $g_0 < g_1 < 1$ such that, for all $g_A \in [g_0, g_1]$, $\beta_A^{MaxUtil} < \beta_A^{DemParity} < \beta$. That is, DemParity causes relative improvement for group A.

2. Similar to the demographic parity, now consider the equal opportunity case. Under the assumption that $\beta_A^{MaxUtil} < \beta < \beta' < \bar{\beta}$ such that $\beta_B^{MaxUtil} > G^(A \to B)(\beta), G^(A \to B)(\beta')$, there exist population proportions $g_2 < g_3 < 1$ such that, for all $g_A \in [g_2, g_3]$, $\beta_A^{MaxUtil} < \beta_A^{DemParity} < \beta$. That is, Equal Opportunity causes relative improvement for group A.

Corollary 3.3 & 3.4

Corollaries 3.3 and 3.4 are comparable and conclude that DemParity and EqOpt can cause harm by being over eager in the selection rates.

For Demographic parity suppose that we fix the selection rate β which is same for the two groups and assume that $\beta_B^{MaxUtil} > \beta > \beta_A^{MaxUtil}$. Then, there exists a population proportion g_0 such that, for all $g_A \in [0, g_0]$, $\beta_A^{DemParity} > \beta$. In particular, when $\beta = \beta_0$, DemParity causes active harm, and when $\beta = \bar{\beta}$, DemParity causes relative harm.

Similarly, taking the transfer function, we assume the same for EqOpt. Suppose that $\beta_B^{MaxUtil} > G^{(A \to B)}(\beta)$ and $G^{(A \to B)}(\beta) > \beta_A^{MaxUtil}$. Then, there exists a population proportion g_0 such that, for all $g_A \in [0, g_0]$, $\beta_A^{EqOpt} > \beta$. In particular, when $\beta = \beta_0$, DemParity causes active harm, and when $\beta = \bar{\beta}$, DemParity causes relative harm.

Corollary 3.5 & 3.6

This section is the comparison of Demographic parity and Equal opportunity, and this section states that there exists scenarios where DemParity performs better than EqOpt and vice versa.

As is evident from the model, in order to compare EqOpt and DemParity, we need to have a knowledge of the full population distributions π_A & π_B which will be used to compute the transfer

function $G^{(A \to B)}$. Hence, the corollaries 3.5 & 3.6 states that if we don't have the knowledge of the function $G^{(A \to B)}$, then EqOpt may avoid active harm where DemParity fails and vice versa.

Fairness Under Measurement Error

This paper initially considers no error in the individual scores. However, it could be the case where the disadvantaged group's scores are systematically underestimated, while the scores for the advantaged group are not. Under such a scenario, this model is comparable to our work.

To define measurement error, the estimate of an individual's score $X \sim \pi$ is prone to errors $e(X)$ such that $X + e(X) := \hat{X} \sim \hat{\pi}$. Since the scores are underestimated for the disadvantaged group, the error for an individual $e(X)$ is negative. In this setting, it is equivalent to consider the CDF of underestimated distribution $\hat{\pi}$ to be dominated by the CDF true distribution π, i.e. $\sum_{x \geq c} \hat{\pi}(x) \leq \sum_{x \geq c} \pi(x)$ for all $c \in C$, where C is the score range.

The paper then suggests a Preposition that given underestimation, the selection rate β_A falls for the disadvantaged group. Suppose $\hat{\beta}$ represents the new selection rate for the underestimated scores then $\beta_A^{MaxUtil} > \hat{\beta_A}^{MaxUtil}$ and $\beta_A^{DemParity} > \hat{\beta_A}^{DemParity}$. Also, if the errors are further such that the true TPR dominates the estimated TPR, it is also true that $\beta_A^{EqOpt} > \hat{\beta}_A^{EqOpt}$.

4.6.5 Comparison with our final results

The concept of measurement error (4.16) is akin to the noisy scores X, which are dependent on talent and environment in our work (3.1.1). Therefore, we compare our model to the measurement error section 4.6.4 where the scores are underestimated only for the disadvantaged group. The two sections below conclude whether our theorems hold in this paper's model.

Theorem 1

Theorem 1(3) holds with the delayed impact paper which states that — If we compare two individuals with the same scores such that one belongs to the advantaged group and the other belongs to the disadvantaged group, hiring the disadvantaged individual is expected to have higher talent value.

$$X + e(X) := \hat{X} \sim \hat{\pi} \tag{4.16}$$

To prove that our first theorem holds in this model, let's consider two individuals with scores $X'_A = X'_B = X'$ where one individual belongs to the disadvantaged group A and the other to the advantaged group B. However since the distribution for group A is $\hat{\pi}(x)$ and for B its $\pi(x)$, the score X'_A is underestimated and the actual true score say $X_A^{true} = X'_A - e(x))$. As e(x) is negative, $X_A^{true} > X'_A$.

From equation 4.15, we know that if $x_1 > x_2$ then $\rho(x_1) > \rho(x_2)$ and hence, $\rho(X''_A) > \rho(X'_B)$ which shows that in the measurement error model, hiring disadvantaged individuals with the same score is expected to have higher success rate.

To conclude, our Theorem 1 holds this paper's model when we assume the measurement error in group A's score measurements.

4.6.6 Worldview Comparison

This research work also considers dynamics where the scores distribution of a population is represented by π_A and π_B for groups A and B respectively and these change overtime. This work considers that the mean of the two population's score distributions is biased against disadvantaged to start with, and when we apply fairness interventions these distributions change. Similar to the dynamics paper (Section [26]), the main aim of this paper is to analyze different fairness criteria and equalize the different population groups.

To sum up, this work considers dynamics or downstream effects into account, i.e. the population distributions are biased against group A to start with, and later on with the fairness intervention the bias could fade and eventually we could achieve equality. Therefore, this model also aligns with our model's Talent distribution equality claim, as over the period of time the distributions reach equality.

4.7 Research Review 6: Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?

4.7.1 Introduction

This paper was one of the first motivation behind our basis for comparing fairness constraints. As we saw in the previous works, several fairness constraints have been proposed in the literature which recognize that certain demographic groups are treated unfairly and propose rules to fix it. However this paper consider a different motivation.

This work suggests that if the training data itself is biased in certain way (including having a more noisy or negatively biased labeling process on members of a disadvantaged group, or a decreased prevalence of positive or negative examples from the disadvantaged group, or both) then applying fairness interventions could actually improve accuracy.

Given the biased training data for the disadvantaged group, the main finding of this work is that an ERM learner subject to the equal opportunity fairness constraint recovers the Bayes optimal hypothesis, making it an attractive choice for decision makers whose overall concern is purely about accuracy on the true data distribution. In deriving this finding, this paper contemplates several fairness interventions such as Demographic Parity, Equalized Odds and data re-weighting.

There are a few assumptions in this work. First, that the Bayes optimal classifiers (h^*_A and h^*_B) classify the same fraction (p) of the respective populations as positive. Second, that both

the population distributions (advantaged and disadvantaged) have the same error rate (bias) η with respect to h_A^* and h_B^* and that these errors are uniformly distributed. Finally, only the training data for the disadvantaged population is then biased with the two biased models discussed.

Considering these assumptions, this work proposes that only equal opportunity constraint will extract the Bayes optimal classifier.

4.7.2 Model

We assume the data lies in some instance space $\mathcal{X}$, such that $\mathcal{X} \in R^d$, and two groups, Group A and Group B, such that $P(x \in A) = 1 - r$ and $P(x \in B) = r$ where $r \in (0, 1)$. To know the group membership, assume that there is a special coordinate of the feature vector x, which denotes group. The data distribution is given by $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$. Assume there exists a pair of Bayes Optimal Classifiers $h^* = (h_A^*, h_B^*)$ where $h_A^*, h_B^* \in \mathcal{H} : \mathcal{X} \to \{0, 1\}$ and $h_A^* \neq h_B^*$ then the true labels for training are generated in such a form:

$$y = y(x) = \begin{cases} \neg h^*(x) & \text{with probability } \eta \\ h^*(x) & \text{with probability } 1 - \eta \end{cases}$$

The labels y after this flipping process are the true labels of the training data.

Assumption 1: $p = P(h_A^*(x) = 1 | x \in A) = P(h_B^*(x) = 1 | x \in B)$ — That is equal positive rates across groups. Hence, η is the same for both the groups and fraction of +ve samples:
$p(1 - \eta) + (1 - p)\eta$

Assumption 2: The paper also considers that the Bayes Optimal Classifier is different for the two groups, i.e., $h_A^* \neq h_B^*$. If h_A^* was also optimal for Group B, then we can just learn h^* for both Groups A and B using data only from Group A and biased data concerns fade away. Thus we are learning a pair of classifiers, one for each demographic group.

Assumption 3: The that h^* is not perfect and independently with probability η, the true label of

x does not correspond to the prediction $h^*(x)$.

4.7.3 Bias in Training Data:

As compared to other works, the bias model of this paper considers only two specific types of biases in the training data only.

1. Under-representation Bias β: In this bias model, the positive examples from Group B (disadvantaged) are under-represented in the training distribution while for Group A, the training data reflects the true data. Hence, this model consider a probability β with which a positively labeled sample is considered in the training data.

2. Labeling Bias ν: Quite similar to the Under-representation Bias, in this model instead of removing the sample from the training data, its label is flipped from positive to negative with a probability ν. Thus, for each pair (x, y), if $x \in B$ and $y = 1$, then independently with probability ν , the label of this point is flipped to negative.

Given these two bias models, this paper then discusses the sampling of training set and then applies the above two biases on it. A classifier h satisfies equal opportunity if $P_{(x,y)\sim D}(h(x) = 1|y = 1, x \in A) = P_{(x,y)\sim D}(h(x) = 1|y = 1, x \in B)$.

4.7.4 Results

This work is mainly about discovering the Bayes Optimal Classifier and compares 4 fairness interventions in the ERM's resultant classifiers which are— Equal Opportunity, Equalized Odds, Demographic Parity and Data Reweighing. Of the four constraints, only equal opportunity is able to extract the Bayes optimal classifier given the model setup and assumptions. Theorem 1 outlines the condition which should hold in order for equal opportunity constraint to extract the Bayes optimal classifier.

Theorem4.1: Assume true labels are generated by $P_{D,r}(h^*, \eta)$ corrupted by both Under Representation bias and Labeling bias with parameters $\beta_{POS}, \beta_{NEG}, \nu$, and assume that

$$(1-r)(1-2\eta) + r((1-\eta)\beta_{POS}(1-2\nu) - \eta\beta_{NEG}) > 0 \tag{4.17}$$

and

$$(1-r)(1-2\eta) + r((1-\eta)\beta_{NEG} - (1-2\nu)\beta_{POS}\nu) > 0 \tag{4.18}$$

Then $h^* = (h^*_A, h^*_B)$ is the lowest biased error classifier satisfying Equality of Opportunity on the biased training distribution and thus h^* is recovered by Equal Opportunity constrained ERM.

On the other hand the other three fairness interventions fail. The paper has simple examples about how they fail.

4.7.5 Comparison with our final results

For the comparison, we now discuss about the distributions for the scores and labels two groups. The distribution of scores in this work is defined as $\mathcal{D}$ and is a pair distributions $(\mathcal{D}_A, \mathcal{D}_B)$, with $\mathcal{D}_A$ determining how $x \in A$ is distributed and $\mathcal{D}_B$ determining how $x \in B$ is distributed. In addition, the Bayes optimal hypothesis $h^* = (h^*_A, h^*_B)$ is such that $h^*_A \neq h^*_B$. Finally, the true labels generated for a x which is drawn from the distribution $\mathcal{D}$ is defined as:

$$y = y(x) = \begin{cases} \neg h^*(x) & \text{with probability } \eta \\ h^*(x) & \text{w. p. } 1-\eta \end{cases} \tag{4.19}$$

Our Theorem 1's findings do not hold in this paper's model and assumptions. This is mainly due to the different bias models in our work and in this paper. The paper's model simply assumes that the distributions for the two groups are different and so is their Bayes hypothesis,

there is no comparison on how the scores distributions are different. For example, it might be possible that the distribution D_A and D_B are Normally distributed where the mean for group B is higher than the mean for Group A and still the findings of this work will hold.

On the other hand, our model 3.1.1 assumes that the environment is biased in such a way that it harms Group B, and hence the scores for group B have a lower distribution than group A. Hence, in this paper's model setting, if we compare two individuals with the same score (x) values from the two groups, it is not guaranteed that the disadvantaged individual is expected to be more talented.

4.7.6 Conclusion

To conclude, this paper shows that equal opportunity constrained ERM will recover from 2 types of training data bias, including Under-Representation Bias and Labeling Bias, in a clean model where the Bayes-Optimal classifiers h_A^*, h_B^* satisfy most fairness constraints on the true distribution and the errors of h_A^*, h_B^* are uniformly distributed.

This paper is limited to only two types of biases, which are not practical, however the findings that equal opportunity constraints recover the Optimal classifiers still hold with our Theorem 2's findings. Furthermore, theorem 1 is not comparable to this work, given this work's bias distribution.

4.7.7 Worldview Comparison

Discussed in section 4.7, this research work starts with the assumption that the true distribution is not biased but its only the training distribution that is biased. It views the world in such a way that the training data we capture is biased due to human error, or historical bias, however the real-world test data is free from such bias.

To conclude, the model discussed in this paper has two aspects, first is the real world

"true" data which is unbiased and represents the true world and the second aspect is that of the biased training data, which has misrepresentation and under-representation of disadvantaged group. Therefore, we could compare our Talent distributions with this model's true real world distributions, both of which are unbiased.

Chapter 5

Conclusion and Future Work

In this chapter we summarise our work and present some concluding remarks. We also discuss the potential future work.

5.1 Conclusion

The work in this thesis can be summarized as follows:

1. First we present an overview of the recent research works in machine learning fairness with the objective to accustom a beginner in the field with the terms and research methodologies commonly used in the literature.

2. We then propose a model with the assumption that the talent distributions for the two groups are equal and its only because of the environment that the individuals of disadvantaged groups perform worse on tests. Furthermore, we hypothesize that the test or performance scores are not simply a representation of an individual's talent but also of the support environment available around her.

3. Third, we propose our main theorem in which we considered for several distributions of Talent

and Environment and showed for which distributions our claim holds. The claim of the theorem being that, given two individuals having similar performance on a screening test, the individual from the disadvantaged group is expected to be more talented than the advantage group.

4. Finally, we have a comparative review of six recent and relevant research works in group fairness where we cover detailed summary of the model, assumptions and main conclusions of each of the research works. In addition, we verify whether our theorem's main claim is consistent in each of the research work's settings. We also analyze the different worldviews, i.e. how the model discussed in these research works view the population distribution of the world. We see that while there are a few models whose worldviews closely align with our model's underlying belief– that talent is the evenly distributed for both the groups– there are others which view the world from a discriminatory standpoint.

5.2 Future Work

There are a few limitations in our work which could be addressed in the future continued works. We briefly discuss three such assumptions:

1. Our first assumption is that the performance scores of an individual are equal to the sum of talents and the environment. However, there could be other factors which could effect the performance of an individual on a test, and we could extend the current model to account for them. For example, if an individual is talented and also has a conducive environment around him, still it is possible for him to perform bad on a test because he is not driven or ambitious. Furthermore, there could be a luck factor associated when measuring the performance score, for instance if a talented and good environment student is just unlucky and arrived late for the exam.

2. Another limitation that our model has is that we always assume that the talent is same for

the two groups. Here, our model could be extended to allow the two groups to have different distributions, which could be useful for real world scenarios. For instance, when applying to armed forces, generally the male would outperform female candidates.

3. We also assume that the performance score is the linear "sum" of talent and environment for an individual, however this study could be extended to consider more complex functions such as a quadratic, or polynomials.

Bibliography

[1] Blum Avrim and Stangl Kevin. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 3:1–3:20, 2020.

[2] Ted Bergstrom and Mark Bagnoli. Log-concave probability and its applications. *Economic Theory*, 26:445–469, 08 2005.

[3] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524. Association for Computing Machinery, 2020.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[5] David Card and Jesse Rothstein. Racial segregation and the black-white test score gap. *Journal of Public Economics*, 91:2158–2184, 02 2007.

[6] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv:2010.04053*, 2020.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 2016.

[8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 10 2016.

[9] Jung Christopher, Kearns M., Neel Seth, Roth Aaron, Stapleton Logan, and Wu Z. Eliciting and enforcing subjective individual fairness. *ArXiv*, abs/1905.10660, 2019.

[10] Dwork Cynthia, Immorlica Nicole, Kalai Adam, and Leiserson Max. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133. PMLR, 23–24 Feb 2018.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 04 2011.

[12] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64:136–143, 2021.

[13] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, 2019.

[14] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2017, page 498–510. Association for Computing Machinery, 2017.

[15] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 269–278. Association for Computing Machinery, 2019.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331. Curran Associates Inc., 2016.

[17] Jiang Heinrich and Nachum Ofir. Identifying and correcting label bias in machine learning. In *AISTATS*, 2020.

[18] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 259–268. Association for Computing Machinery, 2019.

[19] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 240–248. Association for Computing Machinery, 2019.

[20] Kailash Karthik Saravanakumar. The Impossibility Theorem of Machine Fairness – A Causal Perspective. *arXiv e-prints*, page arXiv:2007.06024, July 2020.

[21] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *American Economic Association Papers and Proceedings*, 108:22–27, 05 2018.

[22] Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, page 807–808. Association for Computing Machinery, 2019.

[23] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, pages 43:1–43:23, 2017.

[24] Cohen Lee, Lipton Zachary, and Mansour Yishay. Efficient candidate screening under multiple tests and implications for fairness. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, pages 1:1–1:20, 2020.

[25] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6196–6200. International Joint Conferences on Artificial Intelligence Organization, 2019.

[26] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 359–368. Association for Computing Machinery, 2019.

[27] Zafar Muhammad B., Valera Isabel, Gomez-Rodriguez Manuel, and Gummadi Krishna P. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[28] Arvind Narayanan. 21 fairness definition and their politics by arvind narayanan. `https://fairmlbook.org/`. Accessed: 2018-01-01.

[29] Kilbertus Niki, Gomez-Rodriguez Manuel, Schölkopf Bernhard, Muandet Krikamol, and Valera Isabel. Improving consequential decision making under imperfect predictions. *CoRR*, abs/1902.02979, 2019.

[30] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual amp;group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, page 2239–2248. Association for Computing Machinery, 2018.

[31] Songül Tolan, Marius Miron, Emilia G"omez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, page 83–92. Association for Computing Machinery, 2019.

[32] Y. Yan, W. Wang, X. Hao, and L. Zhang. Finding quasi-identifiers for k-anonymity model by the set of cut-vertex. *Engineering Letters*, 26:150–160, 02 2018.

[33] Samuel Yeom and Michael Tschantz. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 33–41, 08 2018.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment amp; disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180. International World Wide Web Conferences Steering Committee, 2017.