ENSURING FAIRNESS WITH SUPPORT ENVIRONMENT

KARAN DEEP SINGH

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTERS OF SCIENCE

GRADUATE PROGRAM IN DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO
MAY 2021

ENSURING FAIRNESS WITH SUPPORT ENVIRONMENT

by **Karan Deep Singh**

a thesis submitted to the Faculty of Graduate Studies of York University in partial fulfilment of the requirements for the degree of

MASTERS OF SCIENCE

ENSURING FAIRNESS WITH SUPPORT ENVIRONMENT

by **Karan Deep Singh**

Examination Committee Members:

1. Dr. Jeff Edmonds
2. Dr. Ruth Urner
3. Dr. Aijun An
4. Dr. Neal Madras

Abstract

Several fairness constraints have been proposed in the Machine Learning Research literature to rectify the issue of certain demographic groups being treated differently by classifiers. Mitigation of such bias which creeps into the classification systems is the main motivation for our research. This work could broadly be classified into two categories. The first is we propose a new Theorem which states that given two individuals who have the same performance on a test, the individual from the disadvantaged group is expected to be more talented than the one in the advantaged group since in general the advantaged group individual has better support environment around him. In this theorem, we propose a new theoretical model which considers that the performance of an individual is a combination of her "Talent" and the "Environment" around her. Our modest aim with the theorem's statement is that it would help make employers more accurate and fair decisions. The second category is thorough literature review which looks at seven recent research works in Machine Learning Fairness. We analyse the different research works with respect to our theorem and analyze that given the respective model and assumptions, does our theorem's claim hold in research work's setting. Finally, we comment on the different worldviews of research works, i.e. what population distribution is considered by each research work when claiming their results and comparing it with our model, if they assume that the groups have the same Talent distribution.

Given the substantial research in the field of Fairness in Machine Learning, the first chapter of the thesis aims to provide an entry-level overview of the common definitions, fairness interventions and metrics which are essential for a novice reader. This section highlights the terminology common in the field and gives an overview of the current approaches of achieving fairness. The second chapter is a discussion and formal proofs of the first theorem where each section looks into different distributions and conditions for the theorem to hold and a discussion no the distributions where the claim would not hold. The final chapters summarizes the seven research works we looked in detail and discuss the result of theorem 1 with respect to each of the research works, culminating into a discussion of how all the authors view the world in terms of a group's talent distribution.

Table of Contents

1 Introduction

Machine bias is common in several learning applications where the classifier mimics the partial behaviour in the data used in training. With the recent studies proving the vulnerability of these classifiers to the same biases as that of humans, such as the ProPublica Article[6] demonstrating the bias of Compass (an AI Software tool) against black individuals in predicting recidivism, the research in Fairness has increased exponentially[24] in the last decade.

Given the substantial research in the field, this work will cover a comprehensive review (chapter 3) of the state of the art theoretical fairness research works, which is common across a majority of the papers studied. This section would introduce a beginner reader to the terminology and main methods in the Fairness literature in Machine Learning. We discuss generic concepts like- how to find sensitive attributes, the several kinds of approaches to ensure fairness (such as individual and group fairness). Furthermore, we discuss the fairness policies, which are the most common ways of assessing fairness in a machine learning classifier. We conclude this section with a brief discussion about the fairness intervention approaches in which we could ensure fairness for a classifier and are broadly divided in 3 sub-sections i.e. pre-processing, in-processing and post-processing.

In addition, we propose a generic theoretical model (section 4.2) which defines disadvantage for a particular demographic group. This new model is novel as compared to previous works[2, 17, 20, 24] as we also 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, our first result (chapter 4) will 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 prove 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 4.4 and then transition into Gaussian distribution 4.6. Following that, we discuss a property which we call "Sub-exponential" for a curve for which our claim holds. Then, we discuss a some examples and counter-examples for the "Sub-exponential" curves.

Finally, from chapter 5 onwards we summarize the literature review of seven research works [2, 13, 17, 20, 24, 25] which we believe were the most similar to our work. For each of the research work, we have a brief summary of the model outlined, 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 hold in the individual research works. 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 verify if that is the case with the models of these other research works.

2 Contributions

With the recent growth in research in mitigating bias and promote fairness in classifiers, the area is has become complex and hard to penetrate for newcomers to the domain. Hence, the first contribution (Chapter 3) of thesis seeks to provide an overview of the different schools of thought and approaches to mitigating (social) biases and increase fairness in the Machine Learning literature.

Our second major contribution is the bias model we propose. Our motivation behind proposing a new model was that many research works we looked at had a discriminatory worldview where they considered that the disadvantaged groups are inherently less talented. In this work, we consider a different motivation of taking the support environment available to 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 4.2 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 4) 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.

The next section of this work considers the seven most relevant research work (chapter 5 - 5.8) 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 that does in each of the research works, our Theorem 1 still holds.

Finally, (chapter 6) we analyze the different world view of research works, i.e. how the model discussed view the population distribution of the world. We see that while there are a few models[2, 25], 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. This chapter will discuss and compare the research works with our model and our motivation to create our model in a way as discussed.

3 Related Work & Common Literature

3.1 Related Work

Our proposed model is directly related to Kannan et al.[20], which considers a two staged model of screening decision for college admissions— 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 noisy). 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 [17, 20, 24] have considered separate distributions of scores for the two groups, with the disadvantaged group having lower mean 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.

One of the more interesting model was proposed by Friedler et al. [11], which has 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 us. 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 Decision space.

Several recent research works also compare fairness interventions such as demographic parity (DP) and equal opportunity (EO). DP has been considered in numerous recent fairness papers [3, 32] and was proposed in Dwork et al.[9]. A recent work by Hardt et al. (2016)[16] introduced the concept of equality of opportunity. We discuss about the fairness interventions in the section 3.4.

While we majorly looked at papers which considered Group fairness notions, one of the

papers we came across during our research was Roth et. al.[19], which considered individual fairness notion i.e. that all similar individuals should be treated similarly.

3.2 Defining Sensitive Attributes

Almost all fairness interventions 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. [24] 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 [29], 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.

A few approaches do try to make data anonymize by finding correlation between explicitly sensitive data and other features such as graph and network-based[30] methods for discovering proxies. One of the more common approach is to use causal methods find correlation[5].

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, 21] 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. Thus, causal methods are specifically well suited to identifying proxies of sensitive variables[15] as well as subgroup analyses of which subgroups are most unfairly treated and differentiate the types of bias exhibited.

3.3 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 based on individuals such as every individual should be treated similarly.

Group fairness notions tries to equalize the two demographic groups, such as demographic parity where $Pr(R = +|A = a) = P(R = +|A = b) \quad \forall a, b \in A$. Often in group fairness to ensure equality, the members of the disadvantaged group are given an advantage (affirmative action)[20] which comes at the cost of individual fairness being impossible.

Individual fairness, as its name suggests, is individual-based. It was first proposed in Fairness Through Awareness by Dwork et al.[9] in 2012, which is one of the most important foundational papers in the field. The notion of individual fairness emphasizes on 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[1] and cannot co-exist. There exists several studies which have shown this tension between individual and group fairness[7, 22, 28]. Although during our, we looked at one study on Individual fairness [19], our research is based majorly on Group Fairness.

3.4 Group Fairness Policies and Impossibility Theorem

The most common way to assess faitrness is the 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 is about how to compare the output of the model's classifier and compare its result for the two groups.

There are more than 21 definitions[26] 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 Klienberg et. al[23].

This section will now discuss the fairness interventions 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 interventions are the most relevant to our work and the thesis will discuss each in detail.

3.4.1 Demographic Parity

Proposed by Dwork et al. ([9]), Demographic Parity states that the proportion of each segment of a protected class (such as race) should receive the positive outcome with equal probabilities.

$$P(h(x) = 1|x \in A) = P(h(x) = 1|x \in B) \tag{3.1}$$

where $\{A, B\}$ represent the group membership and $h(x)$ represents a binary classifier's output function.

3.4.2 Equal Opportunity

Proposed by Hardt et. al.(2016) [16] is quite similar to Demographic Parity. 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].

$$P(h(x) = 1|y = 1, x \in A) = P(h(x) = 1|y = 1, x \in B) \tag{3.2}$$

3.4.3 Predictive Parity

This metric 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.

$$\forall \hat{y} \in \{0, 1\} \quad P(y = \hat{y}|h(x) = y, x \in A) = P(h(x) = y|y = \hat{y}, x \in B) \tag{3.3}$$

3.4.4 Equalized Odds

Equalized Odds [16] is a similar notion, also introduced in [24]. In addition to requiring Line 1, 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 .

$$\forall \hat{y} \in \{0, 1\} \quad P(h(x) = \hat{y}|y = \hat{y}, x \in A) = P(h(x) = \hat{y}|y = \hat{y}, x \in B) \tag{3.4}$$

3.4.5 Disparate Impact

Disparate impact was first proposed in the fairness literature by [17] Similar to Demographic Parity, it is the ratio of positive classification rate of two groups.

$$\frac{P(h(x) = 1|x \in A)}{P(h(x) = 1|x \in B)} \tag{3.5}$$

3.4.6 Impossibility Theorem

The Impossibility Theorem 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 for a well calibrated classifier and a sensitive attribute capable of introducing machine bias.

It turns out that any two of the three criteria in 3.4.1, 3.4.2 and 3.4.3 are mutually exclusive except in non-degenerate cases. Consider that G is the group membership and Y is is the "true" label distribution then consider the below:

1. Demographic Parity VS Predictive Parity: If G is dependent of Y, then either Demographic Parity holds or Predictive Rate Parity but not both. Proof:

$$G \not\perp Y \text{ and } G \perp Y|\hat{Y}, \text{ then } G \not\perp \hat{Y}$$

.

2. If G is dependent of Y and $\hat{Y}$ is dependent of Y, then either Demographic Parity holds or Equalized Odds but not both. Proof:

$$\hat{Y} \perp G \text{ and } \hat{Y} \perp G|Y, \text{ then either } G \perp Y \text{ or } \hat{Y} \perp Y$$

.

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. Proof:

$$G \perp \hat{Y}|Y \text{ and } G \perp Y|\hat{Y} \text{ implies } G \perp (\hat{Y}, Y) \text{ which implies } G \perp Y$$

.

3.5 Fairness Interventions Approaches

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

3.5.1 Pre-Processing

This approach targets fixing the bias in Data itself. One of the papers (Jiang et al[18]) 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 suggests that there is often an issue is the data itself, 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 as 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. 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.

3.5.2 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.

In our case specifically, we use the Error Minimization Algorithm — Empirical Risk Minimization (ERM) on which we apply a constraint of fairness intervention while training.

3.5.3 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.

4 Theorem 1

4.1 Introduction

Our model considers that every demographic group was 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 eventually become 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 in the model. 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 that under which scenarios can we guarantee that our argument holds.

4.2 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 $g \in \{A, B\}$.

4.2.1 Construct Space

The construct space consists of two distributions:

4.2.1.1 Talent Distribution T

Our main goal is to determine this Talent ($t \in T$) of an individual (where T is the talent distribution). Since we do not have direct access to this space, we want to approximate it using the *Observed Space* discussed later.

4.2.1.2 Environment Distribution E

Unlike the previous models [20, 24], we also take into account the Environment component, which is a measure of how conducive things are around an individual to promote her success. 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 lower distribution, and the subsequent sections will discuss how specifically the environment distributions are shifted.

The environment E 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 defined below.

4.2.2 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.

4.2.3 Score Distribution

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

$$X = T + E \tag{4.1}$$

In general, the environment scores are lesser for the disadvantaged individuals than the advantaged and hence the feature means are also in general lesser for the disadvantaged group.

4.2.4 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 $\vec{x}$, the function h gives a binary screening decision such as whether to hire a candidate or not.

4.3 Motivation

4.3.1 Difference in Expectation

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 4.2 represents the formal equivalent of our main claim:

$$Exp[\, t \,|\, X{=}x \,\&\, G{=}B] > Exp[\, t \,|\, X{=}x \,\&\, G{=}A] \tag{4.2}$$

Or equivalently for all talent thresholds c,

$$\forall c,\ Pr[\, t \geq c |\, X{=}x \,\&\, G{=}B] > Pr[\, t \geq c |\, X{=}x \,\&\, G{=}A] \tag{4.3}$$

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.

$$Exp[\,t\,|X=x\,\&\,G=B] - Exp[\,t\,|\,X{=}x\,\&\,G{=}B] \tag{4.4}$$

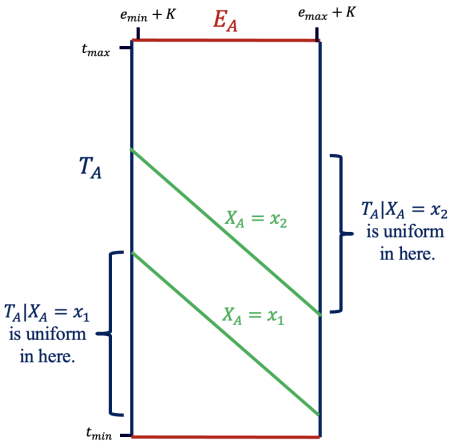
for Uniform (section 4.5), Gaussian (section 4.6) and other 4.9 Distributions of talents and environments. We will show that in the best case the expected difference is positive, and in the worst case it could even fall to negative. We will see that two important issues are whether the range of the Environments E is lower than the Talents T and that the environment has a property that we call sub-exponential, which include Gaussian, Uniform and linear distributions (section 4.12.3).

4.4 Uniform Talent and Environment Distribution: Merging Distributions $\langle T, E_A\rangle$ and $\langle T, E_B\rangle$

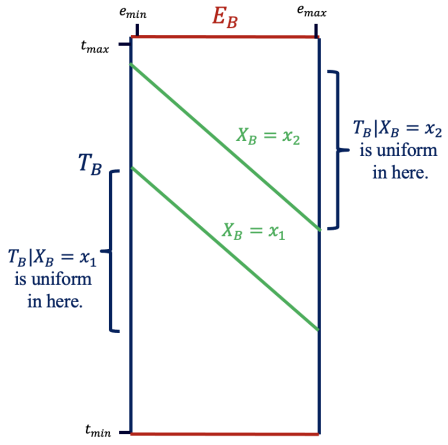
Our first step is to understand the probability spaces for the two groups and for comparison merge them into one. Because it is easier, we will start with all the distributions being uniform. We start with the simple Uniform Distribution of Talent and Environment for a person A and B both receive their talent T from the same uniform distribution $T = \mathcal{U}\,(\,t_{min},\,t_{max}\,)$. The B person receives their environment score E_B from $E_B = \mathcal{U}(\,e_{min},\,e_{max}\,)$ while the A person receives E_A from the shifted distribution $E_A = \mathcal{U}(\,e_{min}+K,\,e_{max}+K\,)$. 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 will represent the full probability space as the $\langle T, E_A\rangle$ vs $\langle T, E_B\rangle$ rectangles in Figure 4.1a and 4.1b. Here talent is on the x-axis and environment on the y. Each tilted green line represents the narrowed probability space when conditioned on the performance score being fixed to $X_A{=}X_B{=}x_i$. 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 and K lower for group A. Because T, E_g, and X_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 $Exp[T_g|X_g{=}x_2] - Exp[T_g|X_g{=}x_1] = x_2{-}x_1$.

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 $E = \mathcal{U}(e_{min}, e_{max})$. Before we advantaged 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



(a) Group A's Distribution



(b) Group B's Distribution

Figure 4.1: Distributions with Low Environment Range

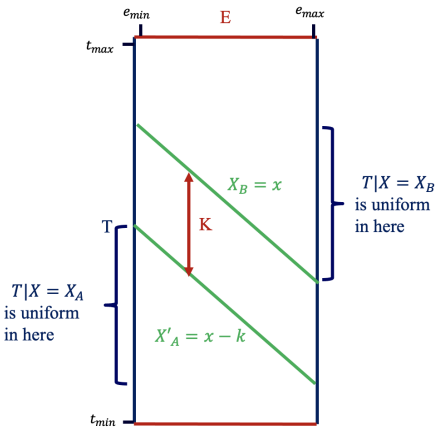


Figure 4.2: Representing X'_A and X_B on the same image

green lines onto the y-axis gives the required result that $Exp[T_B|X_B=x] - Exp[T_A|X_A=x] = K$. The next section will observe that this same result does not occur when the green lines are in the extreme corners.

In the next section, we will also 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 4.3. However unlike the previous figure, here we show two possible values of X which have the same talent values.

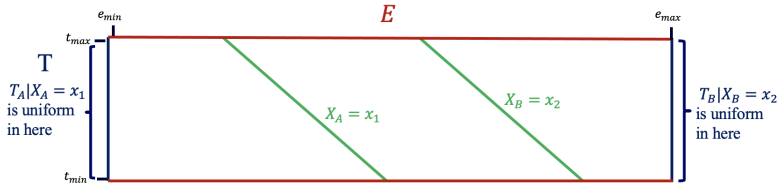


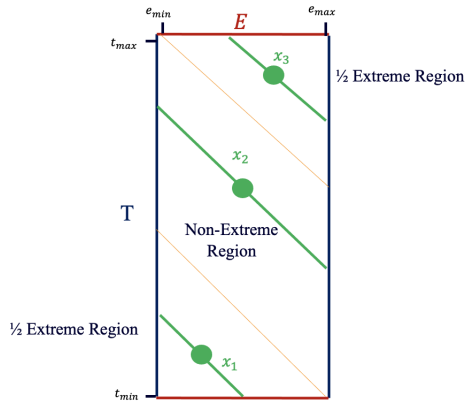
Figure 4.3: Representing X'_A and X_B on the same image

As we can deduce from the figure, that for the pair $x_1 < x_2$, the expected value of talent is *equal* that for x' i.e. $Exp[T \,|\, X=x_2] \geq Exp[T \,|\, X=x_1]$. The next section will observe that this same result does not occur when the green lines are in the extreme corners.

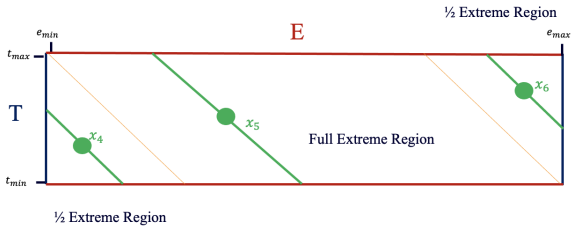
4.5 Extreme x in Uniform Distribution and r-values

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 impossible to estimate the person's talent. 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 $x - e_{max,g}$ is 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 4.4a gives an example of each of the six cases.

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



(a) Narrow Environment Range



(b) Wide Environment Range

Figure 4.4: Extreme Regions: r-value

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 4.3 explains 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

$$Exp[T_B | X_B = x] - Exp[T_A | X_A = x] = \tfrac{1}{2} r(x) K \tag{4.5}$$

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 consider Gausian distributions. Though Gausians have infinite ranges, the mass is concentrated within a few standard diviations. As such, we will show that its effective $r(x)$ value is between 0 and 2 depending on the variences of T, E_A, and E_B.

4.6 Gaussian Distribution Analysis

In this section, we will consider that the Talent and Environment distributions have a Gaussian Distributions and then outline the main claim that given 2 individuals have similar scores, does the individual belonging to the disadvantaged group have a higher expected talent. Hence, we consider that the talent distribution T is Normally distributed and have the same mean and variance across the two groups, i.e. if A stands for the advantaged group and B for the disadvantaged, then

$$T_A = T_B = \mathcal{N}(\mu_t, \sigma_t^2)$$

Similar to the talent distribution, we consider environment distribution to be Normal, however to model disadvantage, we assume that the environments scores of the disadvantaged have a lower mean. This is demonstrated in assumption 1 below.

$$E_A = \mathcal{N}(\mu_{E_A}, \sigma_E^2) \text{ and } E_B = \mathcal{N}(\mu_{E_B}, \sigma_E^2) \tag{4.6}$$

To model disadvantage in the environment, we assume that for Gaussian Distribution,

$$\mu_{E_B} < \mu_{E_A} \tag{4.7}$$

This assumption is realistic and could be proved using real-world datasets. In the dataset[24], we could see that the black individuals in general have a distribution with lower Credit History scores than the white individuals and the dataset follows a Gaussian distribution.

Let $K = \mu_A - \mu_B$. As formalized and proved in section 4.11, the following is true for the above model 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

$$Exp(t_b - t_a | x_b = x_a) = \frac{K}{\sqrt{2}(1 + \frac{\sigma_E^2}{\sigma_T^2})} \tag{4.8}$$

where r is $Exp(E_B - E_A)$, $t_a, \in T$, $t_b \in T$ and $x_a \in X_A$ and $x_b \in X_B$.

Next, we will analyze that what is the impact of the Environment Variance on the overall difference of the expected value of talents. More specifically, what if the environment variance is much larger or much smaller than the talent variance.

4.6.1 Expected Talent Difference between Groups

Let's analyze the expected value derivation with respect to small and large values of the environment variance. From equation 4.8, we know that:

$$Exp(t_b - t_a | x_b = x_a) = Exp(E_A - E_B) * \frac{\sigma_T^2}{\sqrt{2}(\sigma_E^2 + \sigma_T^2)} \tag{4.9}$$

Similar to the analogy we discussed in the section 4.5, where we considered a new variable r which defined the number of intersections with the talent axis, in the case of Gaussians we will also consider a similar analogy.

Considering that $Exp(E_A - E_B) = K$ and we consider r in the case of Gaussians to be $\sqrt{2} * \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$, then with the same equation 4.5 to find the expected difference in talents (i.e. $r * \frac{K}{2}$) as discussed in section 4.5, we could find the difference in the expected values of talents.

We will now consider two cases, first is when the environment distribution has a much higher variance in comparison to variance of Talent Distribution and vice versa.

4.6.1.1 Small Environment Variance

Suppose that the Environment Variance $\sigma_E^2 \ll \sigma_T^2$, then the fraction $\frac{\sigma_E^2}{\sigma_T^2}$ in the equation 4.9 will be very small. This is demonstrated in figure 4.5, the score distribution closely reflects the variance of Talent Distribution and we learn about talents accurately.

If such is the case, then the overall expectation can now be approximately represented as:

$$Exp(t_b - t_a | x_b - x_a = 0) \approx \frac{Exp(E_A - E_B)}{\sqrt{2}} = \frac{K}{\sqrt{2}} \tag{4.10}$$

(a) Talent

(b) Environment (Low Variance)

Figure 4.5: Low Environment Variance compared to Talent Variance

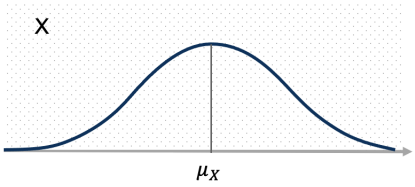


Figure 4.6: Score Distribution

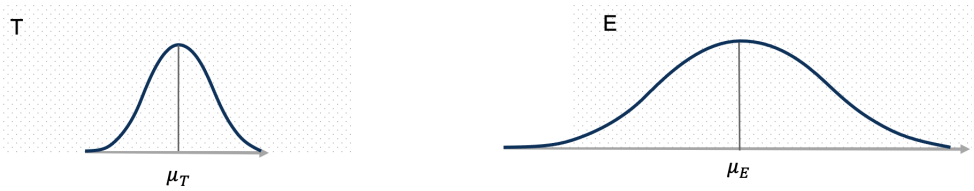
This value is also the maximum possible difference in the expected talent within Group B and Group A. That is in the best case when there is very low variance in the environment distribution E, the maximum difference in the talent distribution you could expect is $\frac{Exp(E_A - E_B)}{\sqrt{2}}$.

4.6.1.2 Large Environment Variance

On the contrary, let's now consider that the Environment Variance $\sigma_E^2 \gg \sigma_T^2$, then the fraction $\frac{\sigma_E^2}{\sigma_T^2}$ in the equation 4.9 will will be very large. This is represented in the Figure 4.7 where the score distribution X will very closely represent the variance of Environment and therefore we do not learn much about the talent distribution from the scores.

If the Variance of Environment is very large, then the overall expectation will be:

$$Exp(t_b - t_a | x_b - x_a = 0) \approx 0 \tag{4.11}$$



(a) Talent

(b) Environment (Low Variance)

Figure 4.7: Low Environment Variance compared to Talent Variance

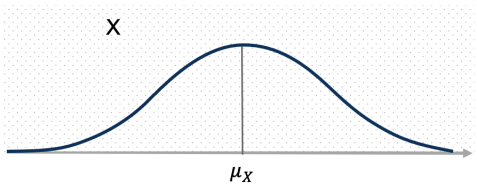


Figure 4.8: Score Distribution

The value $\frac{Exp(E_A - E_B)}{\sqrt{2}}$ being 0 is the minimum possible difference in the expected talent within Group B and Group A. In this case, we could see that when the Environment Noise variance is very high, we effectively learn nothing about the talents for group A and B and hence the expected talent difference is 0.

4.7 Non-Extreme X Values: $r(x) = 2$

Our primary goal is to 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 $e \in E$. 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 non-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.

We say that group A is privileged over group B if their environment distributions are such that when ever $Pr(E_B \geq e_B) = Pr(E_A \geq e_A)$ we have that $e_B \ll e_A$. This says that if the A person received environment value e_A and the B received e_B, 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 $e_B \ll e_A$.

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 is to prove

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

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 pN^{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 $t_B = F(t_A)$ mapping between them and by proving that $t_B > t_A$.

Theorem Here we only consider x_g that are non-extreme performance scores, i.e. $r(x_g) = 2$. Suppose the talent distribution T is uniform. The environment distributions E_A and E_B can be anything. We are assuming the measure of performance is the sum $X_g = T_g + E_g$ of the talent and environment for $g \in \{A, B\}$. It follows that

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

If their environment distributions E_A and E_B are such that group A is privileged over group B, then

$$Pr(T_B \geq t_B | X_B = x_g) = Pr(T_A \geq t_A | X_A = x_g) \;\Rightarrow\; t_B \gg t_A.$$

Suppose further that $E_A = E_B + K$, then $t_B = t_A + K$.
Slightly more generally, if $E_A = d \cdot E_B + k$, then $t_B = t_A + \frac{k}{d} + \frac{d-1}{d} e_A$.

Having a more general performance score computed by $X_g = X(T_g, E_g) = u{\cdot}T_g{+}v{\cdot}E_g{+}x_0$ (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.

Figures 4.21 and 4.9 give an examples of non-uniform distributions in which the results do and do not still hold

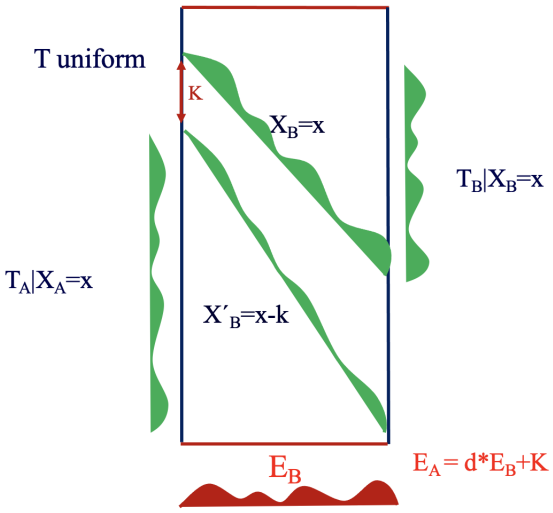


Figure 4.9: The talent distribution T is uniform $E_A = d{\cdot}E_B{+}k$. The green lines are used to project the environment distribution to the talent.

Proof. We will drop the subscript $g{\in}\{A, B\}$, when the statements apply to either group. 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 $Pr(T{\in}[t, t{+}\delta t]) = \delta t{\cdot}P_T(t)$. Similarly, we denote his environment value by e which is drawn from the distribution E with density function $P_E(e)$. Because these two random variables are independent, we can define the cross density function $P(t, e) = P_T(t){\times}P_E(e)$ so that $Pr(T{\in}[t, t{+}\delta t]\ \&\ E{\in}[e, e{+}\delta e]) = \delta t{\cdot}P_T(t){\times}\delta e{\cdot}P_E(t) = \delta t\delta e{\cdot}P(t, e)$. 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 $P(t, e)$.

We will now use the restriction that the talent distribution T is uniform. This means that the density function $P_T(t)$ is constant everywhere in its range and zero elsewhere. In order to be able to ignore ugly constants, Though it is standard to define density functions so that the area under them is one, it is also natural to relax this condition and then to divide by the area when one wants to compute a probability. This allows us to define $P_T(t){=}1$ within the range. We also have the restriction that x is such that $r(x) = 2$. 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. This means that for all values of t that we care about $P_T(t)=1$, giving $P(t, x-t) = P_T(t) \times P_E(e) = P_E(e)$.

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. Because we decided not to care about the area under our density functions, we get the density function $P_x(t) = P(t, x-t) = P_E(x-t)$.

We are now ready to compare the two groups $g \in \{A, B\}$ using this result. Before we can compare this density function $P_{E_g}(x-t_g)$ for the two groups, we need that the area under them is the same. No matter, what the distribution E_A and E_B are, the areas under their density functions $P_{E_g}(e_g)$ are both one. The area under $P_{E_g}(x-t_g)$ will also be one, as long as when one varies over all values of t_g considered, one gets all possible values of e_g. This is the case, because of the restriction that x is such 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. In conclusion,

$$\text{The density function of } Pr(T_g \in [t_g, t_g+\delta t]|X_g=x) \text{ is } P_{E_g}(x-t_g)$$

Due to the linearity of expectation, $Exp(T_g|X_g=x) = x-Exp(E_g)$. The result follows that

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

Recall that our second goal is to define a function

$$F_x(t_A) = t_B \text{ so that } Pr(T_B \geq t_B | X_B = x) = Pr(T_A \geq t_A | X_A = x)$$

$$\text{Or equivalently } Pr(T_B \in [t_B, t_B + \delta t] | X_B = x) = Pr(T_A \in [t_A, t_A + \delta t] | X_A = x)$$

To do this this, it is sufficient to equate their density functions and solve for t_B given a each fixed a value for t_A, namely

$$P_{E_B}(x - t_B) = P_{E_A}(x - t_A)$$

Given that these are also the density functions of this other probability, this says

$$Pr(E_B \in [x - t_B, x - t_B + \delta e]) = Pr(E_A \in [x - t_A, x - t_A + \delta e]).$$

Locally, this does not tell us much. However, because we do this simultaneously for every pair $F_x(t_A) = t_B$, we can integrate and get the global statement

$$Pr(E_B \geq x - t_B) = Pr(E_A \geq x - t_A).$$

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 $t_B \gg t_A$.

In order to be more specific, lets suppose that $E_A = d \cdot E_B + K$, i.e we randomly choose a value E'_A from the distribution E_B and then set $E_A = d \cdot E'_A + K$. Plugging this in gives

$$Pr(E_B \geq x - t_B) = Pr(E_A \geq x - t_A) = Pr(d \cdot E'_A + K \geq x - t_A) = Pr(E'_A \geq d^{-1}(x - t_A - K)).$$

Because E_B and E'_A are drawn from the same distribution, it follows that

$$x - t_B = d^{-1}(x - t_A - K).$$

Solving this gives that

$$F_x(t_A) = t_B = x - \frac{x - t_A - k}{d} = \frac{t_A + K + (d-1)x}{d} = \frac{t_A + K + (d-1)(t_A + e_A)}{d} = t_A + \frac{K}{d} + \frac{d-1}{d} e_A.$$

If further, we set $d = 1$, then we get

$$t_B = t_A + K.$$

□

This section handled the cases in which the talent distribution T is uniform and the values are not extreme. Figures 4.21 and 4.9 give an examples of non-uniform distributions in which the results do and do not still hold. The next section will consider the remaining case.

4.8 Other Distributions

We started with the initial analysis considering Uniform and Gaussian distributions for which our theorem's claim passed. However there exists other more complex distributions for which we can/cannot guarantee that for two individuals with similar scores, the one belonging to the disadvantaged group is more talented. The next two sections will analyze some such distributions in more detail.

4.8.1 Passing Distribution

We first consider a linearly increasing Environment distribution while keeping the Talent Distributions uniform. The figure 4.10 demonstrates that the x-axis represents the linearly increasing Environment Distributions, the y-axis represents the uniform talent distribution and the thickness of the green lines (scores) represents the probability distribution density.

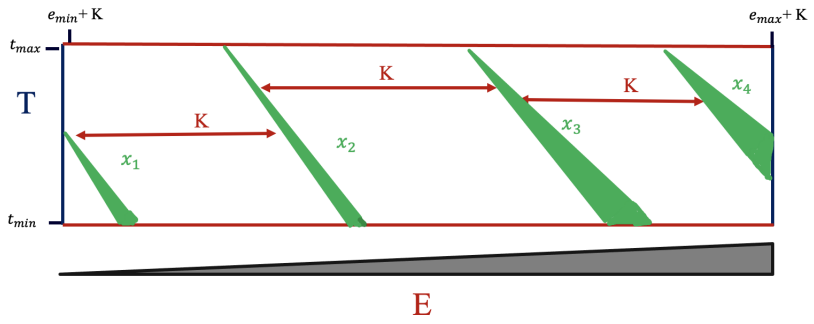


Figure 4.10: Linearly Increasing Environment Distribution

This case is very similar to the figure **??**, the only difference being in the probability densities (thickness) for different x values, which has no effect on our claim. For instance if x_a is x_1 and x_b is x_2, then $Exp[T \mid X = x_B] \gg Exp[T \mid X = x_A]$. On the other hand if x_a is x_2 and x_b is x_3, then $Exp[T \mid X = x_B] = Exp[T \mid X = x_A]$. Hence our claim holds in this case.

Next, we consider a concave function, which increases linearly and then decreases. This is demonstrated in figure 4.11 where the environment PDF is first linearly increasing and then decreasing. Our hypothesis also holds in this case as we can see that for any values of x_1, x_2, x_3 and x_4 the expected values of talent distribution are only increasing.

For instance if x_a is x_1 and x_b is x_2, then $Exp[T \mid X = x_B] \gg Exp[T \mid X = x_A]$ and same is the case when $x_a = x_2$ and $x_b = x_3$, then $Exp[T \mid X = x_B] \gg Exp[T \mid X = x_A]$.

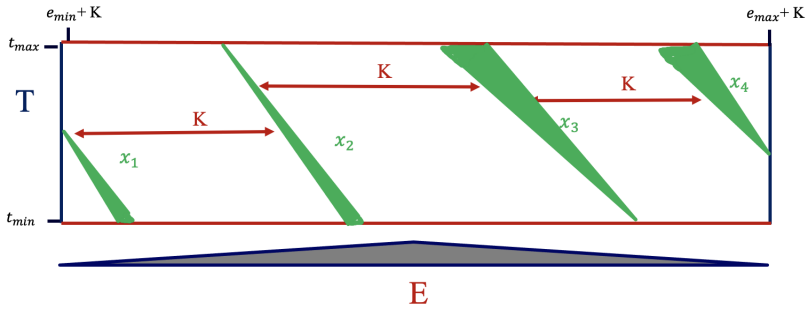


Figure 4.11: Simple Concave Environment Distribution

4.8.2 Failing Distribution

Next we consider an environment distribution for which our theorem's claim fails. Consider the figure 4.12 which shows the environment distribution. We refer to this distribution as "Single Step" function or "Bump" function where the Environment distribution is high in the beginning and then falls steeply as demonstrated.

Figure 4.12: Single Step Distribution

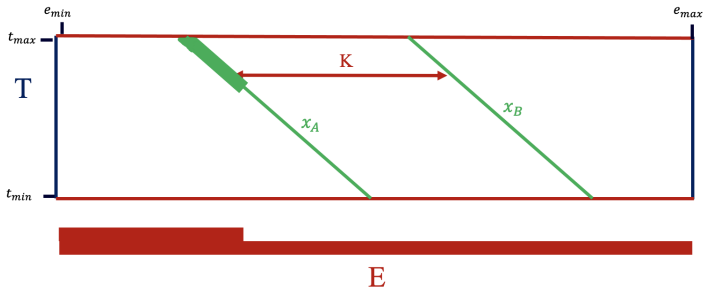


Figure 4.13: Single Step Fall Environment and Uniform Talent

Given such a distribution and Uniform Talent distribution, figure 4.13 and 4.14 represents the overall distribution. 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 4.13, 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 4.14, 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

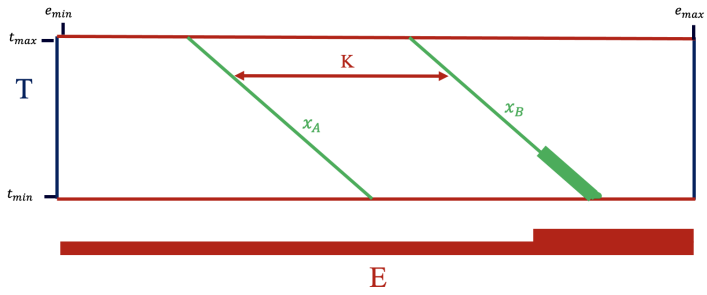


Figure 4.14: Single Step Rise Environment and Uniform Talent

distribution.

Comparison of Single Step to Exponent Function In this part we compare the single step rising function with the exponentially rising function in figure 4.17. We later demonstrate in section 4.9 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.

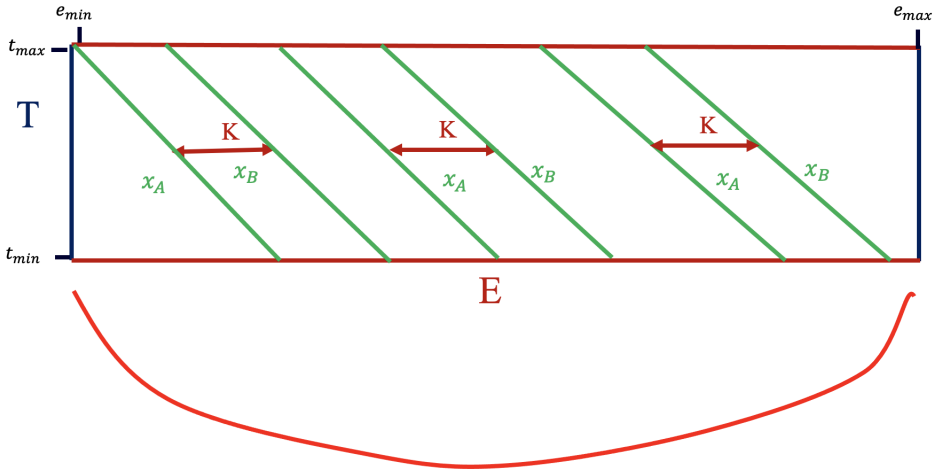


Figure 4.15: A Environment Distribution

There are more interesting distributions for which our main claim fails is suppose that Environment distribution is some convex function as in figure 4.15 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 4.15, 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.

Now that we have discussed the Environment distributions in detail for which our claim fails, we next look at a general case in which we show how our claim would hold for sub-exponential environment distribution.

4.9 Intuition behind Sub-exponential curves and Examples

In lemma 4.12.3, we showed that if the Environment distribution follow a strictly sub-exponential curve, then we can guarantee that the expected value of talents for group B is higher. In this section, we will look into further detail the intuition and idea behind the sub-exponential curves and then look at few examples of sub-exponential and non-sub-exponential curves.

In general, the methodology we discuss in the subsequent sections could be used to figure out whether a given curve is "Sub-exponential" or not.

4.9.1 Broad idea of sub-exponential

In lemma 4.12.3, we defined a function $h(e) = \frac{H_B(e)}{H_A(e)}$ in equation 4.27 and claimed that if the function h is non-decreasing, then the expected talent difference would hold. Here, the functions H_A and H_B are defined as $H_B(t) = E(e - e_B)$ and $H_A(e) = E(e - e_A)$.

Therefore, one could view the functions H_A and H_B as just the environment curves shifted ahead by constants e_A and e_B respectively. Then in order to compare the ratio $\frac{H_B}{H_A}$, consider an example of a Concave function in figure 4.16 where the pair of vertical lines represent H_A and H_B respectively.

Then a curve is considered sub-exponential if the ratio $\frac{H_B}{H_A}$ is falling as we increase our curve. 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. Furthermore, the fall shall not be too steep

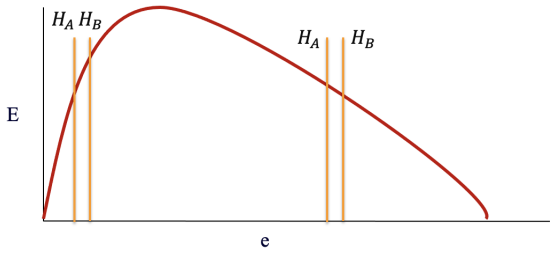


Figure 4.16: An example of Sub-exponential curve

such that it falls faster than an exponent function. Hence, we derived the term "Sub-exponential".

This section covered the intuition behind deducing whether a function is a sub-exponential function or not, in the next section, we will cover more examples.

4.9.2 Examples curves

4.9.2.1 Sub-exponential curves

?? Let's consider a curve which is exponentially increasing as in figure 4.17. For this curve to be sub-exponential, we need the ratio $\frac{H_B}{H_A}$ to be non-decreasing. Hence, 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, the Lemma's claim holds.

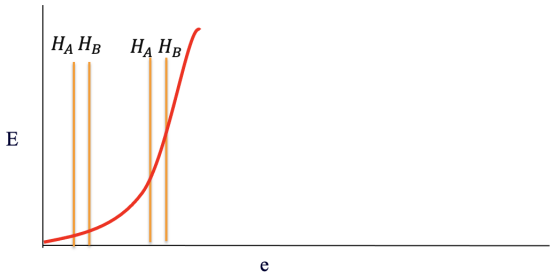


Figure 4.17: Exponentially Increasing

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.

Finally, let's consider the combination of three curves which we have seen so far. First is the concave curve in figure 4.16 for which we have shown that the lemma 4.12.3 would hold. Second,

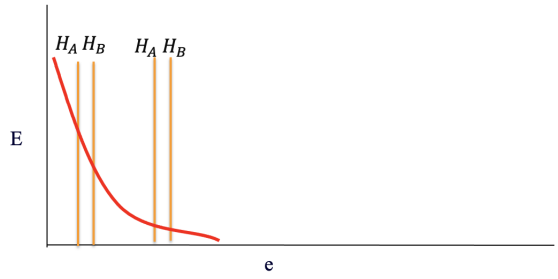


Figure 4.18: Exponentially Decreasing

consider the exponentially increasing and finally the exponentially decreasing. The combination of the three figures is demonstrated in the figure 4.19.

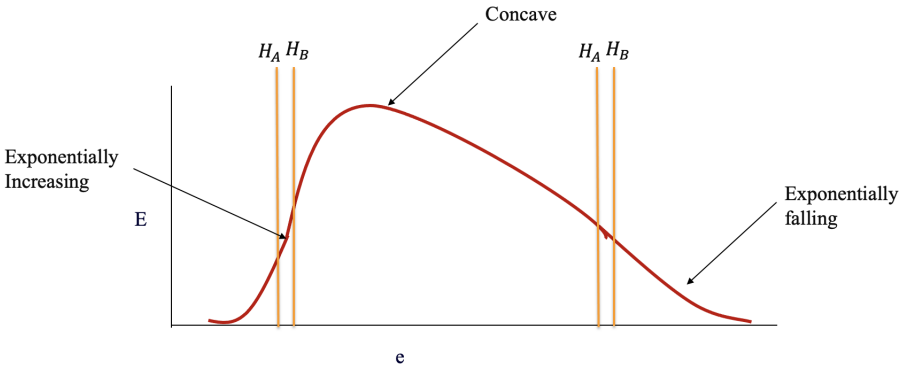


Figure 4.19: Combination Curve

We have shown that for each of the individual curves the lemma would hold, hence lets focus on the transition between the curves 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.

4.9.2.2 Non-Sub-exponential curves

Let's not consider an example of a function which is not Sub-exponential. The figure 4.20 has a concave curve which ends in Uniform edges.

Here, we can clearly deduce that the lemma doesn't hold. For the two pairs highlighted, the fraction H_B/H_A will increase. In the first pair, H_A is still in uniform while H_B begins to rise. Similarly, in the second pair H_B enters the uniform while H_A begins to fall. This makes the ratio H_B/H_A increase and hence the lemma fails to hold.

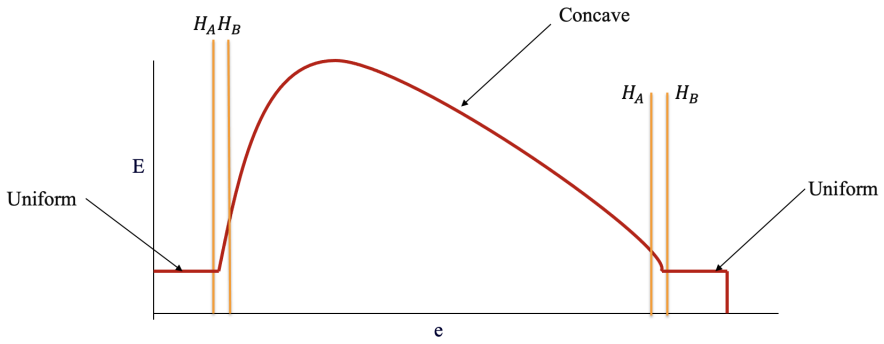


Figure 4.20: Concave Ending in Uniform Functions

4.10 Different Talent Distributions

So far, we have majorly talked about the different Environment distributions while considering the Talent distributions to be either uniform (in section **??**) or Gaussian (in section 4.11). Now we will consider the cases with more complex Talent Distributions.

Our claim is that for the "worst-case" talent distributions (i.e. non-smooth), the theorem's claim will still hold. One intuition behind this claim is that while environment is effectively the noise and its worst case distributions makes the learning problem more difficult, talent distributions is something that we intend to learn for an individual, and hence worst case distributions will still work.

In the subsequent sections, we will discuss three specific cases with different talent distributions and show that the inequality $Exp[t \mid x \,\&\, B] \geq Exp[t \mid x \,\&\, A]$ holds.

4.10.1 Linearly Increasing Talent Distribution

Initially, let us consider that the Talent is linearly increasing. The figure 4.21 illustrates this talent distribution on the y-axis and environment on the x-axis. Similar to the reasoning discussed in section 4.4 we have four scores x_1, x_2, x_3 and x_4. For our main theorem, let's condition on $X_A = X_B = x$. In order to show and compare the two scores on the same figure, let us consider a

new distribution $X'_B = X_B + K$, such that now $X_A = X'_B$. The earlier condition $X_A = X_B = x$ is equivalent to $X_A = x_A$ and $X'_B = x_B = x_A + K$ and it is represented in the figure 4.21. As we know that $x_B = x_A + K$, therefore if x_1 represents x_a, then x_2 will represent x_b and so on.

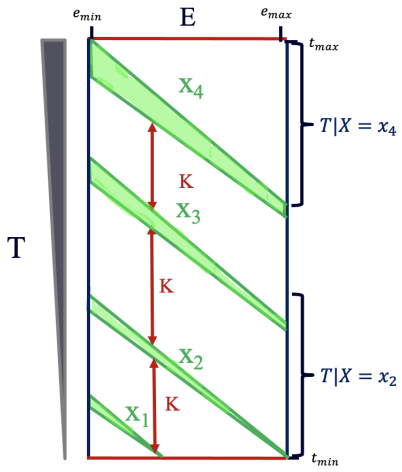


Figure 4.21: Linearly Increasing Talent Distribution

The thickness of the x_i values in the figure represent the probability density for that particular x_i. As is clearly evident, for any increase in x, both the expected value of talent and the probability density increases. Since $x_B = x_A + K$, we can claim that $Exp[t \mid x \,\&\, B] > Exp[t \mid x \,\&\, A]$. (Note: in this case the expected value of talents for group B is strictly greater since we considered low environment range.)

4.10.2 Talent Distribution with a Single Bump

In this section, we consider a talent distribution with a single bump similar to the Bump distribution for environment discussed in the section 4.8.2. While our theorem doesn't hold with bumps in the environment distribution , however the theorem holds perfectly with bumps in the talent distribution.

Consider the figure 4.22 where there is a sudden rise in the talent distribution, while the environment distribution remains uniform. It is evident that for smaller x values (like x_1), the expected value of talent is also small and for larger values (like x_4) the expected value of talent is large. The talent values are the same for medium x values such as x_2 and x_3. Hence, we conclude for this case that $Exp[t \mid x \,\&\, B] \geq Exp[t \mid x \,\&\, A]$.

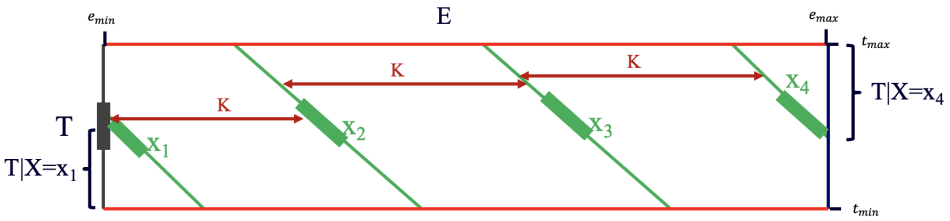


Figure 4.22: Single bump Talent Distribution

4.10.3 Talent Distribution with a Two Bump

Finally, suppose that the talent distributions have two bumps while the environment is uniform as demonstrated in fig 4.23. This case again is very similar to section 4.4 where we discussed Uniform T with small environment distribution. We achieve strict inequality in our theorem in this case, i.e. $Exp[t \mid x \,\&\, B] \gg Exp[t \mid x \,\&\, A]$.

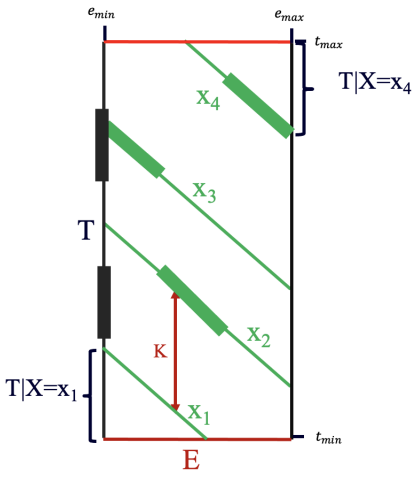


Figure 4.23: Two Bumps Talent Distribution

The figure shows that for any $x' > x$ the expected value of talent is higher for x'. To conclude, talent distribution does not have an impact on the overall finding of our theorem and the inequality $Exp[t \mid x \,\&\, B] \gg Exp[t \mid x \,\&\, A]$ will hold given "Sub-exponential" environment distribution.

4.11 Formalizing Gaussian Distribution Case

In this section, we will formally discuss the case when the Talent and Environments have Gaussian Distributions. We will outline a Lemma to support our claim, and then prove the claim itself.

More specifically, Considering that the *environments* and *talents* for Group A and B are Normally distributed such that $E_A \in \mathcal{N}(\mu_{E_A}, \sigma_E^2)$, $E_B \in \mathcal{N}(\mu_{E_B}, \sigma_E^2)$, $T_A \in \mathcal{N}(\mu_{T_A}, \sigma_T^2)$, $T_B \in \mathcal{N}(\mu_{T_B}, \sigma_T^2)$ and $\mu_{T_A} = \mu_{T_B}$ $(T_A = T_B)$. 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

$$Exp(t_b - t_a | x_b - x_a = 0) = \frac{|r|}{\sqrt{2}(1 + \frac{\sigma_E^2}{\sigma_T^2})} \tag{4.12}$$

where r is $Exp(E_B - E_A)$, $t_a, \in T$, $t_b \in T$ and $x_a \in X_A$ and $x_b \in X_B$.

In order to support our claim for our upcoming theorem for Gaussian Distributions, we consider a Lemma to support our hypothesis.

4.11.1 Lemma: Linear Sum of Mutually Independent Normal Random Variables

[27] If $X_1, X_2 ..... X_n$ are mutually independent normal random variables with means $\mu_1, \mu_2, ... \mu_n$ and variances $\sigma_1, \sigma_2 .... \sigma_n$, then the linear combination:

$$Y = \sum_{i=1}^{n} c_i X_i$$

follows the normal distribution:

$$\mathcal{N}\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

Lemma 2. $\forall c$, $Pr[t \geq c | x \,\&\, G{=}B] > Pr[t \geq c | x \,\&\, G{=}A] \equiv Exp[t | x \,\&\, G{=}B] > Exp[t | x \,\&\, G{=}A]$.

Proof of Lemma 2. $Exp[t|x \,\&\, G{=}B] = \int_{t \geq 0} Pr[t|x \,\&\, G{=}B] t \delta t$
$= \int_{t \geq 0} \int_{c \in [0,t]} Pr[t|x \,\&\, G{=}B] \delta c \delta t$
$= \int_{c \geq 0} \int_{t \geq c} Pr[t|x \,\&\, G{=}B] \delta t \delta c$
$= \int_{c \geq 0} Pr[t \geq c|x \,\&\, G{=}B] \delta c$
$> \int_{c \geq 0} Pr[t \geq c|x \,\&\, G{=}A] \delta c$
$= Exp[t|g \,\&\, 1]$.

4.11.2 Proof for Gaussian Distributions

Considering that the *environments* and *talents* for Group A and B are Normally distributed 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. $T_A = T_B$). 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

$$\mathbb{E}(t_b - t_a | x_b - x_a = 0) = \frac{K}{\sqrt{2}(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})}$$

where K is $\mathbb{E}(E_A - E_B)$, $t_a, \in T$, $t_b \in T$ and $x_a \in X_A$ and $x_b \in X_B$.

Proof: Let Y be the distribution of the talent difference between the two groups A and B such that:

$$Y = T_B - T_A \tag{4.13}$$

Let r be a constant which is defined as the difference between the expected value of E_B and E_A, then from assumption 1, we deduce that $r < 0$.

$$K = \mathbb{E}(E_A) - \mathbb{E}(E_B) > 0 \tag{4.14}$$

Let $y \in Y$ represent the difference in the talent distributions, then we are interested in finding the probability:

$$Pr(Y = y \mid X_B - X_A = 0)$$

Applying Bayes Theorem, the above probability could be written as:

$$Pr(Y = y \mid X_B - X_A = 0) = \frac{Pr(X_B - X_A = 0 \mid Y = y) * Pr(Y = y)}{Pr(X_B - X_A = 0)} \tag{4.15}$$

To determine the individual probabilities on the RHS of **??**, lets first consider $Pr(Y = y)$. The distribution of $Y = T_B - T_A = \mathcal{N}\big(0, 2\sigma^2_T\big)$. Therefore $Pr(Y = y)$ could be stated as:

$$Pr(Y = y) = \frac{1}{2\sqrt{\pi}\sigma_T} e^{-\frac{y^2}{4\sigma^2_T}} \tag{4.16}$$

Similarly to find the probability $Pr(X_B - X_A = 0)$, 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 that:

$$Pr(X_B - X_A = 0) = \frac{1}{\sqrt{2\pi}\sqrt{2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B}}} e^{-\frac{K^2}{2(2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B})}} \tag{4.17}$$

Finally, to find the third probability $Pr(X_B - X_A = 0 \mid Y = y)$, distribution of $X_B - X_A = 0 \mid T_B - T_A = y$

$$X_B - X_A = (T_B - T_A) + (E_B - E_A)$$

We know from 4.13 and 4.14 that:

$$T_B - T_A = y \text{ and } \mathbb{E}(E_B - E_A) = -K$$

Therefore we can induce,

$$\mathbb{E}(X_B - X_A) = \mathbb{E}(T_B - T_A) + \mathbb{E}(E_B - E_A) = y - K$$

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 distribution being equal to y.

$$\sigma^2_{X_B - X_A} = \sigma^2_{E_A} + \sigma^2_{E_B}$$

Finally, the distribution of $(X_B - X_A \mid Y = y)$ could be written as:$(X_B - X_A \mid Y = y) = \mathcal{N}\big(y - K, \sigma^2_{E_A} + \sigma^2_{E_B}\big)$

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

$$Pr(X_B - X_A = 0 | Y = y) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2_{E_A} + \sigma^2_{E_B}}} e^{-\frac{(y-K)^2}{2(\sigma^2_{E_A} + \sigma^2_{E_B})}} \tag{4.18}$$

Substituting Equation 4.16, 4.17, 4.18 into 4.15, we conclude:

$$Pr(Y = y \mid X_B - X_A = 0) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2_{E_A} + \sigma^2_{E_B}}} e^{-\frac{(y-K)^2}{2(\sigma^2_{E_A} + \sigma^2_{E_B})}} * \frac{1}{2\sqrt{\pi}\sigma_T} e^{-\frac{y^2}{4\sigma^2_T}}}{\frac{1}{\sqrt{2\pi}\sqrt{2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B}}} e^{-\frac{K^2}{2(2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B})}}}$$

The above expression can be simplified to:

$$Pr(Y = y \mid X_B - X_A = 0) = \frac{1}{2\sqrt{\pi}\sigma_A} e^{-\frac{1}{2\sigma^2_A}\left(y + \frac{-K}{(\sigma^2_{E_A} + \sigma^2_{E_B})/\sqrt{2}\sigma^2_A}\right)^2} \tag{4.19}$$

where $\sigma_A = \frac{\sigma_T \sqrt{\sigma^2_{E_A} + \sigma^2_{E_B}}}{\sqrt{2\sigma^2_T + \sigma^2_{E_A} + \sigma^2_{E_B}}}$

Equation 4.19 could now be compared to the Probability Density Function of a Normal distribution with mean and variance:

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

Therefore the Distribution is equivalent to:

$$\big(T_B - T_A \mid X_B - X_A = 0\big) = \mathcal{N}\Big(\frac{K}{\sqrt{2}(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})}, \frac{\sqrt{2}\sigma^2_T(\sigma^2_{E_A}+\sigma^2_{E_B})}{\sqrt{2\sigma^2_T+\sigma^2_{E_A}+\sigma^2_{E_B}}}\Big) \tag{4.20}$$

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

To conclude, if $t_a, \in T_A$, $t_b \in T_B$, $x_a \in X_A$ and $x_b \in X_B$, then

$$\mathbb{E}(t_b - t_a | x_b - x_a = 0) = \frac{K}{\sqrt{2}(1+\frac{\sigma^2_{E_A}+\sigma^2_{E_B}}{2\sigma^2_T})}$$

4.12 Extreme X Values: $r(x) = 0$

In this section, we claim that our theorem in general holds for any environment distribution which is "sub-exponential" i.e. any distribution which does not increase at a rate faster than an exponential function at any instance. For example, the two bumps in figure **??** will not fall in this category and hence our hypothesis didn't hold. On the other hand, Uniform, linear and Gaussians are a few examples which are sub-exponential.

4.12.1 Model

Consider some individual, who has a talent $t \in T$ randomly chosen from some fixed distribution with $Pr[t] = T(t)$, which is same for both the Groups. Similar to the model discussed, consider that the group membership is represented by G such that if $G = B$, the individual belongs to the disadvantaged group, else if $G = A$ then the advantaged group.

Suppose that the environment distributions for each of the groups is represented E_A and E_B such that E is our random error variable. Let's represent the probability that the Group B's Environment is u be $e(u)$ i.e. $Pr[E = u] = e(u)$. We require $e(u)$ to be strictly "sub-exponential" in a range $[a, b]$ and zero outside this range. Wolfram alpha shows that sub-exponential includes distributions such as Gaussians, quadratics, polynomials and uniform. The environment distribution for Groups B and A are defined such that $E_B = E$ and $E_A = E + K$, where K is a positive constant. Hence the environment distributions are also "sub-exponential" and Group A's environment shifted ahead by K. Group A is therefore advantaged.

The scores distributions $X = (X_A, X_B)$ are then calculated such that $X_A = T_A + E_A$ and $X_B = T_B + E_B$ and our main goal is to show that if there are two individuals – one in the disadvantaged and other advantaged – have the same scores, the disadvantaged individual is expected be more talented, given the above model and assumption.

4.12.2 Proof

Let's condition on the score $x \in X$ such that this score is the same for two individuals and let's define $t_A = Exp[t|X = x \ \& \ G = A]$ and $t_B = Exp[t|X = x \ \& \ G = A]$ be the expected value of talents for the individuals of group A and B respectively conditioned on score x.

Note that for a disadvantaged person ($G = B$) to get the same score x as an advantaged person ($G = A$), they are expected to have a higher talents , i.e. $t_B > t_A$, or equivalently $t_B = t_A + K$. Because $t_g = t + E_B$, we have that $t = t_g - E_B \in [t_g - a, t_g - b]$, $\forall g \in \{A, B\}$. Let's denote this range as $[t_A^{min}, t_A^{max}]$ and $[t_B^{min}, t_B^{max}]$ for group A and B respectively. Let c denote the threshold for the employer to hire the person.

Now, we will show that for the worst case T and "sub-exponential" environments,

$$Exp[t \,|\, x \,\&\, B] > Exp[t \,|\, x \,\&\, A]$$

Because $t_A \leq t_B$, there are two cases. In the first case, this t interval is disjoint for Group B and Group A, namely $t_A^{min} \leq t_A^{max} \leq t_B^{min} \leq t_B^{max}$. It is trivial to prove the result for this case. Hence, lets assume the second case, namely $t_A^{min} \leq t_B^{min} \leq t_A^{max} \leq t_B^{max}$. Note that abilities outside the range $t \in [t_A^{min}, t_B^{max}]$ are simply not possible given $G{=}g$ and hence will be ignored. Abilities in the range $t \in [t_A^{min}, t_B^{min}]$ only help prove our result because they are low and are possible for Group A but not for Group B. The worst case for the theorem is when the distribution on t simply does not allow such t to arise. Similarly for $t \in [t_A^{max}, t_B^{max}]$. Hence, without loss of generality, lets assume that the distribution on t has range $t \in [t_B^{min}, t_A^{max}]$. If our threshold c is outside this range, then both probabilities in the theorem will either be zero or one. Hence, lets assume $c \in (t_B^{min}, t_A^{max})$.

By definition,

$$Pr[t \ \geq \ c | \, x \,\&\, i] = \frac{Pr[t \ \geq \ c \& \, x \,\&\, i]}{Pr[\, x \,\&\, i]} \tag{4.21}$$

Fix some abilities value t. Let $T(t)dt$ denote the probability of the person having ability t. Because $t_g = t' = t + E$ we have that, $E = t_g - t$ Let's define a function $H_g(t)$ such that:

$$H_g(t) = e(t_g - t) \tag{4.22}$$

This gives $Pr[E = t_i - t] = H_g(t)dt$. In order to solve equation 4.21, we can write that

$$Pr[x\&g] = \int_t T(t)H_g(t)dtdt \tag{4.23}$$

and that

$$Pr[t \geq c \,\&\, x \,\&\, g] = \int_{t \geq c} T(t)H_g(t)dtdt \tag{4.24}$$

Hence, we can use equations 4.24 and 4.23, to substitute in equation 4.21. To simplify notation, let's define a function F, which for any function $r(t)$ gives the the fraction of the area under the curve that is to the right of the $t=c$ is as follows:

$$F(r) = \frac{\int_{t \geq c} r(t)dtdt}{\int_t r(t)dtdt} \tag{4.25}$$

Hence we can write the equation 4.21 as follows:

$$Pr[t \geq c | x \,\&\, g] = F(T \times H_i) \tag{4.26}$$

Now to compare the function H for the two groups A and B, suppose that $h(t) = \frac{H_B(t)}{H_A(t)}$ be the ratio of the H for the groups. Let c' be the constant $h(c)$ and let $h'(t) = h(t)/c'$.

Lemma 2 (4.12.3) proves $h(t_g - t)$ is non-increasing which is equivalent to $h(t)$ is non-decreasing (and strictly increasing for strictly sub-exponential functions).

With the results of the lemma, we conclude that $\forall t \in [t_B^{min}, c)$ we have $h'(t) < 1$ and $\forall t \in (c, t_A^{max}]$ we have $h'(t) > 1$. Hence, multiplying $r(t)$ by $h'(t)$ decreases $r(t)$ for those t before c and increase those after. It follows that the fraction of the area under the curve $r(t)$ that is right of the $t=c$ increases, i.e. $F(r) < F(h'r)$. Similarly, $F(r) = F(c'r)$. The result follows

$$Pr[t \geq c | \, x \,\&\, A] = F(T \times H_A) = F(T \times c' \times H_A) \leq F(T \times c' \times H_B * h') = F(T \times H_B) = Pr[t \geq c | \, x \,\&\, B]$$

This concludes our proof.

4.12.3 Lemma 2

Suppose that $E(u) > 0$ and is "sub exponential" in a range $[a, b]$ and zero outside this range. Let $H_B(t) = E(e - e_B)$ and $H_A(e) = E(e - e_A)$ for $e_A < e_B$. Let

$$h(e) = \frac{H_B(e)}{H_A(e)} \tag{4.27}$$

. Then $h(e)$ is strictly decreasing.

Proof of Lemma 2:

$$h'(t) = \frac{H'_B(e)H_A(e) - H_B(e)H'_A(e)}{H_A^2(e)}.$$

Hence, to prove that $h'(t) < 0$, it is sufficient to prove that $\frac{H_A(e)}{H'_A(e)} < \frac{H_B(e)}{H'_B(e)}$. By the definitions of H_B and H_A, it is sufficient to prove that $\frac{E(e-e_A)}{E'(e-e_A)} > \frac{E(e-e_B)}{E'(e-e_B)}$. Because $e_A < e_B$, it is sufficient to prove that $\frac{E(u)}{E'(u)}$ is strictly increasing.

This can be visualized as follows. In graph of $E(u)$, draw the line though the point $\langle u, E(u)\rangle$ tangent to the curve. Let u_0 be the value of u at where the line crosses the u-axis and $\Delta u(u) = u - u_0$. Note that $E'(u) = \frac{E(u)}{\Delta u(u)}$ and hence $\frac{E(u)}{E'(u)} = \Delta u(u)$. Hence, it is sufficient to prove that $\Delta u(u)$ is strictly decreasing. Now visualize watching $\Delta u(u)$ change as u increases. Note how for concave functions it seem that Δu decreases.

If $r(u) = \frac{E(u)}{E'(u)}$, then $r'(u) = \frac{(E'(u))^2 - E(u)E''(u)}{(E'(u))^2} = 1 - \frac{E(u)E''(u)}{(E'(u))^2}$ To prove that $r'(t) < 0$, it is sufficient to prove that $E''(u) < \frac{(E'(u))^2}{E(u)}$. Suppose this was tight. Then if you tell me $E(0)$ and $E'(0)$, then the whole function is determined. By breaking the domain u into ϵ sized pieces, u effectively becomes an integer and hence we can do induction on it. By way of induction, assume that we have determined $E(u)$ and $E'(u)$. From this we determine $E''(u) = \frac{E'(u))^2}{E(u)}$, $E(u+\epsilon) = E(u) + \epsilon E'(u)$, and $E'(u+\epsilon) = E'(u) + \epsilon E''(u)$. Wolfram alpha gives that $E(u) = c_1 e^{c_2 u}$. Certainly, $h(t) = \frac{H_B(t)}{H_A(t)} = \frac{c_1 e^{c_2(e-e_B)}}{c_1 e^{c_2(e-e_A)}} = e^{c_2(e_A - e_B)}$ is not strictly decreasing but constant as we could expect from things being made tight.

This concludes our lemma.

4.13 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.

5 Related Research Work Review

5.1 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 4). In total, we covered 12 different research works [2, 8, 12, 13, 17, 18, 19, 20, 23, 24, 25, 31]. While we studied each of these works in detail, however in the thesis, we will cover six works of these works, as they are highly correlated with our work of ensuring Group Fairness notions (such as Demographic Parity and Equal Opportunity). Other works such as Roth et al.[19] considered individual fairness, which were impertinent topics as compared to our work and therefore we will not discuss the remaining 6 works in detail. This 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.[20] 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 scores distributions are Gaussian and they use threshold functions as their hypothesis classes. While this work proposes two-staged model, our's discusses a single-staged model. The section 5.2 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. **Disparate Effects of Strategic Manipulation** In this work by Immorlica et. al.[17], the model, scores distributions and the results discussed are very similar to ours. The term "strategic manipulation", analyzes the interaction between a learner and agents in a world where all agents are equally able to manipulate their features in an attempt to "trick" a

published classifier. However, an agent's ability to adapt to an algorithm is not simply a function of her personal interest in receiving a positive classification, but is bound up in a complex web of social factors that affect her ability to pursue certain action responses. This creates a social inequality between the two Groups which is analogous to the Environment Bias in our case. One of the main results of this paper is that Affirmative Action could harm both the groups, although one would suppose that Affirmative Action strictly benefit the disadvantaged. In section 5.3, we discuss the model in detail, outline the assumptions and compare the conclusions.

3. **From Fair Decision Making to Social Equality:** This work by Srebro et. al.[25] is novel in terms of its comparison with the long-term influence of applying fairness intervention 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 such case may help. 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. The section 5.5 will cover this the model and dynamics discussed in the work. The section will also show how the Unconstrained learning is always disadvantaged for Group B.

4. **Delayed Impact of Fair machine learning:** This work (Liu et. al.[24]) 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. There is a discussion that show in one-step feedback model, common fairness criteria in general do not promote improvement over time, and may in fact cause harm in cases where an unconstrained objective would not. This work also considers the delayed impact of three standard criterion similar to our model (i.e. DP, EO and Unconstrained) and also determine their impact on the utility graphs. The section 5.6 will demonstrate the model and a novel tool called *Outcome Curve*, which is helpful in comparison of delayed impact of different fairness constraints.

5. **Recovering from Biased Data:** This work by Blum et. al.[2] discusses in detail about how to extract the Bayes Optimal Classifier, which was very useful in coming up with our second Theorem. This work examines the possibility of extracting the Bayes Optimal Classifier when the Fairness Constraint is applied to the ERM. This paper's has a similar model as compared to our's. 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 environment distribution. In the

section 5.7, 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.

5.2 Research Review 1: Downstream Effects Of Affirmative Action

This research work by Roth et. al.[20], 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 one-staged model, this paper discusses a two-staged model of hiring process:

- Stage 1: Students are admitted to the college on the basis of an entrance exam which is a noisy signal about their qualifications (talent/type).
- Stage 2: Those students who were admitted to college can 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 rational and calculates a posterior distribution on the types/talents of the students conditional on qualifications from Stage 1 and grades from Stage 2. This paper then considers the conditions under which the two fairness definitions can be met or not. Following is the informal definitions of the two fairness criterion discussed in this work:

1. Equal opportunity: 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: Rational employers 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.

5.2.1 Model

This work considers two population distribution 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 $A_i(s) : R \to 0, 1$ 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 $A_i(s)$, where $A_i(s)$ is such that a student is accepted if she is above the β threshold, $S_i \geq \beta_i$.

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:

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

As there are many employers, we consider that the range of possible c values is $C \in [C-, C+]$.

5.2.2 Fairness Definitions

There were two main fairness definitions proposed in this paper:

- Equal Opportunity (EO): holds if and only if the probability of a student being hired by the employer conditional on his type is independent of the student's group. I.e. if for all types t ∈ R,

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

- Irrelevance of Group Membership (IGM): Irrelevance of Group Membership holds if and only if, conditional on admission by the school and on grade g, the employer's decision on whether to hire a student is independent of the student's group. I.e. if for all grades g ∈ 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.

5.2.3 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.

5.2.3.1 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 by only highly selective colleges (i.e. those with very high admissions thresholds) — but 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:

$$Ai(s) = 1 \Leftrightarrow s \geq C+ \tag{5.1}$$

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

5.2.3.2 The Single Employer Threshold Case

Let's assumt 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

$$Exp[T1|G1 = g^*, S1 \geq \beta_1^*] = Exp[T2|G2 = g^*, S2 \geq \beta_2^*] \tag{5.2}$$

5.2.3.3 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.

5.2.4 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 $\forall c$, $Pr[t \geq c|\, x \& B] > Pr[t \geq c|\, x \,\&\, A]$.

This is comparable to the definition of IGM in this work although there is a small caveat, while IGM says that both groups are expected to be equally talented given a threshold C, we say that the disadvantaged is supposed to be more Talented. In addition, the IGM considers only 1 single employer threshold c in its equation, while we consider for every value of scores X.

IGM definition says that given the same grades, is the expected talent value greater than C for both the groups. $\mathbb{E}[T_1|G_1 = g \& A_1 = 1] \geq C \iff \mathbb{E}[T_2|G_2 = g \& A_2 = 1] \geq C$. As we can see, the grade values are the same for both the groups, which is also the assumption in our theorem 1, that the scores of two group's individuals are the same. As as we discussed, IGM could be guaranteed in the case of Noiseless exam scores as well as in the case when we have only a single threshold.

Therefore in our one-staged model, we have a stronger fairness definition where we show that the disadvantaged candidate is more talented. On the other hand, this work shows that both individuals are equally talented in the two staged model for a given threshold.

5.3 Research Review 2: The Disparate Effects of Strategic Manipulation

5.3.1 Introduction

The term "strategic manipulation" in this paper[17] represents the agent's reactivity to a classifier, specifically in the case of college admissions where the agents are the students who manipulate their feature vectors by using the SAT test prep services in an attempt to "trick" the SAT exam. This paper then discusses that this ability of manipulation of features by the agents is not equitably distributed across the groups, but depends on complex web of social factors i.e. in this setting of social inequality, candidate groups face different costs to manipulation.

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 show cases where the subsidy hurts both the groups.

5.3.2 Model and Notion

To briefly state the model, consider that the candidates have innate set of features $x \in X = [0,1]^d$ 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 x to a feature vector y must pay a cost of $c_m(y) - c_m(x)$ where $y \geq x$. To model disadvantage, the study assumes that

$$c_A(y) - c_A(x) \leq c_B(y) - c_B(x) \tag{5.3}$$

i.e. Group A members pay a lower cost than Group B.

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

$$h_A(x) = \begin{cases} 1 & \forall x \text{ such that } \sum_1^d w_{A,i} x_i \geq \tau_A \\ 0 & \forall x \text{ such that } \sum_1^d w_{A,i} x_i < \tau_A \end{cases} \tag{5.4}$$

$$h_B(x) = \begin{cases} 1 & \forall x \text{ such that } \sum_1^d w_{B,i} x_i \geq \tau_B \\ 0 & \forall x \text{ such that } \sum_1^d w_{A,i} x_i < \tau_A \end{cases} \tag{5.5}$$

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[4] show that the scores are skewed for the demographic groups, with disadvantaged having a lower threshold i.e. $\tau_B < \tau_A$.

Given the above setup, Strategic Classification Game with Groups defines that: A candidate from group m pays cost $c_m(y) - c_m(x)$ to move from her original features x to $y \geq x$. There exist true binary classifiers h_A and h_B, for candidates of each group. The true functions upon manipulation are as follows:

$$h'_A(x) = \begin{cases} 1 & \forall y \text{ such that } \sum_1^d w_{A,i} y_i \geq \sigma_A \\ 0 & \forall y \text{ such that } \sum_1^d w_{A,i} y_i < \sigma_A \end{cases} \tag{5.6}$$

$$h'_B(x) = \begin{cases} 1 & \forall y \text{ such that } \sum_1^d w_{B,i} y_i \geq \sigma_B \\ 0 & \forall y \text{ such that } \sum_1^d w_{A,i} y_i < \sigma_A \end{cases} \tag{5.7}$$

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

$$C_{FP} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m} [h_m(x) = 0, f(y) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m \Pr_{x \sim D_m} [h_m(x) = 1, f(y) = 0] \tag{5.8}$$

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

5.3.3 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: 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:

$$f(y) = \begin{cases} 1 & \forall y \geq \sigma \\ 0 & \forall y < \sigma \end{cases} \tag{5.9}$$

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

This strategy is enacted by considering candidates' best-response manipulations. The learner would like to guard against manipulations by candidates with $x < \tau_A$ but still admit candidates with $x \geq \tau_A$, so she considers the maximum manipulated feature y that is attainable by a rational candidate with $x = \tau_A$ who is willing to spend up to a cost of *one* in order to secure a better classification, as in Figure 5.1.

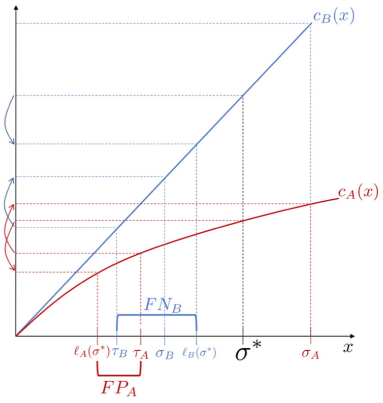


Figure 5.1: Group cost functions for a one-dimensional feature x. τ_A and τ_B signify true thresholds on unmanipulated features for group A and B, but a learner must issue a classifier on manipulated features. The threshold σ_A perfectly classifies group A candidates; σ_B perfectly classifies group B 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).

Hence we deduce theorem 1:

Theorem 1 (d-D Space of Dominant Learner Strategies). In the general d-dimensional Strategic Classification Game with linear costs, there exists a classifier that perfectly classifies group A and a classifier that perfectly classifies group B. All undominated classifiers commit no false negative errors on group A and no false positive errors on group B.

5.3.4 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 5.8 will then

become:

$$C_{FP} \sum_{m\in\{A,B\}} p_m \Pr_{x\sim D_m}[h_m(x)=0, f(y)=1] + C_{FN} \sum_{m\in\{A,B\}} p_m \Pr_{x\sim D_m}[h_m(x)=1, f(y)=0] + \lambda * cost(f,\beta) \tag{5.10}$$

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

Theorem 2 (Subsidies can harm both groups). There exist cost functions c_A and c_B satisfying the cost conditions, learner distributions D_A and D_B, true classifiers with threshold τ_A and τ_B, population proportions p_A and p_B , and learner penalty parameters C_{FN} , C_{FP} , and λ, such that no candidate in either group has higher payoff at the equilibrium of the Strategic Classification Game with proportional subsidies compared with the equilibrium of the Strategic Classification Game with no subsidies, and some candidates from both group A and group B are strictly worse off.

5.3.5 Comparison with our work

5.3.5.1 Comparison with our theorem 1

The model discussed in this work conforms to our Theorem 14. As discussed in section 5.3.2, 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

$$h_A(x) = 1 \implies h_B(x) = 1 \tag{5.11}$$

Using 5.11 we can interpret that for one particular value of x, if the individual belongs to Group B, he is expected to be more talented.

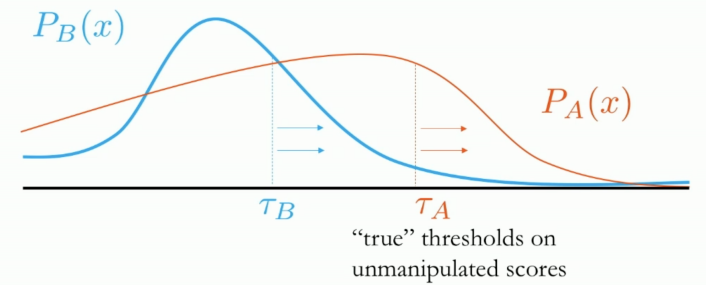


Figure 5.2: Distribution for D_A and D_B with true thresholds τ_A and τ_B

This could also be analyzed graphically as in the figure 5.2, 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 4.

5.4 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 functions and shows two results. First, is that a simplified 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.

5.4.1 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:

$$\frac{\mu(x, A)}{\mu(x, D)} \geq \frac{\mu(x', A)}{\mu(x', D)} \tag{5.12}$$

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 5.4.2 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 5.4.2 will cover one example to illustrate this.

5.4.2 Results

5.4.2.1 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 5.1. 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.

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	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 5.1: Table showing the features and f-values

In the table 5.1, and suppose that the true criterion is the conjunction (dot-product) of $x^{\langle 1\rangle}$ and $x^{\langle 2\rangle}$. To support likelihood condition5.12, 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 5.1. 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 5.2. For all the selection rates $r \leq \frac{5}{18}$ D's representation increases from $\frac{1}{5}$ in table 5.1 to $\frac{1}{3}$ in table 5.2 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 5.3, 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 5.1 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 5.3 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 5.2: Simplification: Not considering $x^{\langle 2 \rangle}$ and γ

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	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 5.3: Simplification: Not considering only $x^{\langle 2 \rangle}$

5.4.2.2 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 5.2 and 5.4, then table 5.3 could be re-written as 5.4 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 5.4: Not Pareto Optimal

5.4.3 General Theorem

The general theorem covers the two results discussed in the section 5.4.2. 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 5.12 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.

5.4.4 Comparison with our Theorem 1

We now compare the general theorem discussed in this paper with our Theorem 1 (section 4) which informally states: Given that the groups have same talent distributions and consider two individuals who have the same scores, the 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 section5.4.2.1 table 5.2.

To conclude, our finding suggests the opposite of the first result of the *Simplicity Creates Inequity paper* 5.4.2 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.

5.4.4.1 Conclusion

Our 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 our's. 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.

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

5.5.1 Introduction

Similar to Liu et al.[24], 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[16, 24] 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.

5.5.2 Model

The model is similar to previous works[16, 20, 24] 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: $\beta(V = v; G = j) = \tau(V = v; G = j).\pi(V = v|G = j)$

5.5.3 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.

5.5.4 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^-)

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

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.

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

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

Assumption 2: The dynamics under 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 $\forall \pi, \pi' \in [0, 1]$:

$$|f(\pi) - f(\pi')| \leq L_{UN} |\pi - \pi'| \tag{5.13}$$

5.5.5 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 5.3

Theorem 1: If equality 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 as demonstrated in the figure 5.4.

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.

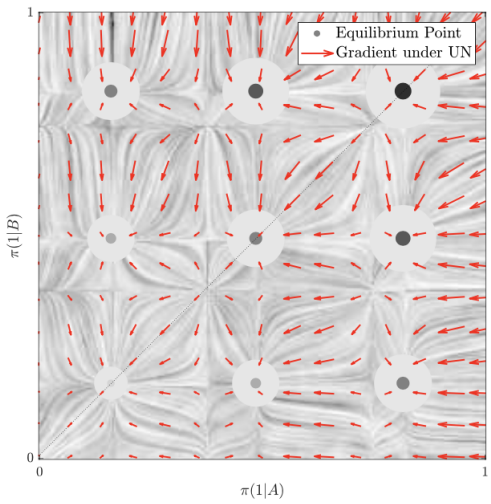


Figure 5.3: Points showing unconstrained equilibriums

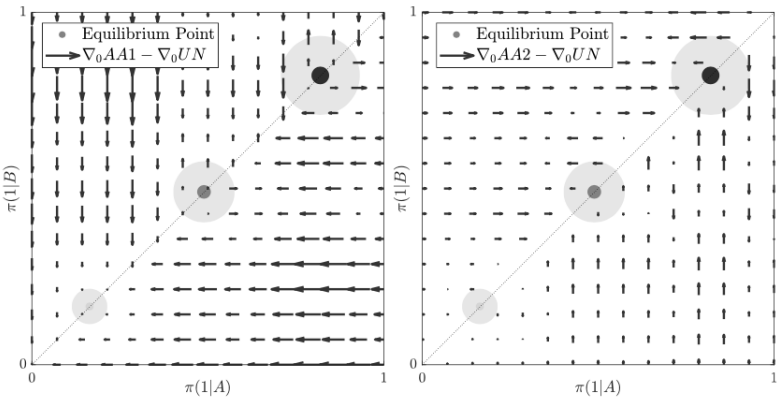


Figure 5.4: Equilibrium points for AA^- and AA^+ respectively

5.5.6 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.

5.5.7 Comparison with our final results

Our theorem 1's (4) 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.

5.6 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 4.2.1 and we also show that the result with measurement error match with our final results in Theorem 1 (**??**). The below section will summarize the population distribution and model setup and how it compares to our findings.

5.6.1 Contributions

Given the one-step feedback model[24] 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 criterion 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 (5.6.2).

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.

5.6.2 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 fraction 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, then there exists a function $u : C \to R$, such that the institution's expected utility for a policy τ is given by:

$$U(\tau) = \sum_{j \in \{A,B\}} g_j \sum_{x \in \chi} \tau(x)\pi_j(x)u(x) \tag{5.14}$$

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:

$$\Delta\mu_j(\tau) = \sum_{x \in X} \pi_j(x)\tau_j(x)\Delta(x) \tag{5.15}$$

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, consider a function $\rho : X \to [0, 1]$ such that individuals of score x succeed with probability $\rho(x)$. This assumption is formally detailed assumptions section 5.6.4.1.

5.6.3 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 makes 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 $\Delta\mu_j^{MaxUtil} = \Delta\mu_j(\tau^{MaxUtil})$).

We say 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 5.5.

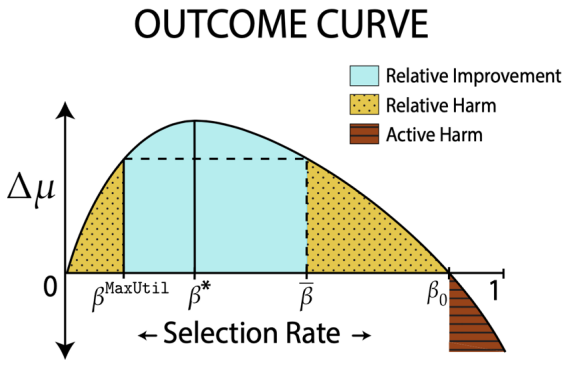


Figure 5.5: Outcome Curve

In the outcome curve figure 5.5, the following selection rates are of interest:

1. We define $\beta^{MaxUtil}$ as the selection rate for A under MaxUtil Policy.
2. β_0 as the harm threshold, such that $\Delta\mu_A(r_{\pi_A}^{-1}(\beta_0)) = 0$;

3. β^* as the selection rate such that $\Delta\mu$ is maximized;

4. $\bar{\beta}$ 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.

5.6.4 Results

Using on the outcome curve (fig 5.5), 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.

5.6.4.1 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.:

$$\frac{u_-}{u_+} < \frac{c_-}{c_+} \tag{5.16}$$

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 $\rho : X \rightarrow [0, 1]$ 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.

$$\text{if } x_1 > x_2 \text{ then } \rho(x_1) > \rho(x_2) \tag{5.17}$$

5.6.4.2 Corollaries

Based on the discussed assumptions, below are the relevant corollaries discussed in this work. For the corollaries, the selection rates must be equal for DemParity, but for EqOpt we can define a transfer function, $G^{(A \rightarrow B)}$, 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.

5.6.4.3 Corollary 3.2

If the assumption 5.16 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.

5.6.4.4 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.

5.6.4.5 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.

5.6.4.6 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 $\beta_A^{EqOpt} > \hat{\beta}_A^{EqOpt}$.

5.6.5 Comparison with our final results

The concept of measurement error (5.18) is akin to the noisy scores X, which are dependent on talent and environment in our work (4.2.1). Therefore, we compare our model to the measurement error section 5.6.4.6 where the scores are underestimated only for the disadvantaged group. The two sections below conclude whether our theorems hold in this paper's model.

5.6.5.1 Theorem 1

Theorem 1(4) 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.

$$X + e(X) := \hat{X} \sim \hat{\pi} \tag{5.18}$$

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 5.17, 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.

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

5.7.1 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 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.

5.7.2 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 $h^*(x)$.

5.7.3 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 $P_{(x,y)\sim D}(h(x) = 1|y = 1, x \in A) = P_{(x,y)\sim D}(h(x) = 1|y = 1, x \in B)$.

5.7.4 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

$$(1 - r)(1 - 2\eta) + r((1 - \eta)\beta_{POS}(1 - 2\nu) - \eta\beta_{NEG}) > 0 \tag{5.19}$$

and

$$(1 - r)(1 - 2\eta) + r((1 - \eta)\beta_{NEG} - (1 - 2\nu)\beta_{POS}\nu) > 0 \tag{5.20}$$

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.

5.7.5 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:

$$y = y(x) = \begin{cases} \neg h^*(x) & \text{with probability } \eta \\ h^*(x) & \text{w. p. } 1 - \eta \end{cases} \tag{5.21}$$

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 4.2.1 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.

5.7.6 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.

5.8 Research Review 7: On (Im)possibility of Fairness

5.8.1 Introduction

The Friedler et. al. paper which introduced the idea of World-views and also compared the previous works by accommodating their models into their own worldview model. To support the concept of worldview, this paper also introduces Spaces— which are the assumptions about the distributions. The main idea of this paper is — to study algorithmic fairness is to study the interactions among different spaces that make up the decision pipeline for a prediction task.

5.8.2 Spaces

Three spaces are introduced by the paper:

- Construct Space: Which contain the true features that we want to make decisions on, for example "grit" or "intelligence" of a student during college selection process. We of course don't have access to these features. The construct space $CS = (P, d_p)$ consists of a pair of individuals and the distance between them. It is assumed that the distance d_p correctly captures closeness with respect to the task.
- Observed Space: In reality, we might not know the Construct Space's features. All we have is what we measure or observe i.e. Observed Space. Examples of Observed space are SAT scores used to measure Grit and High school score for intelligence etc., The observed space (with respect to T) is a metric space OS = (P, d). We assume an observation process $g : P ß \hat{P}$ that generates an entity $\hat{p} = g(p)$ from a person $p \in CS$.
- Decision Space (DS): A decision space is a metric space $DS = (O, d_O)$, where O is a space of outcomes and d_O is a metric defined on O. A task T can be viewed as the process of finding a map from P or P to O.

These spaces interact with each other through Algorithmic Decision Making, which is defined as a mapping between spaces. The desired outcome is a mapping from CS to DS via an

unknown and complex function $o = f(X_1, X_2, ...)$ of features that lie in the construct space.

The paper then defines several definitions between spaces, which are in general used to compare spaces:

Distortion : between two metric spaces. There are many different ways to compare metric spaces using their distances. Distortion is a worst-case notion: it is controlled by the worst-case spread between a pair of distances in the two spaces.

Wasserstein Distance WD The WD finds an optimal transportation between the two sets and computes the resulting distance. It is a metric when d is.

Gromov Wasserstein Distance (GWD): Finally, we need a metric to compare subsets of points that lie in different metric spaces. Intuitively, we would like some distance function that determines whether the two subsets have the same shape with respect to the two underlying metrics.

The above three definitions are used to prove most theorems discussed in the paper. In general, the paper considers the following definition of fairness: A mapping $f : CS \to DS$ is said to be fair if objects that are close in CS are also close in DS. Specifically, fix two thresholds ϵ, ϵ' . Then f is defined as ϵ, ϵ'— fair if for any $x, y \in P$, $d_P(x, y) \leq \epsilon \to d_O(f(x), f(y)) \leq \epsilon'$ Note that the definition of fairness does not require any particular outcome for entities that are far apart in CS.

The paper then considers notion of Worldviews:

- WYSIWYG: the mapping between the construct space and observed space by asserting that the construct space and observed space are essentially the same.
- Structural Bias: which can be informally understood as the existence of more distortion between groups than there is within groups when mapping between the construct space and the observed space, thus identifying when groups are treated differentially by the observation process. WAE: Distribution of each group is the same.The idea is that any difference in the groups' performance (e.g., academic achievement) is due to factors outside their individual control (e.g., the quality of their neighbourhood school) and should not be taken into account in the decision making process.

The final results of the paper include theorems guaranteeing Individual Fairness if the worldview assumption is WYSIWYG and under the WAE, Group Fairness Notion is guaranteed.

5.8.3 Relation to our Model

As discussed in the section 4.2.1, we extend the model proposed in Friedler et. al. [13], having three distribution spaces, the Construct Space, Observed Space and Decision Space. The Construct Space(CS) represents the value of the attribute that is truly relevant for the prediction task, such as intelligence of a student. This value is usually not measurable, so the 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 distribution of the model. The below sections cover our thought process of why and how we extended the Friedler et. al model.

5.8.3.1 Construct Space(CS)

As we do not have access to construct space, in our model's case this evaluates to the Talent and Environment distributions.

Talent Distribution T: We consider two possible group membership for an individual, that is either A or B and the membership is represented by $g \in \{A, B\}$. Our main goal is to determine this Talent $t \sim T$ of an individual (where T is the talent distribution).

We consider Talent in the CS, as we do not have direct access to the talent. In other words, talent is the inherent property of individuals which we believe we could only approximate using the *Observed Space* like SAT scores.

Environment Distribution E Unlike any of the previous models, we also take into account the Environment Variable , which is a measure of how conducive things are around an individual to promote her success. In the real-word data-sets, it is often hard to ascertain whether an individual comes from a family of means or had proper schooling. Hence we consider Environment in Construct space as well.

5.8.3.2 Observed Space

The observed space contains the feature vectors which are an approximation of the construct space for example be SAT score or high school grades, which measure the talent (with an error η).

Since the performance scores X we consider in our model4.2.1 is generally the input to the classification algorithms, we consider X in the Observed space.

5.8.3.3 Decision Space

Finally, the Decision Space represents the distribution of the output of the classifier $\hat{Y} \in \{0, 1\}$ (also called labels), which will be used to validate fairness. For example, a model will satisfy Demographic Parity if: $Pr(\hat{Y} = 1|G = A) = Pr(\hat{Y} = 1|G = B)$. The decision space will contain a set of hypothesis represented by $\mathcal{H} : X \to \{0, 1\}$, where $\mathcal{H}$ is the set of all Threshold functions. Then $h \in \mathcal{H}$ is as shown below.

$$h(x) = \begin{cases} 0 & x \leq \tau \\ 1 & x > \tau \end{cases}$$

where τ is the threshold of the function h.

6 Worldviews of Research works

6.1 Introduction

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 the advantaged group. Hence the assumption of Talent distribution being the same for all groups, supports the idea that demographic groups are 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, with which the see the world and solve the problem of bias against the disadvantaged.

This section will review in detail seven different models from other related fairness literature that we have reviewed from chapter 5 on-wards. We see that while there are a few models[2, 25] that 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. This chapter will discuss and compare the research works with our model and our motivation to create our model in a way as discussed.

6.2 Research Work 1: Downstream Effects of AA

6.2.1 Overview

The paper Downstream Effects of Affirmative Action [20] was the initial motivation for us to consider screening decisions problem where we want to hire a candidate given her scores. There is two limitations of this paper from the bias model's standpoint. First, the Talent Distributions T_i(Refereed as Type) is different for different groups (i represents the group membership). Second,

as discussed in section 5.2, the paper discusses two-staged model where this work assumes that even after going though the university, the disadvantaged group remains at a lower talent distributions. Therefore, we deem that this research work's model is discriminatory against the disadvantaged group.

6.2.2 Bias Model

This section will discuss how the population distributions for the two groups are defined. The two populations of students represented by $i \in \{1, 2\}$, where $i = 1$ is advantaged group and $i = 2$ is disadvantaged. Students have a type/talent 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 $P_i = \mathcal{N}(\mu_i, \sigma_i^2)$ and the assumption is $\mu_1 > \mu_2$ i.e. the disadvantaged group is in general supposed to be less talented.

In addition, the model considers two-staged hiring process where candidates are first hired into a University based on their SAT scores whose distribution is represented as $S_i = T_i + X$ (where X follows a normal distribution with mean 0 and variance 1). A student with SAT Scores is accepted in the university with the probability $A_i(s) \rightarrow [0, 1]$, where $A_i(s)$ is a threshold function. upon successful admission, to further determine Student type, the paper defines another distribution for each student which receives a grade $G_i = T_i + Y$ (where Y follows a normal distribution with mean 0 and variance γ^2).

This raises a limitation that the Talent distribution T_i remains the same when calculating the Grades distribution. The University education does not attempt to address or correct the historical forces. On the contrary, the university is just viewed as a testing mechanism which help the employer in determining the talent of individuals.

6.2.3 Conclusion

Hence, as outlined in section 6.2.2, the bias model of this research has two shortcomings, initially it considers that the talent distribution is different for the two groups with disadvantaged group having a lower talent distribution. This assumption makes this model a prejudiced. In addition, it also assumes that even after getting a university education, the talent distribution of the disadvantaged individuals is the same and worse of then the advantaged group. This adds another level of inherent discrimination.

6.3 Research Work 2: Disparate Impact of Strategic Manipulation

6.3.1 Overview

This research work by Immorilica et. al. [17] as discussed in detail in section 5.3 talks about the case of college admissions where the agents are the students who manipulate their feature vectors by using the SAT test prep services in an attempt to "trick" the SAT exam. This research work considers the world divided into two groups and there are two distributions for each group. One distribution represents the unmanipulated scores and other the manipulated scores and it claims that both are biased against the disadvantaged group B. However it also claims that although the scores for disadvantaged group are lower, but their "true" thresholds are also lower.

Hence, this model is more comparable to our model, as it claims that although the scores are distorted for the disadvantaged group, their true talents are similar to the advantaged group.

6.3.2 Bias Model

Consider that the candidates have set of features represented by $x \in X = [0,1]^d$ such as SATs or Grades, and belong to either 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 x to a feature vector y must pay a cost of $c_m(y) - c_m(x)$ where $y \geq x$. To model disadvantage, the study assumes that

$$c_A(y) - c_A(x) \leq c_B(y) - c_B(x) \tag{6.1}$$

i.e. Group A members pay a lower cost than Group B.

Let τ_A and τ_B represent the "true" thresholds over the unmanipulated features and σ_A and σ_B be thresholds over the manipulated features. This work assumes that $\tau_B < \tau_A$ and $\sigma_B < \sigma_A$. Therefore, this model states that although the scores distribution for the disadvantaged group are biased, the true threshold is also lower for the group. This is demonstrated in figure 6.1, where $\tau_B < \tau_A$ and everyone in Group B who has $x > \tau_B$ has the label $\hat{y} = 1$ and similarly for Group A.

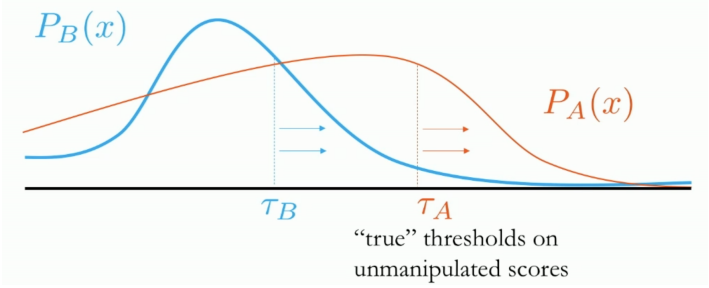


Figure 6.1: True Thresholds τ_A and τ_B

6.3.3 Conclusion

This Bias model discussed in this work alligns more with our Talent and Environment model. Although, the scores distributions are biased against group B, however that 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 are relatively same for the two groups.

6.4 Simplicity Creates Inequity

6.4.1 Overview

This work by Kleinberg et. al.[23] has a similar bias model compared to Research work 1(section 6.2) where the world's assumption is that the disadvantaged group's distribution has in general 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.

6.4.2 Bias Model

To illustrate the partiality in the bias model, consider that $x = (x^{\langle 1 \rangle}, x^{\langle 2 \rangle} ... x^{\langle k \rangle})$ represents the feature vector and 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.

Then the main assumption of this work "Likelihood Condition" is represented as:

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 feature vectors according to the true distribution function f are more represented in Group A.

6.4.3 Conclusion

To sum up, this research work considers a worldview that is 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.

6.5 Research Work 4: From fair decision making to Social Equality

6.5.1 Overview

As discussed in detail the section 5.5, 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.

6.5.2 Bias model

Similar to score distributions in our model, this work proposes the distribution called Qualification Profile(π), which represents the probability distribution for a particular evaluation(V either 0 or 1) 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 $\tau(V = v; G = j) : \{0, 1\} \times \{A, B\} \rightarrow \{0, 1\}$. Then the Institutional Utility $U(\tau)$ is defined considering that $u : \{0, 1\} \rightarrow R; v \rightarrow u(v)$ to be the utility function for *an* individual such that $u(0) \leq 0 \leq u(1)$. Then the Selection Rates(β) for a group j are defined

as: $\beta(G = j) = \sum_{v \in \{0,1\}} u(v).\tau(V = v; G = j).\pi(v|G = j)$. And 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 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.

6.5.3 Conclusion

This work has a worldview which 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. To conclude, this model also aligns with our model's Talent distribution equality claim, as over the period of time the qualification profiles becomes equal.

6.6 Research Review 5: Delayed Impact of Fair Machine Learning

6.6.1 Overview

This research work also considers dynamics similar to what we discussed in section 6.5 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 to start with, and when we apply fairness interventions these distributions change. Similar to the dynamics paper (section 6.5), the main aim of this paper is to equalize the population distributions.

6.6.2 Bias Model

his section briefly discusses the main aspects of the model with respect to population distribution comparison. 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 fraction 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 $(\Delta\mu_j < 0)$.

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:

$$\Delta\mu_j(\tau) = \sum_{x \in X} \pi_j(x)\tau_j(x)\Delta(x) \tag{6.2}$$

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

For population distribution comparison $\Delta\mu_j(\tau)$ is the most important function and in this paper the two populations eventually would reach equality when $\mu_A(\tau) = \mu_B(\tau)$.

6.6.3 Conclusion

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.

6.7 Recovering from Biased Data

6.7.1 Overview

Discussed in section 5.7, 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, however the real-world test data is free from such bias.

6.7.2 Bias Model

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.

Since only training data is effected with the bias discussed, the worldview aligns closely with our model.

6.7.3 Conclusion

To sum up, 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 distributions, both of which are unbiased.

Bibliography

[1] Reuben Binns. On the apparent conflict between individual and group fairness. pages 514–524, 01 2020.

[2] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? 12 2019.

[3] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. pages 13–18, 12 2009.

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

[5] Simon Caton and Christian Haas. Fairness in machine learning: A survey. 10 2020.

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

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

[8] Lee Cohen, Zachary Lipton, and Yishay Mansour. Efficient candidate screening under multiple tests and implications for fairness. 05 2019.

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

[10] Cynthia Dwork, Nicole Immorlica, Adam Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. 07 2017.

[11] Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. 09 2016.

[12] Sorelle Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. pages 329–338, 01 2019.

[13] Sorelle A. Friedler, C. Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *ArXiv*, abs/1609.07236, 2016.

[14] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: Testing software for discrimination. pages 498–510, 08 2017.

[15] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. pages 269–278, 01 2019.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. 10 2016.

[17] Lily Hu, Nicole Immorlica, and Jennifer Vaughan. The disparate effects of strategic manipulation. pages 259–268, 01 2019.

[18] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. 01 2019.

[19] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Wu. Eliciting and enforcing subjective individual fairness. 05 2019.

[20] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. pages 240–248, 01 2019.

[21] Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Improving consequential decision making under imperfect predictions. 02 2019.

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

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. 09 2016.

[24] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. 03 2018.

[25] Hussein Mouzannar, Mesrob Ohannessian, and Nathan Srebro. From fair decision making to social equality. 12 2018.

[26] Arvind Narayanan. 21 fairness definition and their politics by arvind narayanan. 01 2018.

[27] psu.edu. Sums of independent normal random variables. 10 2020.

[28] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna Gummadi, Adish Singla, Adrian Weller, and Muhammad Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. 07 2018.

[29] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. 06 2019.

[30] 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.

[31] Samuel Yeom and Michael Tschantz. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. 08 2018.

[32] Muhammad Zafar, Isabel Valera, Manuel Rodriguez, and Krishna P. Gummadi. Fairness constraints: A mechanism for fair classification. 07 2015.