On Fairness of Classification in Machine Learning

DoWon D. Kim

a senior thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts in Mathematics at
Princeton University

Adviser: Mark Braverman

May 6, 2019

Abstract

As our society becomes increasingly automated, there is a social concern for algorithmic decision-making to be fair and objective. In this thesis, we initiate the study with an overview of several criteria for *group fairness*, their limitations and motivations, and the criterion of *individual fairness*. We start with the case of a single-classifier, and extend the fairness properties to systems using multiple classifiers in composition. We demonstrate how to construct such systems, and find that fairness in social situations varies greatly with context. We find that that classifiers that are *fair-in-isolation* may not necessary yield fair systems in naive composition, and fair systems can be constructed from individually unfair classifiers. Finally, we examine the behavior of *group fairness* criteria under systems of multiple classifiers.

Acknowledgements

First, I would like to thank Professor Mark Braverman for being my adviser for this senior thesis. I'm extremely grateful for your patience and guidance this past year. Thank you for taking in a student who knew very little about the topic at the start and leading him until the end.

To Professor Elias Stein, I want to thank you for inspiring me to pursue Mathematics as my major. When I was unsure about my trajectory, you gave me the courage to continue with Mathematics, and I wish you could have been here with us today.

As the culminating work of my Princeton undergraduate experience, this thesis would not have been possible without the support of my friends throughout my four (and a half) years here. I want to especially thank Chris Donghyuk Choo, Tammy Tseng, Matt Harrington, and David Nie for being a constant source of joy and support. Thank you, Leo Choi, for being a wonderful roommate and friend (and for all the food you generously gave me). Shoutout to the Patton 307 squad, y'all the real ones.

Above all, I am grateful to my parents. Thank you Mom and Dad for your unconditional love and support. I would not be who I am today without you.

And as always, *Food* = *Love*.

Declaration

I declare that I have not violated the Honor Code during the composition of this work. This paper represents my own work in accordance with University regulations.

I authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purposes of scholarly research.

DoWon D. Kim

May 6, 2019

Contents

Chapter 1

Introduction

With the advance of machine learning in recent history, consequential choices are often advised by automated, algorithmic decision-makers. This is evident in modern systems in banking, insurance, medicine, even online advertising as decisions are informed by statistical risk that outline potential rewards and losses of each outcome. For instance, hiring managers may act with bias, by refusing to hire from an equally qualified minority group due to race or gender, or auto insurers may charge a higher premium to male versus female drivers to account for differing accident rates among genders.

There have been legal precedents to stymie such discrimination. In the 1976 US Supreme Court case *Washington vs. Davis*, the court held that under the Equal Protection Clause of the 14th Amendment, employment discrimination under race was unconstitutional, and further defined the standard of "disparate impact" where a practice is considered discriminatory if it has an unjustified adverse effect on a group with or without animus against the group. As such, deeper research into the construction of fair algorithms is increasing dire.

The notion of fairness in classification is broadly divided into **group fairness**, which requires that different demographic, gender, and racial groups experience sim-

ilar treatment within reason, and **individual fairness**, which requires that individuals who are similar with respect to a certain task-metric have similar classification outcomes.

We define a **classification problem** as the mapping of *individuals* to *outcomes*, for instance, assigning job applicants to outcomes $\{0, 1\}$: reject or accept. We limit our study to binary classifiers, C, and define the universe of individuals relevant for a task as U. For $u \in U$, let p_u denote the probability of assigning $u \to 1$. Formally, for a classifier, we can write $C : U \times \{0, 1\}^* \to \{0, 1\}$ where $\{0, 1\}^*$ represents random bits of the classifier. This way, for $u \in U$, we can write $\mathbb{E}_r[C(u)] = p_u$. For distributions on the set of outcomes O of C, we say $M : U \to \Delta(O)$. We say that a classifier or algorithm is **fair-in-isolation**, to mean that it satisfies a certain fairness definition without composition.

In Chapter 2, we lay out the preliminary definitions of group fairness and individual fairness, and demonstrate procedures to construct individually fair classifiers. Chapter 3 dives into behavior of individually fair classifiers in *functional composition*, where multiple classifiers generate a single outcome, and *task-competitive composition* in a single-slot decision problem, where a system of multiple classifiers must choose one outcome from a set of task metrics. Finally, in Chapter 4, we extend the notions of fairness in composition to group fairness criteria, and summarize our findings in Chapter 5.

Chapter 2

Definitions of Fairness

First, we discuss several criteria for *group fairness*. It is natural to pursue a group-based definition, as the absence of proportional representation can signal discrimination against underrepresented minorities. Thus, we must consider both qualifications that encode information about the individuals under consideration, and also set aside a subset of the features as **sensitive attributes**, denoted by $\mathcal{A}$: features such as gender or race that may result in discriminatory treatment of the individual. Consider a classification task that predicts whether a job applicant ought to be hired. Let $\mathcal{X}$ be the non-sensitive qualifications or **stratification set** of the applicant, such as GPA, work experience, test scores, etc., and let $Y = \{0, 1\}$ be the target variable that represents whether the candidate should truly be hired or not. As before, C is a binary classifier that makes a prediction based on some score function $R \in [0, 1]$. Our definitions will follow from the joint distribution of $C, Y, \mathcal{A}$.

2.1 Group Fairness

One of the most popular definitions in literature is *statistical parity*, which seeks to ensure that all groups within the set of sensitive attributes are treated the same.

Definition 1. *A classifier satisfies* ***statistical parity*** *or* ***independence*** *if* $\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_i] = \mathbb{P}[C = 1 \,|\, \mathcal{A} = a_j]$ *for all* $i \neq j$.

Within the context of our example, this implies that the acceptance rates for applicants in two groups a_i, a_j must be equal. Statistical parity is often pursued as a fairness criterion due to the ability to easily relax the constraint by a certain $\varepsilon > 0$ in practice.

$$\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_i] - \mathbb{P}[C = 1 \,|\, \mathcal{A} = a_j] \geq \varepsilon$$

$$\frac{\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_i]}{\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_j]} \geq 1 - \varepsilon$$

However, this criterion is not without its limitations. For instance, it ignores any possible correlations between the sensitive attribute A and the outcome Y. Thus, it opens up the possibility for a callous or lazy decision-maker to be fair and inaccurate. Suppose a company hires diligently from one group $\mathcal{A} = a_1$ at some rate $p_a > 0$ and hires randomly (or lazily) from group $\mathcal{A} = a_2$. While this satisfies independence, it is more likely that more unqualified applicants will be selected from a_2. In hindsight, hired candidates from the latter group may appear to perform worse than the former, establishing a negative track record and widening the bridge between the two groups from the perspective of the company. Furthermore, often a classifier may be fair at the group level, but unfair at the subgroup level. A classifier that satisfies statistical parity with respect to race and gender independently still may fail with respect to the conjunction of race and gender. And often in large feature spaces, sensitive attributes such as race or gender are not sufficiently rich to describe and protect every "at-risk" groups from discrimination. A classifier may protect women from being hired at disproportionately low rates compaerd to men, but may still allow discrimination within the set of women, such as pregnant women or women with disabilities. We will see ways to address such problems in intersections of groups in our discussion of individual fairness.

Remark. *Blindness or ignoring sensitive attributes entirely from classification feature space does not necessarily guarantee fairness either. In large feature spaces, there can be many highly correlated features that serve as proxies for the sensitive attribute. An example may be residence neighborhoods encoding demographic information such as race. Removing sensitive features entirely would result in an essentially similar classifier.*

Definition 2. *A classifier satisfies* ***separation*** *or* ***equality of odds*** *if* $R \perp A|Y$*, i.e. for all groups* $i \neq j$*,*

$$\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_i, Y = 1] = \mathbb{P}[C = 1 \,|\, \mathcal{A} = a_j, Y = 1]$$
$$\mathbb{P}[C = 1 \,|\, \mathcal{A} = a_i, Y = 0] = \mathbb{P}[C = 1 \,|\, \mathcal{A} = a_j, Y = 0]$$

Note that $\mathbb{P}[C = 1 \,|\, Y = 1]$ is the *true positive rate*, the probability that the classifier correctly predicted positive instances, and $\mathbb{P}[C = 1 \,|\, Y = 0]$ is the *false positive rate*, the probability that the classifier assigns positive outcomes to negative instances. Separation thus requires that all groups experience the same true positive and false positive rate.

This appears to solve the laziness problem presented in Definition 1, as it provides incentive to reduce errors uniformly throughout all groups. However, it may not have close the gap for groups that experience discrimination. Suppose we have two groups A, B with 100 applicants in each. Group A has 72 qualified individuals and B has 8. If the company decides to hire a total of 40 qualified applicants, 36 offers will be given to those from A while only 4 will be given to group B. Over time, the welfare gap between the two groups is likely to increase.

The last group fairness criteria we introduce requires that any two groups experience the same *positive predictive value*, the probability that a subject with positive prediction is truly a positive instance.

Definition 3. *A classifier satisfies* ***sufficiency*** *or* ***predictive parity*** *if for $i \neq j$ and all values c in the support of C, $\mathbb{P}[Y = 1 \mid C = c, \mathcal{A} = a_i] = \mathbb{P}[Y = 1 \mid C = c, \mathcal{A} = a_j]$*

The intuition behind sufficiency is similar to that of separation, that the rate of correct positive predictions is the same for all groups. While it promotes accuracy, it does not help bridge the gap due to the same reasoning as separation.

Remark. *These criteria all constrain the joint distribution in different ways such that imposing any two of them in composition can overconstrain the space leaving only degenerate solutions. We examine the proofs for their pairwise incompatibility in Appendix A.*

We now have some intuition that popular group fairness criteria suffer from limitations and are not sufficiently rich to treat all socially relevant subgroups within a population. Here, we turn our attention to the notion of individual fairness, where we consider for any two individuals, how similar or dissimilar they are with respect to a given **task-metric** T. With this standard, we require that people who are similar should be treated similarly as a notion of fairness.

2.2 Individual Fairness

To set up the problem, we now consider *individuals* as objects to be classified properly. Let U denote the universe of individuals, O the set of outcomes. To guarantee fairness, we consider randomized classifiers mapping individuals to distributions over outcomes. In the subsequent definitions we will introduce a distance metric d and randomized mappings $M : U \to \Delta(O)$ from individuals to probability distributions over outcomes. Note that this naturally models a classification scheme. In order to classify $x \in U$, choose an outcome $o \in O$ according to the distribution $M(x)$.

Definition 4. *Let $d : \Delta(O) \times \Delta(O) \to [0, 1]$ denote the total variation distance on distributions over O. Given a universe of individuals U, and a metric $\mathcal{D}$ for a*

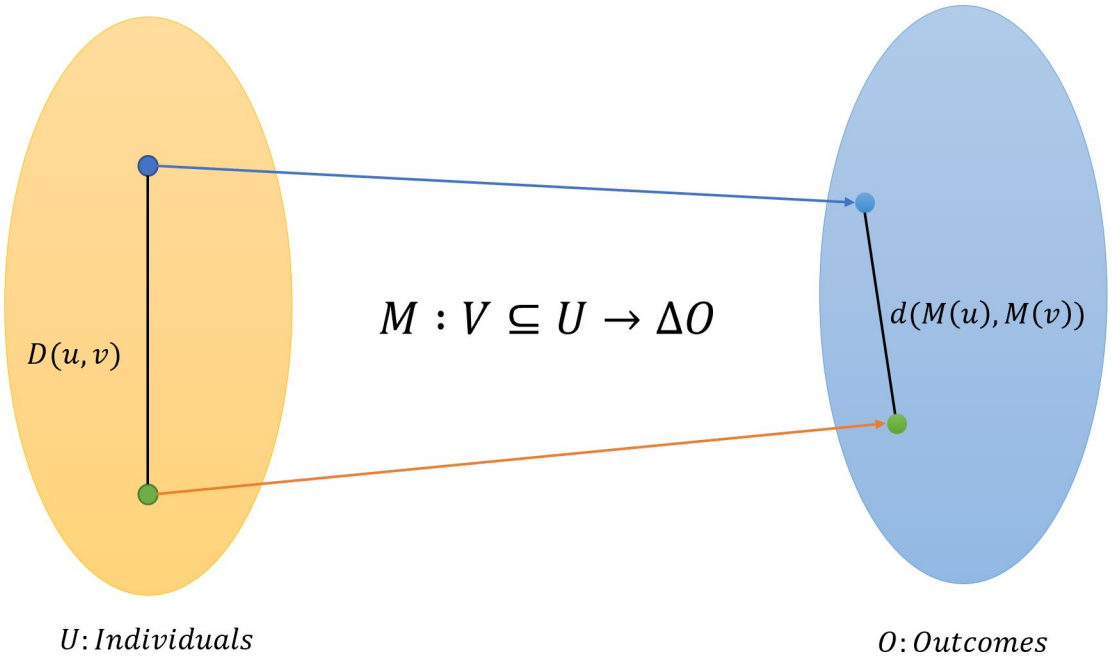

Figure 2.1: Mapping individuals to outcomes in Individual Fairness

classification task T with outcome set O, a randomized classifier $C : U \times \{0,1\}^* \to O$, such that $M : U \to \Delta O$, and a distance measure $d : \Delta(O) \times \Delta O \to \mathbb{R}$, C is ***individually fair*** *if and only if for all $u, v \in U$, $\mathcal{D}(u, v) \geq d(M(u), M(v))$.*[1]

Before we can apply individual fairness in practice, we need to determine a task-specific metric. We address the question of how to find a metric to our discussion of composition, and for now, assume that we have a complete metric for each task. Note in the case of binary classifiers, $|O| = 2$, the total variation distance d can be expressed using probabilities

$$d(M(u), M(v)) = |\mathbb{E}_r[C(u)] - \mathbb{E}_r[C(u)]| = |p_u - p_v| \tag{2.1}$$

[1] ΔO denotes the set of probability distributions on O.

2.2.1 Building Individually Fair Classifiers

Dwork and Hardt showed that individually fair classifiers can be constructed by solving a linear program to minimize the loss of an objective function with respect to the distance constraints. We will use this fact throughout this section to assume we can find such individually fair classifiers given a distance metric. In the following lemma, we show the feasibility of constructing classifiers with positive distance between elements.

Lemma 1. *Let $V \subseteq U$. If there exists an individually fair classifier $C : V \times \{0,1\}^* \to \{0,1\}$ i.e. $\mathcal{D}(x,y) \leq d(M(x), M(y))$ for all $x, y \in V$, then for any $z \in U \setminus V$ there is another individually fair classifier $C' : V \cup \{z\} \times \{0,1\}^* \to \{0,1\}$ such that $\mathcal{D}(u,v) \geq d(M(u), M(v))$ for all $u, v \in U$, with identical behavior to C on V.*

Proof. If $V = \emptyset$, any value p_z trivially suffices to fairly classify z. If $|V| = 1$, for $v \in V$, we can choose any p_z such that that $|p_v - p_z| \geq \mathcal{D}(v,z)$ holds.
For $|V| \geq 2$ we apply the following ClassifierExtension with p_t as the probability of positive classification of z's nearest neighbor in V under C.[2] Let the nearest neighbor be u. Note that the algorithm only modifies $\hat{p_z}$ if a distance constraint is violated. Thus, we need only to check that on each modification, the distance constraints between z and other elements in the opposite direction of the move are not violated.

Without loss of generality, suppose $\hat{p_z}$ is decreased to move within an acceptable neighborhood of u, i.e. $\hat{p_z} \geq p_u$. Then it remains to show that for all v with $p_v > \hat{p_z}$ the distance constraint is remains satisfied. Pick one such v. Since we set $\hat{p_z}$ to p_t, we have $\hat{p_z} - p_u = \mathcal{D}(z,u)$, and by assumption we have $p_v - p_u \leq \mathcal{D}(u,v)$. By the

[2] As before, p_i is the probability that C positively classifies candidate i.

input: metric $\mathcal{D}$, $V \subset U$, target probability p_t, individually fair classifier C, target element $z \in U \setminus V$ to be added
Initialize S **to** V, $\hat{p_z}$ **to** p_t;
for $s \in S$ **do**
  $dist \leftarrow \mathcal{D}(s, x)$;
  if $dist < p_s - \hat{p_z}$ **then**
    $\hat{p_z} \leftarrow p_s - dist$;
  else if $dist < \hat{p_z} - p_s$ **then**
    $\hat{p_z} \leftarrow p_s + dist$;
end
return $\hat{p_z}$

Algorithm 1: ClassifierExtension

triangle inequality and the fact that now we have $p_v \geq \hat{p_z} \geq p_u$:

$$\begin{aligned} \mathcal{D}(u,v) - \mathcal{D}(u,z) &\leq \mathcal{D}(z,v) \\ \mathcal{D}(u,v) - (\hat{z} - p_u) &\leq \mathcal{D}(z,v) \\ (p_v - p_u) - (\hat{z} - p_u) \leq \mathcal{D}(u,v) - (\hat{z} - p_u) &\leq \mathcal{D}(z,v) \\ p_v - \hat{p_z} &\leq \mathcal{D}(z,v) \end{aligned}$$

Thus, the individual fairness criteria is satisfied for z with respect to all $v \in V$ and C' is an individually fair classifier for $V \cup \{x\}$.

□

Lemma 1 allows us to build a fair classifier on U in $O(|U|^2)$ time. The subsequent results also follow from the lemma.

Corollary 1.1. *If there exists an individually fair classifier C on a $V \subset U$, then there exists an individually fair classifier C' on U which is individually fair forllary all $u, v \in U$ and behaves identically to C on V.*

By applying the algorithm inductively to remaining elements in $U \setminus V$, Corollary 1.1 follows immediately.

Corollary 1.2. *There exists an individually fair classifier such that* $d(M(u), M(v)) = \mathcal{D}(u, v)$ *for all* $u, v \in U$

Corollary 1.2 can be seen by starting with a classifier that is fair for a pair of individuals and inductively placing additional individual's probabilities at maximum distance without violating the constraint, then applying ClassifierExtension repeatedly. This is useful when there is an 'axis' within the metric where preserving the distance between extreme points is important in classification.

Corollary 1.3. *Given a metric* $\mathcal{D}$ *and* $\alpha \in \mathbb{R}^+$*, for any pair* $u, v \in U$*, there exists an individually fair classifier* $C : U \times \{0,1\}^* \to \{0,1\}$ *such that* $p_u/p_v = \alpha$*, where* $p_u = \mathbb{E}[C(u)]$ *and* $p_u = \mathbb{E}[C(v)]$

Corollary 1.3 follows from choosing $p_u/p_v = \alpha$ without regard for the difference between p_u and p_v and adjusting afterwards. Let $\beta|p_v - p_v| = \mathcal{D}(u, v)$, and let $\hat{p_u} = \beta p_u, \hat{p_v} = \beta p_v$, such that $|\beta p_v - \beta p_u| = \beta|p_v - p_u| \leq \mathcal{D}(u, v)$, but the ratio $\frac{\beta p_u}{\beta p_v} = \frac{p_u}{p_v} = \alpha$ remains the same.

In the following chapter, we explore how individually fair classifiers can be composed to construct fair systems, and demonstrate the potential dangers of naively combining classifiers.

Chapter 3

Fair Compositions of Classifiers

There are few instances where a single task and classifier describes an entire system. For instance, there may be competition between multiple classifiers for a given decision problem: a candidate may apply to more than one job or university, or an individual's classification may depend on the order or classification outcomes of others. In this chapter, we consider **fairness under composition**, where systems are composed of multiple classifiers in conjunction. We present the two classes of composition [1]: **function composition** and **task-competitive composition**. We consider the construction and limitations of such systems, and provide examples of such systems in our society. Of particular interest is that in certain cases, classifiers that are fair-in-isolation, when composed together do not necessary yield "fair" results, and conversely, unfair classifiers may be constructed into fair systems.

3.1 Functional Composition

Functional compositions are systems with multiple classifiers for a single task generating a single outcome. Two main branches of functional composition are same-task and multiple-task, which considers the intersection of several tasks as one.

3.1.1 Same-task Functional Composition

Same-task functional composition systems may arise in cases such as an individual applying to similar jobs in multiple companies in same sector or a student applying to several universities. In these scenarios, we want to ensure that the individual is accepted by at least one agency; in other words, we are concerned with *OR*-fairness. While there may be some variation between jobs despite being in the same sector, in the below definition we assume that there is one metric for all classifiers and the final outcome.

Definition 5. *Given a universe, task pair (U, T), with a metric $\mathcal{D}$, and a set of classifiers $\mathcal{C} = \{C\}_i$, define the indicator function :*

$$\chi_u = \begin{cases} 1 \text{ if } \sum_{C_i \in \mathcal{C}} C_i(u) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

*which indicates whether $u \in U$ achieves at least one positive classification. Let $\tilde{\chi_u} = \mathbb{P}[\chi_u = 1] = 1 - \prod_{C_i \in \mathcal{C}}(1 - \mathbb{P}[C_i(u) = 1])$. Then a system satisfies **OR-fairness** if $\mathcal{D}(u, v,) \geq d(\tilde{\chi_u}, \tilde{\chi_v})$.*

While it could be argued that acceptance to multiple jobs or universities may have greater positive impact for an individual than being accepted to just one, the marginal impact of a second acceptance is much less than one acceptance is to none. Suppose a task specifies a certain n number of positive classifications for each individual, χ_u in Definition 5 can be adjusted to require at least n. For sake of a clarity, let U be a universe of students applying to college in a given year, each with a qualification metric $q_u \in [0, 1], \forall u \in U$. As before, we can define $\mathcal{D}(u, v) = |q_u - q_v|, \forall u, v \in U$. Each college will be the classifier $C_i \in \mathcal{C}$, which admits students fairly with respect to the $\mathcal{D}$ metric. And the system $\mathcal{C}$ is OR-fair if the indicator variable χ_u satisfies individual fairness under the same metric. We find that despite the assumption that

each college is fair with respect to its applicants, there are sources of unfairness in their composition.

Mixed-Degree Composition

One rather obvious source of unfairness arises from the case when different numbers of classifiers are applied to different subsets of the universe. Not all students apply to the same set of universities, and naturally, those who are able to apply to a greater number of them vastly improve their chances of admission. For any $u, v \in U$ with $q_u = q_v$, in this case it cannot be such that $\mathbb{E}[\chi_u] = \mathbb{E}[\chi_v]$, violating individual fairness.[1]

Theorem 1. *For any (U, T) pair with a non-trivial metric $\mathcal{D}$, there exists a set of individually fair classifiers $\mathcal{C}$ that do not satisfy OR-fairness if each element may be classified by different sets of classifiers with different cardinalities.*

Proof. For each individual $x \in U$, let $\mathcal{C}^x \subseteq C$ denote the set of classifiers acting on x. Consider a set of randomized classifiers $\mathcal{C}$ where all classifiers are identical and assign outcome 1 to $u, v \in U$ with p_u, p_v respectively. Without loss of generality, let $p_u \leq p_v, p_v - p_u = \mathcal{D}(u, v)$, and $p_v > 0$ (we can find such a C via Lemma 1). If $|\mathcal{C}^u| = 1$, and $|\mathcal{C}^v| = 2$, then immediately we see that:

$$\begin{aligned} |\mathbb{E}[\chi_v] - \mathbb{E}[\chi_u]| &= |(1 - (1 - p_v)^2) - p_2| \\ &= |p_v - p_v^2 + (p_v - p_v)| \geq \mathcal{D}(u, v) \end{aligned}$$

which does not satisfy individual fairness. □

This is in line with our intuition, that even for two *identical* students, if one applied to one more school than the other, their probabilities of acceptance will diverge.

[1] In general, aside from the contrived case when all students are admitted with probability $\{0, 1\}$.

Equal-Degree Composition

Now, we assume that the system requires all elements to be classified by the same number and the same set of classifiers. We find that in this case there still are possibilities for a composition of classifiers to violate OR-fairness. The key observation is that for elements with positive distance, the difference in their expectation of acceptance by at least one classifier *does not* diverge linearly in the number of classifiers in the composition. For instance, suppose $q_u = 0.5, q_v = 0.01$. If two classifiers each assign 1 with probability $p_u = q_u, p_v = q_v$, then the probability of positive classifiers by either of the two classifiers will be 0.75 for u and ≈ 0.02 for v, which diverges from their original distance $\mathcal{D}(u, v) = 0.49$.

This problem is exacerbated in situations with a small number of classifiers. As the number of classifiers increases, the probabilties of positive classification by at least one classifier for any pair will eventually converge, approaching 1. However, these situations are not uncommon; in college applications, we would not expect students in most cases to apply to every single institution that exists, rather to a small subset of universities.

Theorem 2. *For any (U, T) pair, with a nontrivial metric $\mathcal{D}$, there exists a set of individually fair classifiers $\mathcal{C}$ which do not satisfy OR-fairness, even if each element in U is classified by all $C_i \in C$.*

Proof. Since $\mathcal{D}$ is non-trivial for $u, v \in U$, let $D(u, v) \in (0, 1)$. Let C be maximally individually fair with $d(M(u), M(v)) = \mathcal{D}(u, v)$ and $\mathbb{E}[C(u)] + \mathbb{E}[C(v)] < 1$. And let our composition be two identical copies of C. Then $\mathbb{E}[\chi_u] = 1 - (1 - p_u)^2, \mathbb{E}[\chi_v] =$

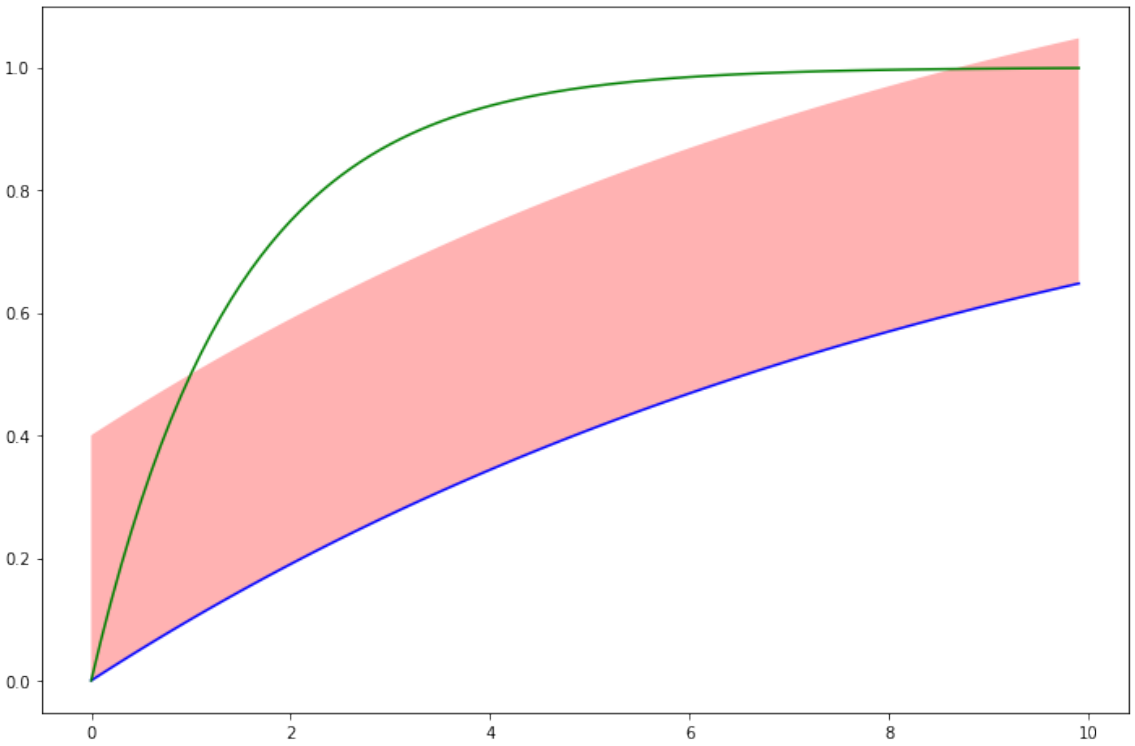

Figure 3.1: Graphical comparison of $1-(1-p_u)^n$ (blue) and $1-(1-p_v)^n$ (green) for $p_u = 0.1, p_v = 0.5$. The shaded pink region denotes the original *fair distance bound*. Observe that the values for p_u surpass the bounded region for $n \leq 8$.

$1-(1-p_v)^2$. Observe that:

$$\begin{aligned} |E[\chi_u] - \mathbb{E}[\chi_v]| &= |(1-p_v)^2 - (1-p_u)^2| \\ &= |(1-2p_v+p_v^2) - (1-2p_u+p_u^2)| \\ &= |2(p_u-p_v) - (p_u^2-p_v^2)| \\ &= |2(p_u-p_v) - (p_u-p_v)(p_u+p_v)| \end{aligned}$$

Since $|p_u - p - v| = \mathcal{D}(u,v)$, without loss of generality, $|\mathbb{E}[\chi_u] - \mathbb{E}[\chi_v]| = |2\mathcal{D}(u,v) - (p_u-p_v)(p_u+p_v)|$. By $p_u + p_v \leq 1$, $(p_u-p_v)(p_u+p_v) < \mathcal{D}(u,v)$, thus we have $|\mathbb{E}[\chi_u] - \mathbb{E}[\chi_v]| > \mathcal{D}(u,v)$ which completes the proof. □

Figure 3.1 demonstrates an example of this problem with small numbers of classifiers. Consider a case where the "worst" element is accepted with probability $\eta << \frac{1}{2}$

with respect to all classifiers. As other elements with probability of acceptance $p \geq \frac{1}{2}$ are classified, their probability of having at least one acceptance will converge to 1, diverging from the acceptance probability of the "worst" element, which increases much more slowly. From the perspective of the classifier, it attempts to maximize the allowed distance for some pairs of elements in order to strengthen the discimina-tory power between good and bad elements for a particular task, thereby increasing the potential for same-task divergence. In settings like loan applications, where an extended loan search with several credit agencies may impact a candidate's credit score, small stretches in distance may impact their eligibility greatly. But on the bright side, the following lemma helps us characterize non-trivial conditions for small sets of classifiers to satisfy OR-fairness.

Lemma 2. *Fix a set of classifiers $\mathcal{C}$. If $\mathbb{E}[\chi_x] \geq \frac{1}{2}$ for all $x \in U$, then the set of classifiers $\mathcal{C} \cup \{C'\}$ satisfies OR-fairness if C' satisfies individual fairness under the same metric and $\mathbb{P}[C'(u) = 1] \geq \frac{1}{2}$ for all $u \in U$.*

Proof. Consider a pair of elements $u, v \in U$. By assumption of individual fairness we have $|\mathbb{E}[\chi_u] - \mathbb{E}[\chi_v]| \leq \mathcal{D}(u, v)$ and $|\mathbb{E}[C'(u)] - \mathbb{E}[C'(v)]| \leq \mathcal{D}(u, v)$. Define a new indicator function χ',

$$\chi'_x = \begin{cases} 1 \text{ if } \chi_x + C'(x) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

This way, $\mathbb{P}[\chi'_x = 1] = 1 - \mathbb{P}[\chi_x = 0 \wedge C'(x) = 0]$. Let p'_u be the probability u is accepted by C' and p_u be the probability that u is accepted by at least one of classifiers in $\mathcal{C}$ (analogously for p'_v, p_v). Then it suffices to show that $|(1 - (1 - p_u)(1 - p'_u)) - (1 - (1 - p_v)(1 - p'_v))| \leq \mathcal{D}(u, v)$ to guarantee individual fairness of the composition. Simplifying we have,

$$|(1 - (1 - p_u)(1 - p'_u)) - (1 - (1 - p_v)(1 - p'_v))| = |p_v p'_v - p_u p'_u + p_u - p_v + p'_u + p'_v|$$

Let $t = p_u - p_v, t' = p'_u - p'_v$. Without loss of generality, assume that either $t, t' > 0$, or t, t' have different signs. Then we have

$$\begin{aligned}
|p_v p'_v - p_u p'_u + p_u - p_v + p'_u + p'_v| &= |p_v p'_v - p_u p'_u + t + t'| \\
&= |p_v p'_v - (p_v + t)(p'_v + t') + t + t'| \\
&= | - p_v t' - p'_v t - tt' + t + t'|
\end{aligned}$$

Since $p_v, p'_v \geq \frac{1}{2}$, $|t|, |t'| \leq \mathcal{D}(u, v)$, therefore we arrive at our desired result:

$$| - p_v t' - p'_v t - tt' + t + t'| \geq |\frac{1}{2}t' + \frac{1}{2}t - tt'| \leq \mathcal{D}(u, v)$$

□

The above lemma gives us a way to determine whether a system is free from same-task divergence. It turns out even when the classifiers do not initially satisfy the initial requirements, the lemma is still useful for detecting same-task divergence. Consider a set of classifiers $\mathcal{C}$ such that $\mathcal{C}' \subset \mathcal{C}$ that do not satisfy the requirements for Lemma 2.If we regroup the classifiers together as an "OR of ORs":

$$C_{a,b}(x) = \begin{cases} 1 \text{ if } \sum_{i \in \{a,\ldots,b\}} C_i(x) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

such that each grouped classifier has $\mathbb{E}[C_{a,b}(x)] \geq \frac{1}{2}$, we can now apply Lemma 2 to the *grouped* classifiers. This notions of "OR of ORs" can be extended to an "OR of heavy ORs" where "heavy" refers to having value one with probability at least $\frac{1}{2}$, in cases where each input to the "OR" is an arbitrary boolean function of classifier outcomes. We expect in practice that this test will be easier to implement than having to fully consider the set of classifiers.

3.1.2 Multiple-Task Functional Composition

From the previous section we found that functional compositions for a single task may violate individual fairness, despite each classifier being fair-in-isolation. We extend this notion to a multiple-task setting, where we consider AND-fairness. Continuing from our college applicant problem, it may be the case to attend college, a student must be (1) accepted by the university *and* (2) be able to pay tuition, through family assistance, scholarships or university financial aid. While this appears to be a single-task problem, note that the task-metric for the problem is two-fold. The university admissions department (such as the one at Princeton University) may want to evaluate applications "need-blind" without considering financial metrics, while the financial aid office wants to maximize the number of students to attend from their pool of available aid.

To make more concrete this example, consider a case of 10 students, with identical academic qualifications, admitted to the university. Suppose two are able to attend without additional aid, three need some partial aid, and five are not able to attend without full tuition assistance from the university. If the university has some limited amount of funding, from their perspective, they could fairly allocate by offering full financial aid package with some probability $p \in (0, 1)$ to all students. However, this would mean students who require aid would be able to attend with probability p. While both admission and aid would be independently fair, we are faced with how to reconcile fairness for the system as a whole. It is not clear, which metric we should use to enforce fairness in this case, so we must turn our attention to systems where the relevant metric for the output is not a subset of the component metrics. Perhaps in this particular scenario, the relevant metric must be weighed more towards qualifications than financial need. Considering this problem, we present the definition of AND-fairness:

Definition 6. *Given a universe U and a set of k tasks $\mathcal{T}$ each with matrics $\mathcal{D}_1, ..., \mathcal{D}_k$,*

and output metric $\mathcal{D}^{\mathcal{O}}$, a system of classifierse $\mathcal{C}$ satisies ***AND-fairness*** *if the indicator*

$$\chi_u = \begin{cases} 1 \text{ if } \prod_{C_i \in \mathcal{C}} C_i(u) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

satisfies $\mathcal{D}^*(u,v) \geq d(\tilde{\chi_u}, \tilde{\chi_v})$ for all $u, v \in U$ where $\tilde{\chi_u} = \mathbb{P}[\chi_u = 1]$

As mentioned before, the output metric $\mathcal{D}^{\mathcal{O}}$ is case dependent, as there are situations where $\mathcal{D}^{\mathcal{O}} \in \mathcal{T}$ and others where $\mathcal{D}^{\mathcal{O}} \notin \mathcal{T}$. The next theorem provides some more insight into choosing such $\mathcal{D}^{\mathcal{O}}$, requiring that the output metric have strictly larger distances than all of the input metrics for all pairs, otherwise individual fairness can be violated in composition.

Theorem 3. *Let $\mathcal{T}$ be a set of k tasks with nontrivial metrics$\mathcal{D}_1, ..., \mathcal{D}_k$, and let $\mathcal{D}^{\mathcal{O}}$ represent the relevant outcome metric. If there exists at least one pair $u, v \in U$ and one pair of tasks T_i, T_j for $i, j \in [k]$ such that*

1. *$\mathcal{D}^{\mathcal{O}}(u,v) \leq \mathcal{D}_i(u,v), \mathcal{D}_j(u,v)$*
2. *$\mathcal{D}_i(u,v), \mathcal{D}_j(u,v) > 0$*

there exists a set of classifiers $\mathcal{C}$ that satsify individual fairness separately, but do not satisfy AND-fairness under composition

Proof. Pick $u, v \in U$, with positive distance for two tasks T_i, T_j, and $\mathcal{D}^{\mathcal{O}} \leq \mathcal{D}_i(u,v), \mathcal{D}_j(u,v)$. Choose an individually fair classifier C for task i such that $p_u - p_v = \mathcal{D}_i(u,v)$ (per Lemma 1). We demonstrate how to select p'_u, p'_v for the classifier C' for task j such that $|p'_u - p'_v| \leq \mathcal{D}_j(u,v)$ but the composition violates *AND*-fairness. Observe that the difference probability of positive classification under *AND*-fairness is equivalent to:

$$d_{AND}(u,v) = |p_u p'_u - p_v p'_v| \tag{3.1}$$

Per the constraints in the theorem, we have two possible cases:

1. Without loss of generality, $\mathcal{D}_i(u,v) > \mathcal{D}^{\mathcal{O}}(u,v)$. This case is immediate: if $|p_u - p_v| = \mathcal{D}_i(u,v) > \mathcal{D}^{\mathcal{O}}(u,v)$, we simply select $p'_v = p'_u = 1$ to violate *AND*-fairness with respect to $\mathcal{D}^{\mathcal{O}}$.

2. $\mathcal{D}_i(u,v) = \mathcal{D}_j(u,v) = \mathcal{D}^{\mathcal{O}}(u,v)$. We choose p'_u, p'_v such that $p'_u - p'_v = \mathcal{D}_j(u,v) = \mathcal{D}^{\mathcal{O}}(u,v)$. Rearranging 3.1, we have

$$d_{AND}(u,v) = p_u p'_u - p_v(p'_u - \mathcal{D}^{\mathcal{O}}(u,v))$$

Then we substitute our original choice for $p_u - p_v = \mathcal{D}_i(u,v) = \mathcal{D}^{\mathcal{O}}(u,v)$, to get

$$d_{AND}(u,v) = (p_v + \mathcal{D}^{\mathcal{O}}(u,v))p'_u - p_v(p'_u - \mathcal{D}^{\mathcal{O}}(u,v)) = p'_u\mathcal{D}^{\mathcal{O}}(u,v) + p_v\mathcal{D}^{\mathcal{O}}(u,v)$$

Thus, choosing p'_u such that $p'_u + p_v > 1$ is sufficient to violate the distance for *AND*-fairness with respect to $\mathcal{D}^{\mathcal{O}}$.

□

One consideration to note is that this theorem is a loose characterization of the cases that violate *AND*-fairness, specifically taking advantage of the distance metric between two classifiers.

3.2 Task-Competitive Composition

In this section, we study cases where multiple tasks are decided by multiple classifiers. First, we need to make concrete the definition of fairness on systems that affect outcomes for different tasks.

3.2.1 Preliminary Definitions

Definition 7. *For a set of k tasks $\mathcal{T}$ with metrics $\mathcal{D}_1, ..., \mathcal{D}_k$, a system of classifiers $\mathcal{S} : U \times r \to \{0,1\}^k$¡ which assigns ouptuts a k-tuple corresponding to the output of task i in the i-th coordinate, satisfies* ***Multiple Task Fairness*** *if for all $i \in [k]$ and for all $u, v \in U$*

$$\mathcal{D}_i(u, v) \geq |\mathbb{E}[\mathcal{S}_i(u)] - \mathbb{E}[\mathcal{S}_i(v)]| \tag{3.2}$$

where $\mathbb{E}[\mathcal{S}_i(u)]$ is the expected outcome for the i-th task in the system $\mathcal{S}$ with expectation ranges the randomness of the system and its components

In contrast to multiple-task *functional* composition, enforcing a strict *multiple-task composition* makes sense when the tasks and their outcomes are distinct and incomparable. We illustrate this with an example: consider an advertising system that shows users ads for either tobacco and children's toys. If two users have similar intentions of purchasing tobacco, they should expect to see a similar proportion of ads for tobacco, irrespective of their preference for children's toys. In short, we do not want to allow positive distance for a task T_i to impact the increase in distance over outcomes for another task T_j.

With multiple possible outcomes, it is natural to consider the element of competition between classifiers. Clearly, if each classifier for a task independently and fairly assigns outputs, the system as a whole will satisfy multiple-task fairness. However, most systems require some degree of trade-offs between tasks; perhaps two advertisers must compete for a slot in a website to show an advertisement. We formalize the competition for a single-slot in *task-competitive composition* below:

Definition 8. *A system $\mathcal{S}$ is said to be a solution to a* ***Single-Slot Composition Problem*** *for a set of k tasks $\mathcal{T}$ with metrics $\mathcal{D}_1, ..., \mathcal{D}_k$, if for all $u \in U$, $\mathcal{S}$ assigns*

outputs for each task $\{y_{u,1}, ..., y_{u,k}\} \in \{0,1\}^k$ *such that*

$$\sum_{i \in [k]} y_{u,i} \leq 1$$

and $\forall i \in [k], \forall u, v, \in U$

$$\mathcal{D}_i(u, v) \geq |\mathbb{E}[y_{u,i}] - \mathbb{E}[y_{v,i}]|$$

Continuing with the advertisement example. The single-slot composition problem captures the scenario when a website may only have one slot to display an ad. Depending on the viewer, the system can choose to show either tobacco or children's toy advertisement, but need not choose either within the scope of the problem. Thus, to solve this problem, we must construct the system such that at most one task-outcome is chosen, while fairness remains preserved for all tasks in the universe.[2] Thus, we need first to consider the notion of breaking a tie.

Definition 9. *A* ***tie-breaking function*** $\mathbb{B} : U \times \{0,1\}^* \times \{0,1\}^k \to [k] \cup \{0\}$ *takes as input an individual* $u \in U$ *and a* k*-bit string* s_u *and outputs the index of a positive classification, 1, in* s_u *if such an index exists, and 0 otherwise.*

Note that $\mathbb{B}$ need not be neutral with respect to all individuals in the universe, as it takes into account $u \in U$ and thus, u's preferences into account. $\mathbb{B}$ can be randomized, although it is not required; with probability $p_{A,B}$, outcome A may be preferred to outcome B. When a preference probability is $0, 1$, we call it a *strict* preference. Also implicit in this definition is that when there is no tie required, the single positive classification is preferred. This is reasonable when both outputs are desirable, as in the advertising case since the platform prefers to generate revenue from displaying as many ads as possible. The tie-breaking function is broad enough to describe situations where ordering of decisions is based on a fixed policy or when

[2]As with our previous discussion of *OR*-fairness, the problem can be extended up to $k-1$ slots by modifying the definition, but we formally treat the single-slot instance in this section.

there exists time pressure to respond to one classifier before another.

Definition 10. *Consider a set $\mathcal{T}$ of k tasks and a tie-breaking function $\mathbb{B}$. Given a set $\mathcal{C}$ of classifiers for the set of tasks, define $y_u = \{y_{u,1}, ..., y_{u,k}\}$ where $y_{u,i} = C_i(u)$. The **Task-Competitive Composition** of the set $\mathcal{C}$ is defined as*

$$y_u^* = \mathbb{B}(u, y_u)$$

for all $u \in U$.

3.2.2 Unfairness in Task-Competitive Composition

Task-competitive composition can describe cases when classifiers are applied in strict ordering until a positive classification is obtained or where classifiers are applied together and a single output is chosen. The former may occur in credit or insurance applications to reflect the process of applying for different loans one at a time, where strict preference models ordering. In the case of advertising, the tie-breaking function can be used to express the probability that one advertiser outbids another, thereby winning the slot. In such multiple task settings, we denote by $\mathbb{B}_u(T)$ as the probability that task T is chosen when T, T' are options.

Before we address the problem of unfairness in a general sense. We consider a simple case in which all $u \in U$ have identical strict preferences for task T.

Lemma 3. *For any two tasks T, T' such that the non-trivial metrics for each task $\mathcal{D}, \mathcal{D}'$ respectively are not identical on a universe U, and if there is a strict preference for T, that is $\mathbb{B}_u(T) = 1 \forall u \in U$, then there exists a pair of classifiers C, C' which are individually fair in isolation, but violate multiple task fairness when combined in task-competitive composition.*

Proof. Recall that task-competitive composition ensures that at most one task can be classified positively for each element. Thus, our strategy to construct C, C' requires

that the distance between a pair of individuals is stretched for the 'second' task. By the non-triviality of $\mathcal{D}$, there exists u, v such that $\mathcal{D}(u, v) \neq 0$. Choose such a pair u, v and let p_u denote the probability that C assigns 1 to u, and similarly for p_v, p'_u, p'_v. For now, we treat these variables as placeholders and demonstrate how to set them to satisfy the lemma.

By the strict preference for T, the probabilities that U, v are assigned 1 for the T' can be written as:

$$\mathbb{P}[\mathcal{S}(u)_{T'} = 1] = (1 - p_u)p'_u$$
$$\mathbb{P}[\mathcal{S}(v)_{T'} = 1] = (1 - p_v)p'_v$$

The distance between the probability is the difference:

$$\begin{aligned}\mathbb{P}[\mathcal{S}(u)_{T'} = 1] - \mathbb{P}[\mathcal{S}(v)_{T'} = 1] &= (1 - p_u)p'_u - (1 - p_v)p'_v \\ &= p'_u - p_u p'_u - p'_v + p_v p'_v \\ &= p'_u - p'_v + p_v p'_v - p_v p'_u\end{aligned}$$

Observe that if $\mathcal{D}'(u, v) = 0$, which implies that $p'_u = p'_v$ and $p_u \neq p_v$,then this quantity is non-zero, yielding the desired contradiction for all fair C' and any C that assigns $p_u \neq p_v$, which can be constructed as per Corollary 1.2.

But if $\mathcal{D}'(u, v) \neq 0$, take C' such that $|p'_u - p'_v| = \mathcal{D}'(u, v)$ and call the distance $|p'_u - p'_v| = m'$. Without loss of generality, let $p'_u > p'_v, p_u < p_v$,

$$\mathbb{P}[\mathcal{S}(u)_{T'} = 1] - \mathbb{P}[\mathcal{S}(v)_{T'} = 1] = m' + p_v p'_v - p_u p'_u$$

Then to violate fairness for T', it suffices to show that $p_v p'_v > p_u p'_u$. Let $p_v = \alpha p_u$

where $\alpha > 1$.

$$\alpha p_u p'_v > p_u p'_u \qquad \alpha p'_v > p'_u$$

Now, it remains to show that we can choose p_u, p_v such that $\alpha > \frac{p'_u}{p'_v}$. By Corollary 1.3, we can choose p_u, p_v to obtain an sufficiently large $\alpha = \frac{p_u}{p_v}$, constrained only by the requirements $p_u < p_v$ and $|p_u - p_v| \leq \mathcal{D}(u, v)$. Therefore, we can find a pair of individually fair classifiers C, C', which when combined with strictly ordered task-competitive composition violate multiple task fairness. □

In a strictly ordered composition, the unfairness rises from the fact that each task inflicts its preferences on subsequent tasks, such as when an equal pair for the second task are unequal for the first. Once the first classifier acts, as long as the distance between the two is positive, the pair have unequal probabilities of even being considered by the second classifier, breaking down their original equality for that task for any classifier. Now, we extend Lemma 3 to the general case, where we do not require a strict preference, and see that unfairness problems persist in general settings as well.

Theorem 4. *For any two tasks T, T' with nontrivial metrics $\mathcal{D}, \mathcal{D}'$ respectively, there exists a set $\mathcal{C}$ of classifiers which are individually fair in isolation but when combined with task-competitive composition violate multiple task fairness for any tie-breaking function.*

Proof. Consider a pair of classifiers C, C' for the two tasks, T, T'. Define p_u, p_v, p'_u, p'_v analogously as before. Note that by our construction of $\mathbb{B}$, for each element $u \in U$, $\mathbb{B}_u(T) + \mathbb{B}_u(T') = 1$, as $\mathbb{B}_u(T)$ refers to the probability of choosing T over T' and $\mathbb{B}_u(T')$ vice versa. Suppose $\mathbb{B}_u(T) = 1$ or $\mathbb{B}_u(T') = 1 \forall u \in U$, then this case is identical to the one described in 3. Thus, we only need consider the following two cases:

1. $\mathbb{B}_u(T) = \mathbb{B}_v(T) \neq 1$. We can express the probabilities that each element is assigned to task T as:

$$\mathbb{P}[\mathcal{S}(u)_T = 1] = p_u(1 - p'_u) + p_u p'_u \mathbb{B}_u(T)$$
$$\mathbb{P}[\mathcal{S}(v)_T = 1] = p_v(1 - p'_v) + p_v p'_v \mathbb{B}_v(T)$$

such that the difference in probabilities after simplifying becomes:

$$\begin{aligned}\mathbb{P}[\mathcal{S}(u)_T = 1] - \mathbb{P}[\mathcal{S}(v)_T = 1] &= p_u(1 - p'_u) + p_u p'_u \mathbb{B}_u(T) - p_v(1 - p'_v) + p_v p'_v \mathbb{B}_v(T) \\ &= p_u - p_v + (p_v p'_v - p_u p'_u)(1 - \mathbb{B}_u(T))\end{aligned}$$

Since we assume that $\mathbb{B}_u(T) \neq 1$, we proceed analogously as the proof of Lemma 3 by choosing C' such that $p_v p'_v > p_u p'_u$ and choosing C such that $p_u - p_v = \mathcal{D}(u, v)$ to attain unfairness for T.

2. $\mathbb{B}_u(T) \neq \mathbb{B}_v(T)$. Without loss of generality, assume $\mathbb{B}_u(T) \neq 1$. From (1), recall that the difference in probability of assignment 1 for T is:

$$p_u - p_v + (p_v p'_v - p_u p'_u)(1 - \mathbb{B}_u(T))$$

Now, choose C such that $p_u - p_v = \mathcal{D}(u, v)$ [3] It suffices then to show that we can choose C' such that $p_v p'_v(1 - \mathbb{B}_v(T)) - p_u p'_u(1 - \mathbb{B}_u(T)) > 0$. Again, for $\alpha > 1$, let $p_u = \alpha p_v$. We need the following:

$$p_v p'_v(1 - \mathbb{B}_v(T)) > \alpha p_v p'_u(1 - \mathbb{B}_u(T))$$
$$p'_v(1 - \mathbb{B}_v(T)) > \alpha p'_u(1 - \mathbb{B}_u(T))$$

[3] In the general case that there is no such individually fair C, choose the individually fair C that maximizes the distance between u and v.

Let $\beta = (1 - \mathbb{B}_v(T))/(1 - \mathbb{B}_u(T))$, which is well-defined since $\mathbb{B}_u(T) \neq 1$, then the above becomes:

$$p'_v\beta > \alpha p'_u \implies \frac{\beta}{\alpha} > \frac{p'_u}{p'_v}$$

Since we are constrained only by $|p'_u - p'_v| \leq \mathcal{D}'(u, v)$, we can choose p'_u, p'_v to be any positive ratio by 1.3. Thus, we have shown that we can find a C' to exceed the allowed distance and violate multiple task fairness.

Thus, we have shown whether the tie-breaking functions are identical for u, v or not, there always exists a pair of classifiers C, C' that are fair in isolation, but when combined in task-competitive composition, do not satisfy multiple task fairness, which completes the proof. □

3.2.3 Building Fair Task-Competitive Composition

Now, we demonstrate how to fairly compose tasks for the single-slot composition problem. While the most obvious solution may be to remove the the conflict in tasks, classifying each task separately and deciding the outcome without influencing other tasks, but in practice, this is rarely feasible. Another approach could be to optimize the classifiers together with knowledge of the tie-breaking function, utility functions, and metrics. This would allow each classifier to respond accordingly, but also would require significant coordination from each deciding agent, which is impractical in application. Thus, we present a general mechanism for the single-slot composition problem which requires no additional information in learning each classifier nor additional coordination between the classifiers by using randomization.

Theorem 5. *For any set of k tasks $\mathcal{T}$ with metrics $\mathcal{D}_1, ..., \mathcal{D}_k$, the system $\mathcal{S}$ described in* RandomizeThenClassify *achieves multiple task fairness for the single-slot composition problem given any set of classifiers $\mathcal{C}$ which are individually fair in isolation.*

input: $u \in U$, set of individually fair classifiers $\mathcal{C}$ acting on U, probability distribution over tasks $\mathcal{Y} \in \Delta(\mathcal{C})$
y **to** $0^{|\mathcal{C}|}$;
$C_t \sim \mathcal{Y}$;
if $C_t(u) = 1$ **then**
  $x_t = 1$;
end
return x

Algorithm 2: RandomizeThenClassify

Proof. Consider the mechanism outlined in RandomizeThenClassify. For each element $u \in U$, the algorithm outputs a single positive classification by construction, as required by the single-slot composition problem. Observe that since the same probability distribution $\mathcal{Y}$ and classifiers $\mathcal{C}$ are used, each element has equal probability of having task T selected, resulting in fair classifications for T. Thus, the probability of positive classification in any task is $\mathbb{P}[C_t \sim \mathcal{Y} = T] \cdot \mathbb{P}[C_T(u) = 1]$. Then the difference in probability of positive classification for a task T between $u, v \in U$ is

$$\begin{aligned}
&\mathbb{P}[C_t \sim \mathcal{Y} = T] \cdot \mathbb{P}[C_T(u) = 1] - \mathbb{P}[C_t \sim \mathcal{Y} = T] \cdot \mathbb{P}[C_T(v) = 1] \\
&= \mathbb{P}[C_t \sim \mathcal{Y} = T](P[C_T(u) = 1] - \mathbb{P}[C_T(v) = 1]
\end{aligned}$$

which immediately fairness by the fact that C_T is individually fair in isolation, as required. □

The algorithm RandomizeThenClassify requires no coordination in the training of the classifiers, and specifically does not require any sharing of objective functions by means of randomization. Furthermore, it retains the ordering of elements by each classifier: if $\mathbb{P}[C_i(u) = 1] > \mathbb{P}[C_i(v) = 1]$, then $\mathbb{P}[\text{RandomizeThenClassify}(u)_i = 1] > \mathbb{P}[\text{RandomizeTheClassify}(v)_i = 1]$. And due to the lack of need for coordination among classifiers, it is easy to implement in the context of the single-slot composition problem.

3.2.4 Summary

We have shown in this chapter that the naive composition of multiple individually fair classifiers may result in unfairness as a whole. While our characterization is not exhaustive, we now have some understanding about how such systems of compositions must interact. In particular, in same-task functional compositions, we have shown that in settings like college applications, where the number of classifiers is small, each entity acting fairly in isolation is not enough to guarantee fairness. In the case of multiple-task compositions, our results still require further research, but roughly are in-line with our intuition that in scenarios where individual metrics do not compose well, such as the ability to afford tuition and admission into university, it may be difficult to ensure fairness as a whole. However, our study is not without promise, as our results indicate that many systems where the number of classifications is large or repeated are less likely to suffer from unfairness, assuming the absence of external sources of bias. Furthermore, in the single-slot task-competitive composition, we have shown that the algorithm RandomizeThenClassify can be used to correct unfair systems composed of individually fair classifiers. While a downside of this mechanism is that it may reduce the number of positive allocations, it gives us further room for research to find alternative algorithms to correct for unfairness in composition.

Chapter 4

Extensions to Group Fairness

The examples discussed in the previous chapter provide some intuition that solutions to compositional problems are not easy to construct without coordination between classifiers or situationally tailored compositions. In this chapter, we extend the composition results to group fairness notions, following the generalized form of *Statistical Parity*, which we will call *Conditional Parity*.

Definition 11. *A classifier C satisfies* ***Conditional Parity*** *with respect to a stratification set $\mathcal{X}$ and for sensitive attributes $\mathcal{A}$, if for all $a_1, a_2 \in \mathcal{A}$ and for all $x \in X$:*

$$\mathbb{P}[C = 1 \mid \mathcal{A} = a_1, \mathcal{X} = x] = \mathbb{P}[C = 1 \mid \mathcal{A} = a_2, \mathcal{X} = x]$$

Note that this definition is similar to the statistical parity criterion, with the inclusion of the condition on the set of non-sensitive stratification set X. In our discussion, we say that a classifier "satisfies group fairness" if it satisfies the above definition.

To extend our results of individual fairness composition to group fairness composition, we first show that many classifiers satisfy conditional parity in isolation, but do not under composition.

4.1 Functional Compositions

We find that the results from individual fairness for functional compositions largely extend in the group fairness setting, with a few adjustments to the definitions.

Same-Task Functional Composition

Consider a pair of classifiers C, C' which both satisfy conditional parity with respect to the same set of sensitive attributes and stratification set:

$$\mathbb{P}[C(u) = 1 | \mathcal{A} = a_1, \mathcal{X} = x] = \mathbb{P}[C(u) = 1 | \mathcal{A} = a_2, \mathcal{X} = x]$$
$$\mathbb{P}[C'(u) = 1 | \mathcal{A} = a_1, \mathcal{X} = x] = \mathbb{P}[C'(u) = 1 | \mathcal{A} = a_2, \mathcal{X} = x]$$

where the probability is taken over u and randomness of the classifier C, for all $a_1, a_2 \in \mathcal{A}, x \in \mathcal{X}$. Continuing from our college admissions example, $\mathcal{A}$ may be the set of genders, and $\mathcal{X}$ are the set of qualifications for admission. Suppose we apply C, C' to all members of the universe. Then we can write the conditional partiy constraints as sums over all individuals. For sake of notational clarity, let $U_{a_i,x}$ denote all $u \in U$ with $\mathcal{A} = a_i, \mathcal{X} = x$, and as before, let p_u be the probability that $C(u) = 1$.

$$\frac{1}{|U_{a_1,x}|} \sum_{u \in U_{a_1,x}} p_u = \frac{1}{|U_{a_2,x}|} \sum_{u \in U_{a_2,x}} p_u \tag{4.1}$$
$$\frac{1}{|U_{a_1,x}|} \sum_{u \in U_{a_1,x}} p'_u = \frac{1}{|U_{a_2,x}|} \sum_{u \in U_{a_2,x}} p'_u \tag{4.2}$$

Recall the *OR*-fairness requirement; summing over the universe in a similar fashion we have:

$$\frac{1}{|U_{a_1,x}|} \sum_{u \in U_{a_1,x}} \mathbb{P}[C(u) = 1 \vee C'(u) = 1] = \frac{1}{|U_{a_2,x}|} \sum_{u \in U_{a_2,x}} \mathbb{P}[C(u) = 1 \vee C'(u) = 1] \tag{4.3}$$

Since randomness of the classifiers is independent, we can rewrite (4.3) as:

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u + (1-p_u)p'_u = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u + (1-p_u)p'_u$$

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u - p'_u + p_u p'_u = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u - p'_u + p_u p'_u \tag{4.4}$$

As C, C' both satisfy conditional parity. Thus, combining (4.1), (4.2) with (4.4), we get:

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u p'_u = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u p'_u \tag{4.5}$$

Thus, our goal reduces to characterizing when (4.1), (4.2) imply (4.5). In other words, if classifiers independently satisfy conditional parity, we want to describe when they imply the composed *OR*-fairness conditional parity condition.

There are two cases for us to consider: when elements with $\mathcal{X} = x$ are treated the same and when they are not. First consider the former: any classifier that treats all elements with identical the qualification set equally satisfies conditional parity since (5.3) is satisfied, as all $u \in U_{a_1,x} \cup U_{a_2,x}$ are classified positively with probability p_u.

Proposition 6. *If a pair of classifiers C, C' treat all elements with $\mathcal{X} = x$ identically, and if C, C' satisfy conditional parity in isolation, then the "OR" of the two satisfies conditional parity under same-task composition.*

However, in the latter case, when elements with $\mathcal{X} = x$ are not treated equally, there is no guarantee that conditional parity will be satisfied under *OR* composition. Consider a simple example based on college admissions. Suppose based on certain admission qualifications, each $x \in \mathcal{X}$ corresponds to a range of acceptance probabilities. For simplicity, suppose we have p_h, p_m, p_l corresponding to high, medium, low likelihood of admission. Each group may consist of different mix of individuals mapped to p_h, p_m, p_l. Let group A only contain medium qualified applicants mapped to p_m, and

group B be composed of half highly qualified and half with low qualification. For a single classifier or university, choosing $p_h = 0.9, p_m = 0.5, p_l = 0.1$, conditional parity is satisfied as $\frac{p_h+p_l}{2} = p_m$. But note that in composition with two university applications, the squares in each group diverge: $1-(1-p_m)^2 \neq \frac{1}{2}(1-(1-p_h)^2)+\frac{1}{2}(1-(1-p_l)^2)$, violating conditional parity in composition. To satisfy conditional parity under *OR*-composition, one could treat all individuals with $\mathcal{X} = x$ equally as before, but this ignores other information about individuals in group A for the sake of technicality and may sacrifice accuracy.

4.2 Multiple Task Composition

We discussed two types of multiple-task composition: functional composition and single-slot task-competitive composition. When extended to group fairness, the results in the single-outcome setting are similar to the multiple outcome setting, thus we focus our discussion onto the latter. As before, we show how to extend the results from individual fairness to show that for a set of multiple tasks and tie-breaking functions, classifiers which independently satisfy group fairness can result in unfair task-competitive compositions.

4.2.1 Extension of Single-Slot Composition Results

Consider two tasks T, T', with sensitive attribute sets A, A' and stratification sets $\mathcal{X}, \mathcal{X}'$, composed under task-competitive composition with a tie-breaking function $\mathbb{B}$ for the single-slot decision problem. Recall the advertising problem with tobacco and children's toy ads, and suppose our sensitive attributes are race and gender, and the stratification sets are the respective interests purchasing either of the product groups. For an individual $u \in U$, let the p_u, p'_u denote the probabilities of positive classification, and $\mathbb{B}_u(T)$ the preferential probability of choosing T over T'. For a

system to satisfy multiple task conditional parity, we need the probabilities of positive classification to satisfy for all $a_1, a_2 \in \mathcal{A}, x \in \mathcal{X}$ and respectively for $\mathcal{A}', \mathcal{X}'$:

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u(1-p'_u) + p_u p'_u \mathbb{B}_u(T) = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u(1-p'_u) + p_u p'_u \mathbb{B}_u(T)$$
$$\frac{1}{|U_{a'_1,x'}|}\sum_{u\in U_{a'_1,x'}} p'_u(1-p_u) + p_u p'_u \mathbb{B}_u(T') = \frac{1}{|U_{a'_2,x'}|}\sum_{u\in U_{a'_2,x'}} p'_u(1-p_u) + p_u p'_u \mathbb{B}_u(T')$$

As the classifiers individually satisfy conditional parity, we can simply the constraints with (4.1) and (4.2) to yield:

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u p'_u(\mathbb{B}_u(T)-1) = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u p'_u(\mathbb{B}_u(T)-1) \tag{4.6}$$
$$\frac{1}{|U_{a'_1,x'}|}\sum_{u\in U_{a'_1,x'}} p_u p'_u(\mathbb{B}_u(T')-1) = \frac{1}{|U_{a'_2,x'}|}\sum_{u\in U_{a'_2,x'}} p_u p'_u(\mathbb{B}_u(T')-1) \tag{4.7}$$

To show that multiple conditional parity does not hold in task-competitive composition, it suffices to show that we can construct C, C' such that (4.6) and (4.7) do not hold.

Theorem 7. *For any two tasks T, T' with sensitive attribute sets $\mathcal{A}, \mathcal{A}'$ and qualification sets $\mathcal{X}, \mathcal{X}'$ such that $\cup_{x\in\mathcal{X}} u_x = U_x$ and $\cup_{x'\in\mathcal{X}'} u_x = U_{x'}$, and a tie-breaking function $\mathbb{B}$ such that there exist $a_1, a_2 \in \mathcal{A}, x \in \mathcal{X}, x' \in \mathcal{X}'$ such that at least one of $U_{a_1,x} \cap U_{x'}$ and $U_{a_2,x} \cap U_{x'}$ is nonempty, and*

$$\frac{1}{|U_{a'_1,x'}|}\sum_{u\in U_{a_1,x}\cap U_{x'}} \mathbb{B}_u(T) \neq \frac{1}{|U_{a'_2,x'}|}\sum_{u\in U_{a_2,x}\cap U_{x'}} \mathbb{B}_u(T)$$

there exists a pair of classifiers C, C' which satisfy conditional parity in isolation but do not under task-competitive composition.

Proof. We construct C, C' such that each element with $\mathcal{X} = x$ and each element with $\mathcal{X}' = x'$ is treated identically under C, C' respectively, and suppose every element

has probability of positive classification < 1. Suppose that $C,' C$ satisfy multiple conditional parity under task-competitive composition, then by (4.6) and (4.7) we should have

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p_u p'_u(\mathbb{B}_u(T)-1) = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p_u p'_u(\mathbb{B}_u(T)-1)$$
$$\frac{1}{|U_{a'_1,x'}|}\sum_{u\in U_{a'_1,x'}} p_u p'_u(\mathbb{B}_u(T')-1) = \frac{1}{|U_{a'_2,x'}|}\sum_{u\in U_{a'_2,x'}} p_u p'_u(\mathbb{B}_u(T')-1)$$

By construction, we have that each element u with $\mathcal{X} = x$ is treated identically, thus has the same classification probability p_u. So we can simply the above to obtain:

$$\frac{1}{|U_{a_1,x}|}\sum_{u\in U_{a_1,x}} p'_u(\mathbb{B}_u(T)-1) = \frac{1}{|U_{a_2,x}|}\sum_{u\in U_{a_2,x}} p'_u(\mathbb{B}_u(T)-1) \tag{4.8}$$

We can rewrite (4.8) in terms of intersecting sets in $\mathcal{X}'$, where $p'_{x'}$ denotes the probability of positive classification for elements with $\mathcal{X}' = x'$ under C'

$$\frac{1}{|U_{a_1,x}|}\sum_{x'\in\mathcal{X}'}\sum_{u\in U_{a_1,x}\cap U_{x'}} p'_{x'}(\mathbb{B}_u(T)-1) = \frac{1}{|U_{a_2,x}|}\sum_{x'\in\mathcal{X}'}\sum_{u\in U_{a_2,x}\cap U_{x'}} p'_{x'}(\mathbb{B}_u(T)-1) \tag{4.9}$$

However, by assumption there exists a_1, a_2 and $x' \in \mathcal{X}'$ such that

$$\frac{1}{|U_{a'_1,x'}|}\sum_{u\in U_{a_1,x}\cap U_{x'}} \mathbb{B}_u(T) \neq \frac{1}{|U_{a'_2,x'}|}\sum_{u\in U_{a_2,x}\cap U_{x'}} \mathbb{B}_u(T)$$

which breaks the equality in (4.8) and thus (4.9), completing the proof. □

Thus, group fair classifiers in composition under task-competitive setting can fail to satisfy multiple conditional parity. However, as before, the algorithm RandomThenClassify can remedy this issue.

Theorem 8. *For any two tasks T, T' with $\mathcal{A}, \mathcal{A}'$ and $\mathcal{X}, \mathcal{X}'$ respectively, if classifiers C, C' satisfy conditional parity in isolation but not under task-competitive composi-*

tion, applying RandomThenClassify *on the classifiers will satisfy conditional parity in composition.*

Proof. Applying RandomThenClassify, the probability that the composed system classifies u positively for task T is

$$\mathbb{P}[S(u)_T = 1] = \mathbb{P}[C_t \sim \mathcal{Y} = T] \cdot \mathbb{P}[C(u) = 1]$$

By the randomness of the algorithm, the probability that $C_t \sim \mathcal{Y}$ is the same for all $a \in \mathcal{A}, x \in \mathcal{X}$, and since the original classifier satisfies conditional parity, we have $\forall a_1, a_2 \in \mathcal{A}, \forall x \in \mathcal{X}$:

$$\mathbb{P}[S(u)_T = 1 \,|\, \mathcal{A} = a_1, \mathcal{X} = x] = \mathbb{P}[S(u)_T = 1 \,|\, \mathcal{A} = a_2, \mathcal{X} = x]$$

The proof for $S(u)_{T'}$ follows analogously. □

Hence, the same strategy to fix unfairness in task-competitive composition for individual fairness is also effective in group fairness settings.

Chapter 5

Conclusion

Much of past work on fairness in algorithmic systems has been very broad. From a brief glance, our immediate understanding of fairness does not seem to compose well with machine learning, where we are often concerned with some objective. Maximizing revenue, optimizing for performance, etc., become the primary concern. However, fairness is essential, if our society is to be more and more automated, with decisions being informed by statistical measures and classification. There is a social concern to prevent the repeat discrimination for protected groups, such as race or gender, for which we have suffered the costs in the past. But also it is worth studying how we can create objective systems that can promote widespread welfare and maximize social utility as a whole.

In this thesis, We have provided an overview of current prevailing notions of fairness. We demonstrated that group fairness criteria may not be rich enough to quantify and study fairness. Thus, we turned our attention to the notion of individual fairness and showed how to construct individually fair classifiers given a complete distance metric. This turned out to be very useful when studying systems of compositions of classifiers, where the individual fairness criterion could be used as a standard to see if multiple classifiers violated or preserved fairness in composition.

Interestingly, we showed results that compositions of systems built from individually fair classifiers in isolation do not necessarily result in systems that are fair in composition. Specifically, we analyzed the existence of such pitfalls for *functional composition* and *task-competitive composition*. In general, there is no guarantee of fairness under composition, for either individual fairness. However, our results were not without promise. We also provided means of detecting if a system was free from same-task divergence in a same-task functional composition and analyzed the use of randomness to correct an unfair system. Finally, we extended the results of individual fairness in composition to the generalized group fairness criterion of conditional parity, and found that conditional parity experience similar pitfalls of naive composition, but fairness violations can still be corrected.

There still remains work to be done on fairness in classification. Much of our work was focused on scenarios where classifiers act on the entire universe of individuals with each individual's outcome being decided independently. However, many applications this is not the case, as classifiers may be acting as a selection mechanism, in some arbitrary order, or in ways where the outcomes received by the individuals are dependent. Unlike other social concerns in technology, such as invasions of privacy, fortunately, fairness is correctable, and repairable after the fact, to build systems that are more and more fair. This gives us promise that we can find more mechanisms for fair composition than the ones introduced in this thesis.

Appendix A

Independence vs. Sufficiency

We show that in general the group fairness criteria *independence* and *sufficiency* are mutually exclusive. We make the assumption that the sensitive characteristics and the target outcome Y are not independent, which is to say that group membership does have some effect on the distribution of the target variable.

Proposition 9. *Suppose A and Y are not independent, then independence and sufficiency cannot both hold*

Proof. Suppose both independence and sufficiency hold, then we have

$$A \perp C \text{ and } A \perp Y \,|\, C \Rightarrow A \perp (Y, C) \Rightarrow A \perp Y$$

a contradiction to the assumption that A and Y are independent. □

Independence vs. Separation

A similar mutual exclusion holds for *independence* and *separation*. In this case, we make the restriction for $Y \in \{0, 1\}$, and that the predictor C is not independent of the target Y. The latter assumption is not out of ordinary, as any useful predictive measure should be correlated with the target.

Proposition 10. *For $Y \in \{0,1\}$, A not independent of Y, and C not independent of Y, independence and separation are mutually exclusive.*

Proof. Let $Y \in \{0,1\}$. We prove the contrapositive:

$$A \perp C \text{ and } A \perp C \,|\, Y \implies A \perp Y \text{ or } C \perp Y$$

By the definition of conditional probability we have:

$$\mathbb{P}[C = c \,|\, A = a] = \sum_y \mathbb{P}[C = c \,|\, Y = y]\mathbb{P}[Y = y \,|\, A = a] \tag{1}$$

Since by assumption, $A \perp C, A \perp C \,|\, Y$ respectively, (A.1) simplies to two equations:

$$\mathbb{P}[C = c] = \sum_y \mathbb{P}[C = c \,|\, Y = y]\mathbb{P}[Y = y \,|\, A = a] \tag{2}$$

$$\mathbb{P}[C = c] = \sum_y \mathbb{P}[C = c \,|\, Y = y]\mathbb{P}[Y = y] \tag{3}$$

From (A.2) and (A.3) we get:

$$\sum_y \mathbb{P}[C = c \,|\, Y = y]\mathbb{P}[Y = y \,|\, A = a] = \sum_y \mathbb{P}[C = c \,|\, Y = y]\mathbb{P}[Y = y]$$

Let $p = \mathbb{P}[Y = 1]$, $p_a = \mathbb{P}[Y = 1 \,|\, A = a]$, and $c_y = \mathbb{P}[C = c|Y = y]$, then the above simplies to:

$$pr_1 + (1-p)r_0 = p_a r_1 + (1-p_a)r_0 \implies p(r_1 - r_0) = p_a(r_1 - r_0)$$

Observe that this holds if $r_0 = r_1$, which implies that $C \perp Y$ or if $p = p_a$ for all $a \in A$, which implies that $A \perp Y$, and the proof is complete. □

Separation vs. Sufficiency

Here we see that *separation* and *sufficiency* imposed together only results in degenerate cases. For sake of practicality, suppose we have a binary classifier with nonzero false positive rate (FPR) and true positive rate (TPR).

Proposition 11. *Assume Y is not independent of A, and let C be a binary classifier with nonzero FPR and TPR, then separation and sufficiency cannot both hold.*

Proof. As $A \not\perp Y$, for two groups $a, b \in A$, we have

$$p_a = \mathbb{P}[Y = 1 \mid A = a] \neq \mathbb{P}[Y = 1 \mid A = b] = p_b$$

Assume separation holds, then since the classifier is imperfect, all groups have the same nonnegative FPR, and the same postivive true positive rate TPR. Recall, for a binary classifier, sufficiency requires all groups have the same positive predictive value (PPV). The PPV for group a can be expressed as:

$$PPV_a = \frac{TPR \cdot p_a}{TPR \cdot p_a + FPR \cdot (1 - p_a)}$$

In order for $PPV_a = PPV_b$, only if either $TPR = 0$ or $FPR = 0$, neither of which can hold due to our assumption. Thus, separation and sufficiency imposed together overconstrain the solution space leaving only degenerate cases. □

Bibliography

[1] C. Dwork and C. Ilvento, "Fairness Under Composition," pp. 1–72, 2018.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2018. http://www.fairmlbook.org.

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," 2011.

[4] S. Verma and J. Rubin, "Fairness Definitions Explained," pp. 1–7, 2018.

[5] R. Bartlett, U. C. Berkeley, A. Morse, U. C. Berkeley, R. Stanton, and N. Wallace, "Consumer-Lending Discrimination in the Era of FinTech *,"

[6] P. Shi, "Prediction and Optimization in School Choice," no. 2010, 2016.

[7] C. Prendergast, "The Allocation of Food to Food Banks," 2015.

[8] P. A. Pathak and T. Sönmez, "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review*, vol. 98, no. 4, pp. 1636–1652, 2008.

[9] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," no. Ec, 2018.