Fairness in selection problems

Vitalii Emelianov

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Vitalii EMELIANOV

Thèse dirigée par **Patrick LOISEAU**, Chercheur, Inria et
codirigée par **Nicolas GAST**, Chercheur, Inria

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Equité dans les problèmes de sélection

Fairness in selection problems

Thèse soutenue publiquement le **13 juin 2022**,
devant le jury composé de :

Monsieur Manuel Gomez Rodriguez
Professeur, Max Planck Institute for Software Systems, Rapporteur
Madame Nicole Immorlica
Chercheuse, Microsoft Research, Membre
Monsieur Jean-Michel Loubes
Professeur, Université Toulouse Paul Sabatier, Rapporteur
Madame Marie-Christine Rousset
Professeure, Université Grenoble Alpes, Présidente
Monsieur Alexis Tsoukiàs
Professeur, Université Paris Dauphine, Membre
Monsieur Nicolas Usunier
Chercheur, Facebook, Membre

Acknowledgements

This work would not have been possible without the support of my supervisors, Patrick Loiseau and Nicolas Gast. I would like to thank you for your scientific guidance, your patience, and care throughout the journey.

I would also like to thank my Ph.D. defense committee: Alexis Tsoukiàs, Jean-Michel Loubes, Manuel Gomez Rodriguez, Marie-Christine Rousset, Nicolas Usunier, and Nicole Immorlica. It was a great pleasure and quite a unique chance to discuss my work with such amazing researchers.

I thank my colleagues at LIG for their scientific and moral support. Additionally, special thanks to Inria, who financed my research for almost four years.

Finally, I wish to thank my friends and family, especially my parents and Kamilia, who always believe in me and can find the best words of encouragement, giving me the strongest motivation to continue.

Abstract

Data-driven decision-making algorithms are increasingly applied in many domains with high social impact, such as hiring, lending, or criminal justice. Recently, it was shown that such algorithms could lead to discrimination against certain demographic groups (e.g., they can discriminate by race, gender, or age). This led to a recent active line of research—called *algorithmic fairness*—which studies how to develop efficient algorithms with fairness guarantees. Most of the decision problems with high social impact mentioned above are essentially selection problems. In selection problems, the decision-maker must select a *fixed fraction* of the best candidates given their characteristics. The notion of a selection budget contrasts selection problems with classification problems typically studied in the algorithmic fairness literature.

In this thesis, we study the causes of discrimination in selection problems and the impact of fairness mechanisms on the utility of selection. Our first contribution considers a selection problem with candidates whose quality is measured with a group-dependent noise—a phenomenon called *differential variance*. We study the impact of differential variance on group representations and how standard group fairness mechanisms affect the selection utility in the presence of differential variance. Our second contribution proposes a game-theoretic model of a selection problem with differential variance. We assume *strategic candidates* who maximize the individual utility by making a costly effort. The effort induces their quality, measured by a (group-fair) decision-maker with group-dependent noise. We characterize the equilibrium of such a game. In our third contribution, we consider a multistage selection problem. We extend classical group fairness notions to a multistage setting and propose the notions of local (per stage) and global (final stage) fairness. We then introduce and study the *price of local fairness* which is the ratio of utilities induced by the globally fair algorithm to that of the locally fair algorithm.

Résumé

Les algorithmes de prise de décision basés sur des données sont de plus en plus appliqués dans de nombreux domaines à fort impact social, tels que l'embauche, le crédit ou la justice pénale. Récemment, il a été démontré que ces algorithmes pouvaient entraîner une discrimination à l'encontre de certains groupes démographiques (par exemple, une discrimination fondée sur la race, le sexe ou l'âge). Cette constatation a donné naissance à une ligne de recherche active récente, appelée *équité algorithmique*—qui étudie comment développer des algorithmes efficaces avec des garanties d'équité. La plupart des problèmes de décision à fort impact social mentionnés ci-dessus sont essentiellement des problèmes de sélection. Dans les problèmes de sélection, le décideur doit sélectionner une *fraction fixe* des meilleurs candidats compte tenu de leurs caractéristiques. La notion de budget de sélection contraste les problèmes de sélection avec les problèmes de classification typiquement étudiés dans la littérature sur l'équité algorithmique.

Dans cette thèse, nous étudions les causes de la discrimination dans les problèmes de sélection et l'impact des mécanismes d'équité sur l'utilité de la sélection. Notre première contribution considère un problème de sélection avec des candidats dont la qualité est mesurée avec un bruit dépendant du groupe—un phénomène appelé *variance différentielle*. Nous étudions l'impact de la variance différentielle sur les représentations du groupe et comment les mécanismes standards d'équité de groupe affectent l'utilité de la sélection en présence de variance différentielle. Notre deuxième contribution propose un modèle de théorie des jeux d'un problème de sélection avec variance différentielle. Nous supposons des *candidats stratégiques* qui maximisent l'utilité individuelle en faisant un effort coûteux. L'effort induit leur qualité, mesurée par un décideur (juste de groupe) avec un bruit dépendant du groupe. Nous caractérisons l'équilibre d'un tel jeu. Dans notre troisième contribution, nous considérons un problème de sélection en plusieurs étapes. Nous étendons les notions classiques d'équité de groupe à un cadre à plusieurs étapes et proposons les notions d'équité locale (par étape) et globale (étape finale). Nous introduisons et étudions ensuite le *prix de l'équité locale* qui est le rapport des utilités induites par l'algorithme globalement équitable à celui de l'algorithme localement équitable.

Contents

Chapter 1

Introduction

1.1 Motivation and Context

The use of machine learning algorithms is increasingly omnipresent. These algorithms find applications in recommendation systems, fraud detection, and language translation. Learning algorithms can outperform human experts at some tasks both in terms of speed and accuracy, for example, in face and object recognition (Dooley et al., 2021; Geirhos et al., 2018).

Machine learning algorithms are also used in decision-making problems with high social impact, such as hiring, college admission, lending, or criminal justice (Finocchiaro et al., 2021). Ethics is an essential component of such decisions as it directly affects how societies work and develop. There is a lot of evidence that human decision-makers, implicitly or explicitly, can make biased decisions based on salient demographic attributes of individuals such as gender (Ahmed et al., 2021; Human Rights Watch, 2020), race (Bertrand and Mullainathan, 2004; Gersen, 2019; Larson et al., 2016; Quillian et al., 2017) or age (Cohen, 2019). The rise of algorithmic decision-making was believed to solve the problem of discrimination as machines are neutral—their decisions cannot be governed by prejudices like for humans. However, we frequently observe that the algorithms tend to have the same biases as the humans have (Lambrecht and Tucker, 2018; Larson et al., 2016). A recent line of research on the fairness of algorithms—called *algorithmic fairness*—studies the solutions to the problem of discrimination caused by automated decision-making systems (Barocas et al., 2019).

The algorithmic fairness literature proposes several fairness definitions and studies how to design efficient algorithms that satisfy these fairness notions. In this literature, fairness is typically considered as an additional constraint that introduces a tradeoff between fairness metrics and prediction accuracy. Most works in algorithmic fairness focus on *classification problems*: given the characteristics of individuals, the goal is to train an algorithm (based on historical data) that can correctly assign a *class label* to each individual. For example, in lending, a bank is interested in estimating the probability of an individual's default. Hence, given the characteristics of an individual (e.g., income, marital status, etc.), an algorithm must assign a label (e.g., repay or default) with as

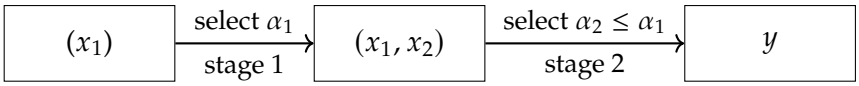


Figure 1.1 – **Illustration of a two-stage selection process.** At the first stage, an attribute x_1 of individuals is observed, and α_1 proportion of them is selected. At the second stage, an additional attribute x_2 is observed, and the α_2 of candidates is subselected. The true quality y of an individual is observed only after the final stage decision. The decision-maker aims to maximize the expected quality of selected candidates.

high accuracy as possible while providing some fairness guarantees.

Many decision-making problems with high social impact mentioned above are essentially *selection problems*. For example, in college admissions, the number of admitted candidates is limited by the capacity of a school. Similarly, in lending, the number of loans should meet the budget of a bank. In the general formulation of a selection problem, a decision-maker is given a number of candidates and their characteristics, and they must select a *fixed* fraction of the best of them. Selection is also often performed in multiple stages: at each stage, the decision-maker learns more information about the selected candidates, for example, by conducting consecutive interviews that evaluate different candidates' expertise. At the same time, the budget constraints, like time or the number of interviewers, do not allow the decision-maker to select all candidates at each stage. Hence, the selection procedures assume obtaining more and more information and selecting fewer candidates with each stage (see Fig. 1.1). Surprisingly, the selection problem is rarely considered in the algorithmic fairness literature, which raises the following questions:

> What is the **cause of discrimination** in selection problems, and what is the **impact of fairness mechanisms** on the utility of selection?

In this thesis, we focus on simple theoretical models that provide answers to the above questions. We also illustrate our results by experimenting on real data. We believe that the qualitative results given by our models can provide an aid to decision-makers in designing ethical selection procedures.

1.2 Main Questions and Contributions

We elaborate on three different aspects of the selection problem in our contributions. It is important to note that all three contributions study essentially the same selection problem shown in Fig. 1.1 but from different perspectives.

1.2.1 What is the impact of noise on fairness in selection problems?

The standard explanation for discrimination in selection problems is *implicit bias* (Kleinberg and Raghavan, 2018). The implicit bias model assumes that decision-makers have

internal preferences for some demographic groups of individuals compared to others; hence, they evaluate them as more qualified. In the economics literature, the authors identify another type of discrimination called *statistical discrimination* where the lack of information about individuals leads to discrimination (Arrow, 1973; Fang and Moro, 2011; Phelps, 1972). We argue that statistical discrimination is an equally (if not more) frequent in data-driven solutions to selection problems since having the biases is inefficient and can be fixed, while fixing the statistical discrimination is more difficult as it requires collecting more data.

In Chapter 3, to capture the effects of statistical discrimination on selection, we propose a model of selection problem where candidates' quality estimation is affected by a phenomenon called *differential variance*. In the model of differential variance, the candidates' quality is measured by *an unbiased but noisy* estimate which has differing noise variances for different demographic groups. In this setting, the qualities of individuals are fixed but observable by a decision-maker through these noisy estimates. We show that differential variance leads to the underrepresentation of the high-noise groups for all small-enough selection sizes if the decision-maker is Bayesian (i.e., knows the joint distribution of qualities and estimates). In contrast, we show that differential variance leads to the underrepresentation of the low-noise groups for all small-enough selection sizes if the decision-maker is group-oblivious (i.e., cannot use the group information). After, we study how some classical group fairness mechanisms—the demographic parity and the γ-rule[1]—affect the quality of selection. We show that the Bayesian decision-maker is optimal, hence, no fairness mechanism can lead to an increase in the selection quality; we also show a bound on the decrease of selection quality. Interestingly, we identify model parameters in which the aforementioned fairness mechanisms improve the selection quality if the decision-maker is group-oblivious; hence, there is no quality-fairness tradeoff for this decision-maker. This chapter is based on our publication (Emelianov et al., 2020) and its extended version (Emelianov et al., 2022a).

1.2.2 What is the impact of strategic behavior on fairness in selection problems?

In most of the works studying fairness in selection problems, including our previous contribution, it is assumed that individuals' characteristics are fixed and do not depend on the selection procedure (Celis et al., 2020; Kleinberg and Raghavan, 2018). However, there is a lot of evidence that individuals involved in selections can behave strategically if the decision rule is common knowledge: the individuals aim at maximizing the probability of a successful outcome by changing their attributes at some cost (Patro et al., 2022). For instance, in hiring, people try to work on a better CV to increase the chance of getting a good job; at the same time, they may not put more effort than needed for getting this job.

[1] A fairness mechanism that imposes that the ratio of group selection rates cannot be smaller than a predefined parameter $\gamma \in [0, 1]$. We discuss group fairness notions in greater detail in Chapter 2.

In Chapter 4, we propose a model of a selection problem with *strategic individuals* by extending the model of selection from Chapter 3. We assume that individuals obtain qualities by making an effort that maximizes the individual payoff. The payoff is equal to a fixed positive reward (e.g., a salary) if an individual is selected, minus the quadratic effort-dependent cost. The decision-maker who performs the selection is assumed Bayesian and chooses candidates based on their expected quality. This defines a population game between candidates, and we prove that there is a unique Nash equilibrium in this game. We study the two regimes of rewards for this game—small and large—and we identify that one of the groups is always underrepresented at equilibrium in both cases. For small rewards, we show that the high-noise group is underrepresented for all small-enough selection sizes (as in the non-strategic setting studied in Chapter 3). In contrast, as the reward goes to infinity, we show that the low-noise group is underrepresented for all selection sizes. Finally, we study how the demographic parity mechanism affects the selection quality. We show that it can lead to an improvement of selection quality in the strategic setting even for a Bayesian decision-maker who is optimal in the non-strategic setting (see Chapter 3). This chapter is based on our publication (Emelianov et al., 2022b).

1.2.3 What is the cost of per-stage fairness in multistage selection problems?

As we mentioned before, selection procedures are often performed in multiple stages where with each stage more information about the candidates is collected, but fewer candidates are selected. In hiring, for instance, decision-makers study the applicant's CVs at the first stage and invite the most qualified for a second-stage interview. The multistage nature of selection adds another dimension to the selection problem: First, it is unclear how to extend existing fairness notions to the multistage setting. Second, because some information becomes available only at later stages, there is a question on how the time at which the sensitive group information is available affects the utility/fairness of a selection.

In Chapter 5, we propose a model of *multistage selection* that captures the above effects. We extend two classical notions of fairness—demographic parity and equal opportunity—to a multistage setting. We call, thus, a *locally fair* an algorithm that satisfies per-stage fairness, and a *globally fair* an algorithm that satisfies only final stage fairness. We study, both theoretically and numerically, the ratio of the selection quality by the locally fair to that of a globally fair algorithm which we call *the price of local fairness*. First, we prove that the price of local fairness is bounded, i.e., per-stage fairness cannot harm much the selection quality compared to the final stage fairness. Second, we perform a series of experiments on synthetic and real data, in which we show that the later the sensitive attribute is observed, the larger the price of local fairness. In other words, decision-makers can impose local fairness if the sensitive attribute is observed early. This chapter is based on our publication (Emelianov et al., 2019).

Organization of the Manuscript We discuss the theoretical and empirical background on fairness and review the related work in Chapter 2. Our contributions are presented in Chapters 3, 4 and 5. We conclude with a discussion on implications and perspectives of the thesis in Chapter 6.

Chapter 2

Background and Related Work on Algorithmic Fairness and Fairness in Economics

Our work lies at the intersection of *algorithmic fairness* and *fairness in economics*. On the one hand, we study the problem of an optimal selection with fairness constraints. On the other hand, we consider the impact of selection procedures on individuals, especially on their *representation* in selection. Hence, we separate the related work into two main parts. In the first part (Sections 2.1, 2.2 and 2.3), we present the current progress of algorithmic fairness research. In the second part, we present the literature on the economic theories of discrimination (Section 2.4) and fairness in computational social choice (Section 2.5).

In this chapter, we review the related work in the space of fairness broadly to position the overall contribution of our thesis. We defer a more in-depth description of the most closely related work to each of the individual chapters dedicated to our three contributions, where we will be able to provide a more precise comparison.

2.1 Fairness in Learning

It is worth mentioning that algorithmic fairness is a relatively new research area. The first works studying fairness in learning problems are by Pedreshi et al. (2008) and Dwork et al. (2012). Since then, there have been many works on defining fairness and constructing algorithms that respect those definitions mainly for the case of *supervised learning* (Chouldechova, 2017; Corbett-Davies et al., 2017; Hardt et al., 2016b; Kleinberg et al., 2017; Lipton et al., 2018; Zafar et al., 2017b).

Supervised learning. In *supervised learning problems* (Hastie et al., 2001, p. 9), the goal is to find an algorithm that can correctly assign labels $\hat{y}$ to unseen data points x, where such an algorithm is trained on the historical labeled data $(x_i, y_i)_{i=1}^{l}$ drawn independently from some fixed distribution. If labels are continuous (e.g., $y \in \mathbb{R}$), then the corresponding problem is called *regression problem*. If labels are categorical (e.g., $y \in \{0, 1, \ldots, m\}$), then the corresponding problem is called *classification problem*.

Most of the works on fairness in supervised learning focus on studying binary classification problems, where $y \in \{0, 1\}$. Here $y = 1$ typically represents that an individual is good for the purpose of classification (repays the loan), and $y = 0$ represents that an individual is not good for the purpose of classification (defaults). Similarly, the binary decision $\hat{y} = 1$ means a positive outcome for an individual (accept the loan), whereas $\hat{y} = 0$ means a negative outcome for an individual (reject the loan). Individuals are represented by multidimensional vectors $x = (x_1, \ldots, x_d)$, where one of the attributes of such a vector is called a sensitive attribute that we denote by g. The sensitive attribute g represents the belonging of an individual to some demographic group, e.g., to a particular race, gender, or age. For simplicity of exposition, we assume only two demographic groups denoted by $g \in \{A, B\}$. Further, we will use capital letters $\hat{Y}$, Y, X and G to denote random variables corresponding to $\hat{y}$, y, x and g.

The authors in algorithmic fairness literature propose various fairness definitions that are typically based on equating some combinations of the classification outcome for the demographic groups. For example, the classical **demographic parity** criterion ensures that positive outcome must occur with equal probability for all demographic groups:

$$\mathrm{P}(\hat{Y} = 1|G = A) = \mathrm{P}(\hat{Y} = 1|G = B).$$

Note that the demographic parity requires an exact equality of the acceptance probabilities. In reality, it is sometimes hard to satisfy the exact equality, and a relaxed version of the above definition is more desirable. For example, if we assume that the acceptance rate ratio should not be inferior to some fixed positive parameter $\gamma \in [0, 1]$, this defines the γ**-rule** criterion:

$$\frac{\mathrm{P}(\hat{Y} = 1|G = A)}{\mathrm{P}(\hat{Y} = 1|G = B)} \geq \gamma \text{ and } \frac{\mathrm{P}(\hat{Y} = 1|G = B)}{\mathrm{P}(\hat{Y} = 1|G = A)} \geq \gamma.$$

For $\gamma = 0.8$, the above criterion corresponds to a classical *four-fifths rule* used for hiring/college admission (it is part of the "Uniform Guidelines on Employee Selection Procedures" (1978)). For $\gamma = 1$, the above notion coincides with the notion of demographic parity.

In the above definitions of demographic parity and γ-rule, the actual label y is not taken into account. Hence, the above definitions require an (almost) equal probability of acceptance independently on the "goodness" of an individual. More recent, the condition called **equal opportunity** (Hardt et al., 2016b) ensures that the *true positive* outcome must occur with equal probability for both groups, e.g., a hiring probability of a good male engineer should be the same as for a good female engineer:

$$\mathrm{P}(\hat{Y} = 1|Y = 1, G = A) = \mathrm{P}(\hat{Y} = 1|Y = 1, G = B).$$

Note that the perfect classifier $\hat{Y} = Y$ satisfies the notion of equal opportunity whereas it might not satisfy the notions of demographic parity and γ-rule.

The above notions of fairness are called *group fairness* since they define fairness for a group as a whole. Another approach to defining fairness in algorithmic fairness literature is to require a fair decision per each individual—a notion called *individual fairness*.

It can be done by guaranteeing that similar individuals must have close probabilities of a positive outcome (Dwork et al., 2012), or by not favoring less qualified individuals over good qualified (Kearns et al., 2017, meritocratic fairness). Recently, a new direction of *counterfactual fairness* has been proposed, and it ensures that predictions for an individual should not change in a world where an individual's sensitive attribute (e.g., their race) is different while other attributes remain the same (Kilbertus et al., 2017; Kusner et al., 2017).

There are multiple approaches to enforcing fairness. We can distinguish three of them which are mainly *pre*-processing, *in*-processing and *post*-processing. In the pre-processing approach, the goal is to transform the data so that using a classifier on such data would guarantee some notions of fairness. For example, we could learn a new representation of data that would guarantee statistical independence with demographic groups (Gordaliza et al., 2019; Locatello et al., 2019; Zemel et al., 2013). In in-processing, fairness is enforced during the training phase of a classifier. This is usually implemented by solving a constrained optimization problem of loss minimization with fairness constraints (Agarwal et al., 2018; Romano et al., 2020; Zafar et al., 2017a; Zhang et al., 2018). Finally, post-processing techniques adapt existing (possibly unfair) classifiers by transforming their predictions in such a way that they satisfy some notions of fairness. For example, Hardt et al. (2016b) propose a post-processing technique that guarantee group fairness notions; Petersen et al. (2021) propose an algorithm for individual fairness.

Reinforcement learning and bandit problems. The simplest example of a reinforcement learning problem is a *stochastic multi-arm bandit problem* (Slivkins, 2019): at every moment t, the decision-maker is offered a finite set of arms (actions) to choose. A chosen arm a gives some random reward r_a which is a realization of a random variable from some fixed but unknown distribution $\mathcal{D}_a$ associated with this arm a. The goal is to obtain as large a cumulative reward as possible in a finite time. This type of sequential decision-making has found many applications in, for example, online advertisement, where a_t could represent an ad displayed to users at every moment t.

Patil et al. (2020) propose a solution to the fair stochastic bandit problem where each arm should be pulled some required fraction of times that can be seen as a variant of demographic parity. Joseph et al. (2016) and Gillen et al. (2018) study a problem of fairness in *contextual bandits* (Slivkins, 2019, p. 93)—a modification of the stochastic multi-arm bandit problem where at each time t, a context vector x_t is observed prior to choosing an arm $a_t \in A$. For example, the context x_t can be seen as a compact representation of an individual, and a_t as the ad displayed for such individual. The authors call a policy fair if arms with higher long-term rewards are selected with a higher probability than arms with lower long-term rewards which can be seen as a variation of individual fairness. Zhang and Liu (2021) provide an extensive survey on the current progress on fairness in sequential and reinforcement learning.

Learning with strategic individuals. In the learning literature, particularly in the classification literature, it is usually assumed that the feature distribution of individuals is fixed. Recently, it has been shown that individuals may respond to a classification rule by manipulating their features (at some cost) to get a better outcome if the classification rule becomes common knowledge. The classification problem with strategic individuals was first studied by Hardt et al. (2016a). This work was followed by multiple works exploring different aspects of strategic classification problems (Braverman and Garg, 2020; Dong et al., 2018; Kleinberg and Raghavan, 2019; Tsirtsis and Gomez Rodriguez, 2020).

To our knowledge, the first model considering fairness in classification problems with strategic individuals has been studied by Hu et al. (2019). In this work, it is assumed that individuals are separated into two different groups—advantaged and disadvantaged—and the cost of feature manipulation is group-dependent, making it harder for disadvantaged candidates to change their features. The authors show that, at equilibrium, the decision-maker mistakenly accepts some advantaged group members while erroneously rejecting the disadvantaged group members. In addition, the authors study the subsidizing mechanism that allows disadvantaged candidates to manipulate the features at a lower cost; they show that this mechanism increases the utility of the decision-maker, but it can decrease the utility for both groups of candidates. Estornell et al. (2021) study the effect of different group fairness mechanisms in strategic classification such as equating positive rates, true positive rates, or false positive rates.

Positioning of our work. *In this thesis, we assume that the number of individuals is large, and there is enough statistical information to perform the decision-making; hence, the learning aspect of the selection problem is not considered. This thesis focuses on group notions of fairness, particularly on the notions of demographic parity and equal opportunity. We want to emphasize that in this thesis, we study the selection problem for which the size of the selection budget governs the number of positive predictions. This budget constraint makes selection problems different from classification problems typically studied in the algorithmic fairness literature. For example, in (Emelianov et al., 2022b), we show that in selection problems with strategic individuals, due to limited selection size, there is competition among individuals, which is not the case for strategic classification problems.*

2.2 Fairness in Selection Problems

In selection problems, the decision-maker is given some candidates, where each candidate is specified with an estimate of a quality it possesses. The decision-maker has to select a subset of the best candidates given a fixed budget. This budget constraint makes the selection problem different from the classification problem.

The classical explanation for discrimination in selection is *implicit bias* (Greenwald and Krieger, 2006; Kleinberg and Raghavan, 2018). Implicit bias model assumes that the decision-maker observes estimates X_i of quality Y_i per each individual i, and

these estimates are biased for disadvantaged individuals and unbiased for advantaged individuals:

$$X_i = \begin{cases} Y_i/\beta & \text{if individual } i \text{ is disadvantaged,} \\ Y_i & \text{if individual } i \text{ is advantaged,} \end{cases} \quad \text{where } \beta > 1.$$

The problem of fairness in selection under the presence of implicit bias was first studied by Kleinberg and Raghavan (2018). In this work, the authors show that implicit bias naturally leads to the underrepresentation of disadvantaged candidates. After, they consider a fairness mechanism, called *Rooney rule*, which implies that at least one candidate from the disadvantaged group must be selected (Collins, 2007). The authors show that the Rooney rule can improve selection quality under certain conditions. An extension of the Rooney rule is studied by Celis et al. (2020), where the authors consider the ranking problem in the presence of implicit bias; the authors obtain similar results that fairness mechanisms can improve the utility of a ranking. Long-term effects of the Rooney rule in selection with implicit bias are studied by Celis et al. (2021). In this model, the authors explore how the multiplicative implicit bias parameter evolves with time as the decision-maker learns better about the candidates by comparing the latent and the observed selection utilities.

The aforementioned papers (Celis et al., 2021, 2020; Kleinberg and Raghavan, 2018) essentially analyze one-stage selection problems. In practice, many selection problems are performed in two (or more) stages where the first-stage decision is refined in a second stage with access to a more precise information. To our knowledge, the first work studying fairness propagation in a two-stage setting is by Bower et al. (2017). The authors show that an approximate notion of equal opportunity is multiplicative, i.e., if the first stage selection satisfies $(1+\varepsilon)$-equal opportunity and the second stage satisfies $(1+\delta)$-equal opportunity in isolation, then the whole pipeline satisfies $(1+\varepsilon)(1+\delta)$-equal opportunity. Kannan et al. (2019) consider a two-stage selection problem in application to college admission where group-dependent qualities of candidates are observed with group-independent noise. The authors study if certain notions of fairness, irrelevance of group membership and equal opportunity, can be guaranteed simultaneously.

The selection problem is a particular case of aggregation of decisions: in multi-stage selection problems, applicants rejected at earlier stages cannot be presented at later stages. We can imagine several different compositions of dependent or independent decisions. Dwork and Ilvento (2019) study how individually (un)fair decisions compose if aggregated as OR/AND operations. Arunachaleswaran et al. (2021) propose a model of fairness propagation based on a directed acyclic graph, where each starting node represents a certain group, transition from layer to layer depicts the transition between social steps. The authors study how to adjust transitions given a fixed budget to maximize the social welfare and the welfare of the worst-off group.

Finally, one-stage selection problems can be seen as a special case of a ranking problem. In ranking problems, the goal is to sort items (e.g., employees) to guarantee the most relevant ranking for a user (e.g., an employer). Each ranked item gets some exposure, and the higher the item in the list, the higher the exposure. Selection is

thus a ranking where every individual at the top of the list gets the same exposure, while individuals not on the top list get zero exposure. A recent survey on fairness in ranking problems can be found in (Pitoura et al., 2021).

Positioning of our work. *Our work complements those studies on fairness in selection problems. In addition to implicit bias, we consider the effects of group-dependent noise on selection in (Emelianov et al., 2022a, 2020, 2022b). In (Emelianov et al., 2022b), we propose a model of selection with strategic candidates, which is, to our knowledge, is the first such a model in the literature. Finally, in (Emelianov et al., 2019), we consider a multistage selection problem, and study the price of fairness in this model.*

2.3 Long-Term Effects of Fairness Mechanisms

In all the aforementioned literature, the impact of fairness mechanisms is studied only in the short term. In reality, current decisions affect the future decisions that must not be overlooked when designing decision-making policies.

There are several works studying the long-term impact of fairness mechanisms. Liu et al. (2020) study a dynamical model of classification where agents decide to obtain a binary qualification in response to a known classification rule. The qualification induces a feature distribution that the decision-maker observes, and the decision-maker chooses the classification rule which maximizes his utility. The authors identify the equilibria of this dynamical process; they also study how several interventions—subsidizing the cost of efforts and decoupling the decisions for groups—affect the qualification rates at equilibria. A similar model for demographic parity and equal opportunity fairness criteria is proposed by Zhang et al. (2020).

Heidari and Kleinberg (2021) study the optimal algorithm that allocates opportunities in a society consisting of advantaged and disadvantaged groups. The individuals of advantaged groups have higher chances of successfully fulfilling the opportunity compared to disadvantaged candidates; fulfilling the opportunity gives a positive reward to the decision-maker. At each period, a new generation of individuals appears, where each new individual inherits the group of its parent with some probability. The authors identify that the optimal policy—the one that maximizes the cumulative discounted reward—naturally performs affirmative actions. Similarly, Hu and Chen (2018) and Jehiel and Leduc (2021) study whether affirmative actions should be imposed permanently or can be lifted.

Positioning of our work. *While understanding the importance of studying the long-term effects of fairness, we do not explicitly model this setting in this thesis. Nevertheless, our results can be used to provide intuitions on the long-term impact of group fairness mechanisms. In (Emelianov et al., 2022a, 2020), we identify cases when fairness mechanisms that guarantee equal representation also lead to a higher short-term utility. Since equal representation helps obtain better statistical information about individuals, it might also lead to removing statistical discrimination in the future while not sacrificing the selection utility. In (Emelianov et al.,*

2022b), we show that fairness mechanisms incentivize the individuals from disadvantaged groups to make a more significant effort. This affects intergenerational mobility since great efforts at the current generation can potentially lower effort costs in the next generations.

2.4 Economic Models of Discrimination

In algorithmic fairness, the cause of discrimination is rarely modeled. The economic literature proposes and studies several models of discrimination that we discuss in this section.

Taste-based discrimination. The simplest and possibly the first model of discrimination was proposed by Becker (1971). In this model, which is developed to explain wage discrimination, the decision-maker assigns wages to individuals as if the difference in wages for the majority (W_A) and the minority (W_B) groups must be not smaller than a fixed positive value d, i.e., $W_A - W_B \geq d$. In other words, the decision-maker acts as minority workers should compensate for their employment (because of a distaste). The model of implicit bias introduced by Kleinberg and Raghavan (2018), hence, can be seen as a variant of this model, since the estimates of qualities X, given a fixed latent quality Y, differ by a multiplicative parameter β for majority and minority individuals.

Statistical discrimination. The theory of statistical discrimination, initiated by Phelps (1972) and Arrow (1973), argues that the lack of information about individuals is a source of discrimination. Phelps (1972) proposes a model of discrimination where each individual i possesses a latent quality Y_i drawn from a fixed group-independent distribution, assumed normal $\mathcal{N}(0,1)$. A Bayesian decision-maker observes a noisy estimate X_i of the candidate's quality Y_i, where the noise is symmetric and zero-mean with a group-dependent variance $\sigma^2_{G_i}$:

$$X_i = Y_i + \varepsilon_i \cdot \sigma_{G_i}, \text{ where } \varepsilon_i \sim \mathcal{N}(0,1). \tag{2.1}$$

The decision-maker assigns a wage W_i equal to the expected posterior quality

$$W_i = \mathrm{E}(Y_i|X_i) = \rho^2_{G_i} X_i,$$

where ρ_{G_i} is a group-dependent correlation coefficient between X_i and Y_i. This model is assumed to explain the inequality of wages as an outcome of group-depended variance of the noise. From the equation above, we observe that for a given estimate X, the higher noise variance leads to a lower correlation coefficient value, hence, a lower wage. In other words, two employers with equal estimates but different noise variances will be assigned different wages.

Lundberg and Startz (1983) extend Phelps' model to a strategic setting by assuming two groups of workers $G \in \{A, B\}$, where each group of workers chooses the value of effort m_G according to a quadratic costs Cm^2_G. The effort induces a quality Y_G

that is assigned randomly according to a normal distribution $\mathcal{N}(m_G, 1)$. The decision-maker observes noisy estimate X_G of a quality Y_G where the noise variance is group-dependent (as in (2.1)); similarly, the decision-maker assigns the wage equal to the expected quality of a candidate. The authors show that, at equilibrium, the high-noise candidates make a lower effort and are paid less on average compared to the low-noise candidates. However, if the decision-maker is restricted to not using the group information for wage assignment, the effort is equal for both groups.

In many statistical discrimination models, the authors assume that individuals of different groups have identical *a priori* characteristics, but the decision-maker uses their group-based beliefs when facing imperfect information to assess the performance of individuals of a particular group (Aigner and Cain, 1977; Arrow, 1973; Coate and Loury, 1993; Fang and Moro, 2011). In some cases, these beliefs (stereotypes) lead to equilibria in which these discriminating beliefs are fulfilled.

Positioning of our work. *In our (Emelianov et al., 2022a), we do model the effect of taste-based discrimination on selection by introducing the group-dependent implicit bias. In addition to implicit bias, we also consider the impact of statistical discrimination by introducing the group-dependent noise as in (Phelps, 1972). Note that in (Emelianov et al., 2022a, 2020, 2022b), we use a similar model to (Lundberg and Startz, 1983; Phelps, 1972) but in the context of selection problems where a fraction of candidates receives a fixed reward rather than assigning wages to individuals.*

Discrimination and norms. Discrimination can also occur due to self-selection which is usually governed by social norms. For example, women (men) tend to participate in women- (men-) associated activities. Akerlof and Kranton (2000) were the first to motivate and model this phenomenon where an identity of an individual governs his decisions. There are multiple follow-up works, including, e.g., (Benabou and Tirole, 2011; Carvalho and Pradelski, 2019).

Positioning of our work. *We note that in our game-theoretic model of selection in (Emelianov et al., 2022a,b), we do not consider that individuals take into account their identity when making strategic decisions. The underrepresentation in our models is due to group-dependent cost-of-effort and statistical discrimination.*

2.5 Fairness in Computational Social Choice

Finally, in *computational social choice* literature, the authors study how to aggregate individual preferences in order to maximize the social welfare and some notion of fairness. In this section, we present some of the most classical problems in social choice literature: *fair division* and *voting* problems. Recently computer scientists started to look at data-driven problems from a social choice perspective and use the results from this discipline. For example, Balcan et al. (2019) and Hossain et al. (2020) propose and study notions of fairness for classification problems based on the notion of envy-

freeness developed in fair division literature. Chakraborty et al. (2019) propose a fair voting mechanism applied for user-content recommendation systems.

Fair Division. In fair division literature (Suksompong, 2021), the agents must share a good in a fair way. The fair division algorithms find their applications in many situations, such as dividing lands or allocating computational resources to users (Suksompong, 2021). A classical definition of fairness in division literature is *envy-freeness*. If a good can be represented as a set A, then its division among n agents is a tuple $(A_1, \ldots, A_n)$, where $A_i \subseteq A$ and $A_i \cap A_j = \emptyset$. Each agent i possesses a utility u_i. The allocation is called envy-free if for all agents i and j, we have that the agent i does not envy a share of the agent j, i.e., $u_i(A_i) \geq u_i(A_j)$.

The above formulation is the simplest one, and there are multiple variants of the above problem. For example, the good can be *divisible* (e.g., being a segment $[0, 1]$), or *indivisible* (e.g., being a finite set). In addition, we might want to impose additional constraints on the allocation, e.g., a connectedness. We refer the reader to a recent survey by Suksompong (2021) on different variations of the fair division problem.

Voting. In voting literature (Brandt et al., 2016), each agent (voter) is given a list of alternatives (or candidates) to rank. Each agent i has its ordered list of preferences, denoted as a binary relation $\succcurlyeq_i$. The aim of the decision-maker is to aggregate agents' decisions to obtain a collective decision $\succcurlyeq$ that satisfies some efficiency and fairness properties. There is a variety of fairness notions defined in the voting literature. For example, an aggregation rule is called *weakly Paretian* if $a \succcurlyeq_i b$ for all agents implies that $a \succcurlyeq b$. An aggregation is *independent of irrelevant alternatives* if a relation of every two alternatives $a \succcurlyeq b$ in the resulting aggregated profile will not change if adding a new alternative c. As in the algorithmic fairness literature, there is no a unique notion of fairness for voting, and it might be hard to satisfy some of them simultaneously. There is, for example, a famous impossibility theorem by Arrow (Brandt et al., 2016, p. 6) which states that when there are three or more alternatives, then every aggregation rule that is weakly Paretian and independent of irrelevant alternatives must be a dictatorship: the only possible aggregation rule is such that there is a dictating agent d, for which if $a \succcurlyeq_d b$, then $a \succcurlyeq b$.

Chapter 3

Selection with Differential Variance

This chapter is based on our publication (Emelianov et al., 2020) and its extended version (Emelianov et al., 2022a). To have a consistent notation across all chapters, we slightly modified the notations by changing W to Y, $\hat{W}$ to X, $\tilde{W}$ to $\tilde{Y}$, and x to r.

The code to generate all figures can be found at:
https://gitlab.inria.fr/vemelian/differential-variance-code

Abstract Discrimination in selection problems such as hiring or college admission is often explained by implicit bias from the decision maker against disadvantaged demographic groups. In this chapter, we consider a model where the decision maker receives a *noisy* estimate of each candidate's quality, whose variance depends on the candidate's group—we argue that such *differential variance* is a key feature of many selection problems. We analyze two notable settings: in the first, the noise variances are unknown to the decision maker who simply picks the candidates with the highest estimated quality independently of their group; in the second, the variances are known and the decision maker picks candidates having the highest expected quality given the noisy estimate. We show that both baseline decision makers yield discrimination, although in opposite directions: the first leads to underrepresentation of the low-variance group while the second leads to underrepresentation of the high-variance group. We study the effect on the selection utility of imposing a fairness mechanism that we term the γ-rule (it is an extension of the classical four-fifths rule and it also includes demographic parity). In the first setting (with unknown variances), we prove that under mild conditions, imposing the γ-rule increases the selection utility—here there is no trade-off between fairness and utility. In the second setting (with known variances), imposing the γ-rule decreases the utility but we prove a bound on the utility loss due to the fairness mechanism.

3.1 Introduction

Discrimination in selection and the role of implicit bias Many selection problems such as hiring or college admission are subject to discrimination (Bertrand and Mullainathan, 2004), where the outcomes for certain individuals are negatively correlated with their membership in salient demographic groups defined by attributes like gender, race, ethnicity, sexual orientation or religion. Over the past two decades, implicit bias—that is an unconscious negative perception of the members of certain demographic groups—has been put forward as a key factor in explaining this discrimination (Greenwald and Krieger, 2006; Kleinberg et al., 2017). While human decision makers are naturally susceptible to implicit bias when assessing candidates, algorithmic decision makers are also vulnerable to implicit biases when the data used to train them or to make decisions was generated by humans.

To mitigate the effects of discrimination on candidates from underrepresented groups, various fairness mechanisms[1] are adopted in many domains, either by law or through softer guidelines. For instance, the *Rooney rule* (Collins, 2007) requires that, when hiring for a given position, at least one candidate from the underrepresented group be interviewed. The Rooney rule was initially introduced for hiring American football coaches, but it is increasingly being adopted by many other businesses in particular for hiring top executives (Cavicchia, 2015; Passariello, 2016). Another widely used fairness mechanism is the so-called $^4/_5$-*rule* (Holzer and Neumark, 2000), which requires that the selection rate for the underrepresented group be at least 80% of that for the overrepresented group (otherwise one says that there is adverse impact). This rule is part of the "Uniform Guidelines on Employee Selection Procedures" (1978).[2] A stricter version of the $^4/_5$-rule is the so-called *demographic parity* constraint, which requires the selection rates for all groups to be equal. An overview of these and other fairness mechanisms can be found in (Holzer and Neumark, 2000).

Fairness mechanisms, however, have been the subject of frequent debates. On the one hand, they are believed to promote the inclusion of deserving candidates from underrepresented groups who would have otherwise been excluded in particular due to implicit bias. On the other hand, they are viewed as requiring consideration of candidates from underrepresented groups at the expense of candidates from overrepresented groups, which may potentially decrease the overall utility of the selection process, i.e., the overall quality of selected candidates.

[1] These mechanisms are sometimes termed "positive discrimination" (e.g., in Germany, France, China, or India) or "affirmative actions" (in the USA), often referring to their justification as corrective measures against discrimination suffered in the past by disadvantaged groups. In our work, we analyze the effect of these mechanisms in a particular setting of selection problems (with differential variance) independently of their motivation, hence we use the more neutral term "fairness mechanisms."

[2] A set of guidelines jointly adopted by the Equal Employment Opportunity Commission, the Civil Service Commission, the Department of Labor, and the Department of Justice in 1978 ("Uniform Guidelines on Employee Selection Procedures" 1978).

Formal analysis of fairness mechanisms in the presence of implicit bias Perhaps surprisingly, the mathematical analysis of the effect of fairness mechanisms on utility in the context of selection problems was initiated only recently by Kleinberg and Raghavan (2018) (see also an extension to ranking problems by Celis et al. (2020)). Kleinberg and Raghavan (2018) assume that each candidate i has a true latent quality Y_i that comes from a group-independent distribution. They model implicit bias by assuming that the decision maker sees an estimate of the quality $X_i = Y_i$ for candidates from the well-represented group and $X_i = Y_i/\beta$ for candidates from the underrepresented group, where $\beta > 1$ measures the amount of implicit bias. The factor β is unknown (as it is implicit bias) and the decision maker selects candidates by ranking them according to X_i. Then Kleinberg and Raghavan (2018) show that, under a well-defined condition (that roughly qualifies scenarios where the bias is large), the Rooney rule improves in expectation the utility of the selection (measured as the sum of true qualities of candidates selected for interview). This result contradicts conventional wisdom that fairness considerations in a selection process are at odds with the utility of the selection process. Rather, it formalizes the intuition that, in the presence of strong implicit bias (which makes it hard to compare candidates across groups), considering the best candidates across a diverse set of groups not only improves fairness but it also has a positive effect on utility.

The phenomenon of differential variance and its role in discrimination In this chapter, we identify and analyze a fundamentally different source of discrimination in selection problems than implicit bias. Even in the absence of implicit bias in a decision maker's estimate of candidates' quality, the estimates may differ between the different groups in their *variance*—that is, the decision maker's ability to precisely estimate a candidate's quality may depend on the candidate's group. There are at least two main reasons for group-dependent variances in practice. The first arises from *candidates*: different groups of candidates may exhibit different variability when their quality is estimated through a given test. For instance, students of different genders have been observed to show different variability on certain test scores (Baye and Monseur, 2016; O'Dea et al., 2018).[3] The second arises from the *decision makers*: decision makers might have different levels of experience (or different amounts of data in case of algorithmic decision making) judging candidates from different groups and consequently, their ability to precisely assess the quality of candidates belonging to different groups might be different. For instance, when hiring top executives, one may have less experience in evaluating the performance of female candidates because there have been fewer women in those positions in the past (in France for instance, there was only one woman CEO amongst the top-40 companies in 2016-2020 (*Isabelle Kocher, seule*

[3]Note that, while this indicates that observed performance is more variable for one group than the other, it is impossible to tell whether this comes from different underlying distributions or from different measurement variances—or (more likely) from both. In fact, the general "variability hypothesis" is subject to a number of controversies. Nevertheless, this indicates potential differences between groups in the variance of the observed signals and our model can flexibly incorporate both different prior distributions and different measurement noises.

femme dirigeante du CAC 40 n.d.)). The quality estimate's variance might also change from one decision maker to another. For example, in college admissions, recruiters might be able to judge candidates from schools in their own country more accurately than those from international schools.

We refer to the above phenomenon as *differential variance* as the variance of the quality estimate is group-dependent. We posit that differential variance is an omnipresent and fundamental feature affecting selection problems (including in algorithmic decision making). Indeed, having different variances for the different groups is mostly inevitable and hardly fixable. In this chapter, we model the differential variance phenomenon by assuming that the decision maker sees of an estimate of the quality of a candidate X_i that is equal to the candidate's true latent quality Y_i (possibly with an additional bias term) plus an additive noise[4] whose variance depends on the group of the candidate.

We distinguish between two notable settings. In the first setting, the noise variance is assumed to be unknown to the decision maker—we then call it *implicit variance*. In this case, a natural baseline decision maker is the *group-oblivious* algorithm[5] that simply selects the candidates with the highest estimated quality X_i, irrespective of their group. The group-oblivious selection algorithm can represent not only a decision maker unaware of the implicit variance in their estimates, but also a decision maker determined to not use group information—as it may be the case for instance in college admission based on standardized tests. In the second setting, the noise variance is known to the decision maker. In this case, a natural baseline is the *Bayesian-optimal* algorithm: this decision maker can use the group information as well as the knowledge of the distributions of latent quality and noise to select the candidates that maximize the expected quality given the noisy estimate.

As a first cornerstone, our analysis shows that, in the presence of differential variance, both the group-oblivious and the Bayesian-optimal algorithms lead to discrimination (although in opposite directions, see the overview of our results below). A natural way to address this representation inequality is to adopt fairness mechanisms proposed to address discrimination in selection such as the ones discussed above; but this poses the same question that was investigated by Kleinberg and Raghavan (2018) in the case of implicit bias: *what is the effect of fairness mechanisms on the quality of a selection in the presence of differential variance?*

Our model and overview of our results To answer this question, we propose a simple model with two groups of candidates A and B: for each candidate i, the decision maker receives a noisy (and possibly biased) quality estimate $X_i = Y_i - \beta_{G_i} + \sigma_{G_i}\varepsilon_i$, where G_i is the group to which the candidate belongs and ε_i is a standard normal random variable. The estimator has an additive bias β_{G_i} and a variance $\sigma^2_{G_i}$ that depend on the

[4]This noise may be a property of the decision maker getting a noisy perception of the candidate's quality or a property of the candidate (i.e., the variability in the candidate's performance).

[5]Throughout the chapter, we use the term 'algorithm' for the selection procedure, irrespective of whether it is algorithmic decision making or not.

candidate's group. We assume that the true quality Y_i comes from a distribution—assumed normal in our analytical results—that may be group-dependent. The decision maker then selects a fraction α (called selection budget) of the candidates.

The key feature of our model is the variance $\sigma^2_{G_i}$ that depends on the candidate's group—to model differential variance. In its general version, we also allow a bias and a latent quality distribution that depend on the candidate's group. Using this general model, we first show (Section 3.4.1) that both the group-oblivious and the Bayesian-optimal selection algorithms systematically lead to underrepresentation—i.e., lower selection rate—of one of the groups of candidates. Specifically, we identify a cutoff budget such that the group-oblivious selection algorithm leads to underrepresentation of the low-variance group for any budget α smaller than the cutoff (the most common case) and underrepresentation of the high-variance group for any budget α larger than the cutoff. Conversely (and for a different cutoff), the Bayesian-optimal algorithm leads to underrepresentation of the high-variance group for low budgets and of the low-variance group for high budgets. In fact, we show (Section 3.5.1) that this is true even in the absence of bias and with group-independent latent quality distributions—that is, if the noise variance is the only thing that depends on the candidate's group. In this particular case, the cutoff budget for both algorithms is $\alpha = 1/2$.

Then we investigate how the utility of the group-oblivious and the Bayesian-optimal baselines are affected when imposing a fairness mechanism. Specifically, we study a generalization of the $^4/_5$-rule that we call *γ-rule*, which imposes that the selection rate for a given group is at least γ times that of the other group for some parameter $\gamma \in [0, 1]$. This includes both the $^4/_5$-rule ($\gamma = 0.8$) and demographic parity ($\gamma = 1$) as special cases. In the general model, we identify conditions under which the γ-rule never decreases the utility of the group-oblivious algorithm (Section 3.4.2)—that is, there is no trade-off between fairness and selection quality for this baseline. The utility even strictly increases for γ close enough to one, including for demographic parity. Interestingly, in the special case without bias and with group-independent latent quality distributions—that is, with only implicit differential variance—, this result *always* holds for any parameters (Section 3.5.1). Compared to the Bayesian-optimal baseline, the γ-rule cannot increase the utility (since Bayesian-optimal is already optimal given the available information). We prove, however, a bound on the ratio of the utility of the Bayesian-optimal algorithm with and without the γ-rule imposed, which limits the decrease of utility due to imposing a fairness mechanism in this setting. Our bound is valid in the general model (Section 3.4.3) but takes a particularly simple form in the special case without bias and with group-independent latent quality distributions (Section 3.5.1).

A typical case of differential variance is when the decision maker has more uncertainty about one group, due to lack of statistical confidence (e.g., in hiring). In such a case, the high-variance group naturally corresponds to the minority group. The group-oblivious algorithm would then overrepresent the minority group (for small selection budgets), and the fairness mechanism would lead to selecting fewer of the minority group—which is counter-intuitive. We stress, however, that those are

typically cases where the relevant baseline is the Bayesian-optimal algorithm, which behaves very differently. Through the Bayesian posterior quality computation, this baseline would disregard candidates for which the observed quality estimate is uninformative, that is the high-variance group. As mentioned above, we indeed find that the Bayesian-optimal algorithm underrepresents the high-variance group (i.e., the minority), and that the fairness mechanism increases the proportion of selected high-variance candidates—which is coherent with intuition for that case. The group-oblivious baseline is meaningful in other scenarios, typically when the decision maker is not allowed to use the group information (e.g., in college admission based on standardized tests). In such cases, the high-variance group may not be a minority group (and our model does not require that it is).

At a high-level, our results indicate that, with differential variance, the two decision makers (group-oblivious and Bayesian-optimal) lead to nearly opposite outcomes in terms of discrimination; and that the effect of imposing fairness mechanisms can be very different for both. These results imply that a policy-maker considering fairness mechanisms for a given problem should first evaluate to which decision maker the selection rule corresponds, and then choose whether or not to recommend the γ-rule based on it. Note that this should be fairly easy to distinguish between the two in practice, since one conditions on group identity while the other does not.

Organization of the chapter The rest of the chapter is organized as follows. We present the model in Section 3.3. We give all the results in the most general case in Section 3.4. Due to their generality, those results are sometimes complex. In Section 3.5, we analyze three notable cases for which the results are easier to interpret: the case without bias and with group-independent latent quality distributions (Section 3.5.1), the case with bias but with group-independent latent quality distributions (Section 3.5.2), and the case without bias but with group-dependent latent quality distributions (Section 3.5.3). Through numerical simulations in Section 3.6, we extend our analytical results, in particular to cases where the latent quality distribution does not follow a normal law. We conclude in Section 3.7. We provide all omitted proofs in Section 3.8.

3.2 Related Work

The problem of selection under the presence of implicit bias (Greenwald and Krieger, 2006) is first considered by Kleinberg and Raghavan (2018). In their work, the authors study the Rooney rule (Collins, 2007) as a fairness mechanism and show that under certain conditions, it improves the quality of selection. An extension of the Rooney rule is studied under a similar model by Celis et al. (2020), where the authors investigate the ranking problem (of which the selection problem can be seen as a special case) also in the presence of implicit bias and obtain similar results. In both papers, simple mathematical results expressing conditions under which the Rooney rule improves utility are obtained in the limit regime where the number of candidates is very large;

we use the same limit regime in our work. In contrast to those papers that only consider bias, we introduce in addition the notion of differential variance to capture the difference in precision of the quality estimate for different groups. We also consider an additive bias rather than a multiplicative one as it makes more sense for normally distributed qualities. Although our model incorporates both an additive bias and differential variance (in Section 3.4), we purposely restrict it in Section 3.5 to the simplest possible form of differential variance so as to show its effect on the selection problem independently of bias. In our work, we also consider the $^4/_5$-rule (Holzer and Neumark, 2000) (or rather an extension of it that we call the γ-rule and that includes demographic parity) rather than the Rooney rule. The main difference between the two is that the $^4/_5$-rule imposes a constraint on the *fraction* of selected candidates from the underrepresented group whereas the Rooney rule or its extension in (Celis et al., 2020) imposes a constraint on the *number* of selected candidates from the underrepresented group.

Implicit bias, or simply bias (possibly from an algorithm trained on biased data) in the evaluation of candidates quality is certainly a primary factor of discrimination; but it is also one that may reasonably be fixable through the use of algorithms combined with appropriate debiasing techniques and ground truth data (Raghavan et al., 2020) (e.g., by learning fair representations of data (Locatello et al., 2019; Zemel et al., 2013)). The effects of bias can be also fixed by introducing some fairness constraints on learned prediction models. For example, in (Wick et al., 2019), the binary classification problem in the presence of label bias is studied and it is shown that adding a demographic parity constraint to an empirical risk minimization problem can lead to better generalization. Similarly, Blum and Stangl (2019) study the effects of label bias on binary classification and they show that equal opportunity fairness criterion (that ensures that true positives are equal across the groups) can reduce the bias in prediction for most of the reasonable cases, as well as improve the accuracy of classification. Dutta et al. (2020) quantify a fairness-accuracy trade-off using an information theoretic approach and, in addition, they show that for the majority of traditional fairness criteria (like equal opportunity and demographic parity) there exists an ideal data distribution for which fairness and Bayesian optimality are in accordance.

The notion of differential variance first appeared (with different terminology) in the seminal work of Phelps (1972) to explain racial inequality in wages. There, a Bayesian decision maker observes noisy signals of productivity of each worker. Productivities are assumed to be drawn from a common distribution while precisions of estimation differ across races. Phelps shows that a Bayesian decision maker that assigns wages equal to the expected productivity of a worker leads to inequality of wages: in the region of high values of signals the low-precision workers receives lower wages. Our model is similar that of Phelps, with additional bias and possibly group-dependent prior distributions. We also study cases where the variance is implicit—hence the decision maker cannot use Bayes' rule to estimate expected quality given noisy estimates—, and focus on utility for our main results.

This chapter is an extended version of our paper "On Fair Selection in the Presence

of Implicit Variance" (Emelianov et al., 2020). We extend it by considering the general model with bias and group-dependent latent quality distributions, and by analyzing in parallel the two baselines of the group-oblivious and Bayesian-optimal algorithms (whereas in (Emelianov et al., 2020) we only look at the group-oblivious baseline, that is at implicit variance). On the other hand, we do not include the results on two-stage decisions makers for conciseness. Following (Emelianov et al., 2020), Garg et al. (2021) studied a similar model (using the term differential variance that we also adopt here). The authors propose a model of school admission with students of two groups: advantaged and disadvantaged. Each student has an intrinsic quality which is not observable to schools: only noisy signals of the quality are available. The advantaged and disadvantaged students differ in the level of precision of their signals, and can also differ in their ability to access the tests. The authors consider the case of a Bayesian school of limited capacity. They study how different policies made by the school (group-aware and group-unaware) affect the diversity level, individual fairness and overall merit of admitted students. The authors also study how dropping test scores and different abilities to access tests affects the above characteristics.

Recently, we have found a work by Temnyalov (2019) in which the author studies a general selection problem with a finite number of candidates and multiple (ranked) positions, each of a small capacity. Each candidate has an unobservable type, a noisy signal of the type, and a social group (e.g., race or gender), where the type distribution is assumed group-independent. The author shows that the surplus maximizing decision-maker will use differential treatment: in particular, when signals are unbiased, the author shows that high-noise candidates will be preferred over low-noise candidates if the surplus function has convex differences in types (in contrast, low-noise candidates will be chosen over high-noise candidates if the surplus function has concave differences in types). In addition, the author discusses an example of a selection problem with candidates of Gaussian types and noises, and shows that high-noise candidates will be preferred over low-noise if the surplus function has convex differences in types. The main similarity of our results with that of (Temnyalov, 2019) is that the differential variance can lead to differential treatment (even in the absence of implicit bias). The main difference of our work is that we consider the effects of quota-based fairness mechanisms on the utility of selection in the presence of differential variance: we show that fairness mechanisms can increase the selection utility if the baseline decision-maker is group-oblivious; we also show that fairness mechanisms decrease the selection utility if the baseline decision-maker is Bayesian, but we derive the bounds on the utility decrease. In our work, we also assume that the latent quality distribution can be group-dependent.

3.3 Model and Selection Algorithms

3.3.1 The model of selection with differential variance

We consider the following scenario. A decision maker is given n candidates, out of which a subset of size $m = \alpha n$ is selected, $\alpha \in (0, 1)$. We assume that the set of

candidates can be partitioned in two groups: group A and group B. There are n_A candidates from group A and $n_B = n - n_A$ candidates from group B. We refer to them as A-candidates and B-candidates.

Each candidate $i \in \{1, \dots, n\}$ is endowed with a true latent quality Y_i. We assume that the qualities Y_i are drawn *i.i.d.* from an underlying probability distribution that can be group-dependent.[6] For our analytical results, we assume that this distribution is a normal distribution of mean μ_{G_i} and variance $\eta^2_{G_i}$, where $G_i \in \{A, B\}$ is the group of candidate i.

The goal of the decision maker is to maximize the expected quality of the selected candidates: $\mathrm{E}\left[\sum_{i \in \text{selection}} Y_i\right]$. When making the selection decision, the decision maker has access to a (possibly biased) noisy estimator of the true quality. We denote the estimator of the quality of candidate i by X_i. We assume that the bias and the variance of the estimator may depend on the group: for a candidate i that belongs to group $G_i \in \{A, B\}$, its estimated quality is

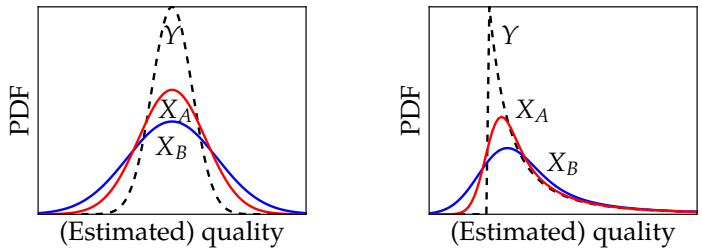
$$X_i = \begin{cases} Y_i - \beta_A + \sigma_A \cdot \varepsilon_i & \text{if } i \text{ is an } A\text{-candidate,} \\ Y_i - \beta_B + \sigma_B \cdot \varepsilon_i & \text{if } i \text{ is a } B\text{-candidate,} \end{cases} \tag{3.1}$$

where ε_i is a centered random variable from $\mathcal{N}(0, 1)$—the standard normal distribution, of mean 0 and variance 1. The variables ε_i are assumed independent and identically distributed. Note that we model the bias as an additive parameter in contrast to the multiplicative parameter in (Kleinberg and Raghavan, 2018). This is more suitable for our model of qualities as normally distributed random variables, which can be negative (a multiplicative bias on a negative quality would turn into a positive effect, which is not meaningful).

We denote by $\hat{\sigma}^2_{G_i} = \sigma^2_{G_i} + \eta^2_{G_i}$ the variance of the estimate X_i. Without loss of generality, we assume that the estimates' variance is larger for A-candidates than for B-candidates, that is $\hat{\sigma}^2_A > \hat{\sigma}^2_B$. We note that none of our results require that A is also the minority group, i.e., that $n_A < n_B$. It is possible to think of scenarios where the minority group has lower variance in cases where the difference in variances arises from the candidates. In the example of students' tests scores (see Section 3.1), for instance, one could potentially observe that males have greater variability in topics in which they are in majority. If the difference in variances arises from the decision maker and has a statistical nature, the minority group (for past selections) will have higher variance due to less data points to build the estimator.

Throughout the chapter, we refer to this difference in variance as *differential variance* because we assume that the variance of the estimators differs across groups. Fig. 3.1 illustrates the resulting distribution of quality estimates for groups A and B for different distributions of the true latent quality (by abuse of notation, we denote by X_A a variable that has the same distribution as $Y_i + \sigma_A \varepsilon_i$ and similarly for B).

[6] We present here the model in its most general form. We will analyze special cases, in particular when the quality distribution is group-independent, in Section 3.5.



(a) $Y \sim \mathcal{N}$ (b) $Y \sim$ Pareto

Figure 3.1 – **Probability density function of the true latent quality Y and the estimated quality X.** To the purpose of illustration, the underlying distribution is assumed group-independent and the estimation is unbiased.

3.3.2 Selection algorithms

Candidates are selected in a one-stage process: for each candidate i, the decision maker observes the quality estimate X_i as well as its group $G_i \in \{A, B\}$. The decision maker then selects m candidates out of those n. The goal of the decision maker is to maximize the expected quality of the selected m candidates. In this chapter, we distinguish and study the following two baseline selection algorithms. Each baseline is a natural selection algorithm in a situation when the decision maker knows the parameters of the model (μ_{G_i}, $\eta^2_{G_i}$, β_{G_i} and $\sigma^2_{G_i}$) or not.

Group-Oblivious Algorithm One of the most natural selection rules is to sort the candidates according to X_i irrespective of their group and to keep the best m. We call this the *group-oblivious* selection algorithm. Typical examples of the group-oblivious algorithm could be admission processes in colleges where the selection is performed with respect to standardized test results (no group information is taken into account), or selection processes where the decision maker does not know the model's parameters, and in particular where it does not know the variance of the estimator (hence the name *implicit variance* in that case). This selection algorithm might be also seen as a fair treatment because the selection does not use the group label. Yet, because of the differential variance or bias, this might lead to discrimination. We will discuss that in Theorem 3.4.1.

Bayesian-Optimal Algorithm When the variance of the noise is known, an alternative selection algorithm is what we call the *Bayesian-optimal* algorithm. This algorithm knows all the parameters of the problem (the quality distribution, the variances of noise σ^2_G, and the biases β_G) and chooses the candidates with the largest expected quality given the estimate X_i. Since (Y_i, X_i) is a bivariate normal random vector, then using the property of conditional expectation for normal random vectors, the expected quality of candidate given its quality estimate can be expressed as:

$$\widetilde{Y}_i = \mathrm{E}(Y_i|X_i) = \frac{\eta^2_{G_i}}{\sigma^2_{G_i} + \eta^2_{G_i}}(X_i + \beta_{G_i}) + \left(1 - \frac{\eta^2_{G_i}}{\sigma^2_{G_i} + \eta^2_{G_i}}\right)\mu_{G_i}. \tag{3.2}$$

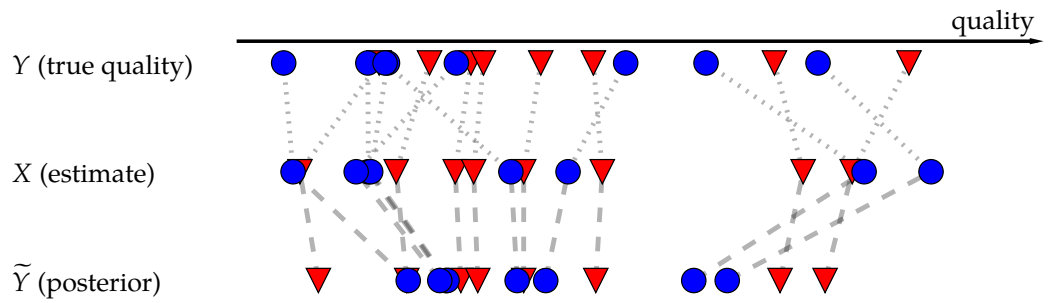


Figure 3.2 – **Illustration of the baseline selection algorithms.** Here, there are $n_{blue} = 8$ blue and $n_{red} = 8$ red candidates, and the decision maker wants to select $m = 2$ candidates. The quality is group-independent and there is no bias. The estimator variance is three times higher for the blue candidates. Here, the group-oblivious algorithm would select the 2 blue candidates. Yet, because blue candidates have higher variance, the Bayesian-optimal algorithm would select 2 red.

Note that $\widetilde{Y}_i$ converges to $X_i + \beta_{G_i}$ as $\sigma^2_{G_i}$ tends to 0 (i.e., there is no noise) and it converges to μ_{G_i} as $\sigma^2_{G_i}$ tends to ∞. Intuitively, all candidates appear similar to the decision maker as the precision of estimation degrades. We denote by $\tilde{\sigma}^2_{G_i} = \eta^4_{G_i} / \left(\eta^2_{G_i} + \sigma^2_{G_i}\right)$ the variance of the expected quality $\widetilde{Y}_i$.

Perhaps more surprisingly, the Bayesian-optimal algorithm also leads to discrimination (although in the opposite way as for the group-oblivious algorithm) as we show in Theorem 3.4.2. We illustrate how the two decision making algorithms work with an example depicted in Figure 3.2. In this example, the blue candidates have higher variance $\sigma_{blue} = 3\sigma_{red}$. This implies that the posteriors $\widetilde{Y}_i$ are more shrank towards the mean for blue candidates than for red candidates: as a result, the Bayesian-optimal tends to select fewer blue candidates compared to red candidates. Note that the Bayesian-optimal is only optimal *in expectation* given the information available; it needs not be optimal for a given realization (on Figure 3.2 the optimal selection ex-post would be one red and one blue).

3.3.3 The γ-rule fairness mechanism

For a given algorithm alg $\in \{\text{obl}, \text{opt}\}$, we denote by r_A^{alg} (and r_B^{alg}) the proportion of the A-candidates (and B-candidates) that are selected, where obl stands for group-oblivious and opt for Bayesian-optimal. A selection algorithm might favor one group or the other, that is $r_A^{\text{alg}} \gg r_B^{\text{alg}}$ or $r_B^{\text{alg}} \gg r_A^{\text{alg}}$. To mitigate the inequality, the decision maker can introduce selection quotas. One example is the 4/5-rule (Holzer and Neumark, 2000) that imposes that $r_A \geq \frac{4}{5} r_B$ and $r_B \geq \frac{4}{5} r_A$.

In this chapter, we consider a generalization of the 4/5-rule that is parameterized by $\gamma \in [0, 1]$. We say that a selection satisfies the *γ-rule* if

$$r_A \geq \gamma r_B \qquad \text{and} \qquad r_B \geq \gamma r_A. \tag{3.3}$$

A selection algorithm satisfies this constraint if and only if it picks at least $m\gamma n_A/(n_B + \gamma n_A)$ A-candidates and at least $m\gamma n_B/(n_A + \gamma n_B)$ B-candidates. Indeed, the total number of selected candidates is $m = x_A n_A + x_B n_B$ which means that $r_B = (m - r_A n_A)/n_B$. The constraint $r_A \geq \gamma r_B$ is therefore true if $r_A \geq \gamma x_B = \gamma(m - r_A n_A)/n_B$ which is true if and only if $r_A \geq \gamma m/(n_B + \gamma n_A)$. Similarly, the constraint $r_B \geq \gamma r_A$ is true if and only if $r_B \geq \gamma(m - r_B n_B)/n_A$.

This means that one can easily transform a baseline into a γ-fair algorithm by first selecting at least $m\gamma n_A/(n_B + \gamma n_A)$ A-candidates and at least $m\gamma n_B/(n_A + \gamma n_B)$ B-candidates and then filling the remaining positions according the best estimated candidates (candidates with largest X_i if the baseline algorithm is group oblivious and with largest $\widetilde{Y}_i$ if the baseline algorithm is Bayesian-optimal), irrespective of their group. This is what defines the *γ-fair group-oblivious* and *γ-fair Bayesian-optimal* algorithms.

When $\gamma = 0$, the γ-fair version of a baseline algorithm reduces to the original unconstrained algorithm (the algorithm that does not to take into account fairness). When $\gamma = 1$, the γ-rule mechanism corresponds to the classical notion of *demographic parity* (Zafar et al., 2017b) that mandates that the selection rates be equal across different groups. We highlight the demographic parity mechanism as a special and important case of the γ-rule. Note that because n_A, n_B and m are integer variables, it might be impossible to satisfy the constraints in (3.3) when γ is too close to 1. In such a case, we say that an algorithm is γ-fair if the constraint (3.3) is satisfied up to one candidate.

3.3.4 Simplification of the selection problem for large n and m

In the remainder of the chapter, we study the selection problem when the number of candidates is large. That is, we assume that there exist fixed fractions $p_A, \alpha \in (0, 1)$ such that

$$n_A = \lfloor p_A n \rfloor \qquad n_B = \lceil (1 - p_A) n \rceil \qquad m = \lfloor \alpha n \rfloor,$$

and let n grow. Our theoretical results are obtained in the limit where n goes to infinity (similarly to (Celis et al., 2020; Kleinberg and Raghavan, 2018)). In Section 3.6.3 we show numerically that our results for $n = \infty$ continue to hold for finite selection sizes. Note that p_A represents the fraction of A-candidates in the population while α represents the global selection ratio (or budget).

For a finite n, the selection algorithms presented in Sections 3.3.2-3.3.3 are hard to analyze because the probability for a candidate to be selected depends on all other candidates. As we prove below, characterizing the performance of a selection problem is simpler when the number of candidates n is infinite because there is an equivalence between the algorithms presented in the previous sections and threshold-based algorithm. A threshold-based algorithm uses two thresholds θ_A and θ_B and selects all G_i-candidates, such that $X_i \geq \theta_{G_i}$.[7] For given thresholds θ_A and θ_B, we denote the

[7]Note that the Bayesian-optimal algorithm can also be written that way, with appropriate thresholds, because *within a given group* the expected qualities $\widetilde{Y}_i$ are in the same order as the signals X_i.

expected utility of the corresponding selection by $\mathcal{V}(\theta_A, \theta_B)$:

$$\mathcal{V}(\theta_A, \theta_B) = \mathrm{E}\left[Y_i \mid X_i \geq \theta_{G_i}\right].$$

Hence, the selection of a candidate does not depend on the qualities of the other individuals. Also, as we show in the next theorem, the fraction of A-candidates that are selected becomes deterministic as n goes to infinity.

Lemma 3.3.1. *For any of the selection algorithms presented in Sections 3.3.2-3.3.3,*

1. *there exists a deterministic fraction $r_A \in [0, 1]$ such that the fraction of A-candidates that are selected by the algorithm converges (in probability) to r_A as n grows;*

2. *there exist deterministic thresholds θ_A, θ_B such that the expected utility of this algorithm converges to $\mathcal{V}(\theta_A, \theta_B)$.*

Proof Sketch. The above result is essentially a direct consequence of the law of large numbers. By the Glivenko-Cantelli theorem, the empirical distribution of the estimated qualities of the G-candidates converges to the distribution of X_G as $n \to \infty$. This shows that taking the best $\lfloor np_A r_A \rfloor$ A-candidates or taking all A-candidates above the r_A-quantile of the distribution X_A is asymptotically equivalent as $n \to \infty$. □

For these given thresholds θ_A, θ_B, the fractions of selected candidates are $\mathrm{P}(X_i \geq \theta_{G_i})$. Using the above definition, we denote by $\mathcal{U}(r_A)$ the expected utility of a threshold-type selection algorithm that selects A-candidates with probability r_A and that satisfies the selection size constraints in expectation:

$$\mathcal{U}(r_A) = \mathcal{V}(\theta_A, \theta_B), \text{ where } \theta_A, \theta_B \text{ are such that } \begin{cases} \mathrm{P}(X_i \geq \theta_A \mid G_i = A) = r_A, \\ \mathrm{P}(X_i \geq \theta_{G_i}) = \alpha. \end{cases} \tag{3.4}$$

Note that combining the constraints in (3.4) immediately gives that such an algorithm selects B-candidates with probability $r_B = (\alpha - r_A p_A)/(1 - p_A)$. Hence, it is sufficient to describe the algorithm with r_A.

The above definition of expected quality is not directly applicable to the selection algorithms presented in Section 3.3.2 because those algorithms are defined neither in terms of fraction of selected candidates nor in terms of thresholds. In fact, for a given selection algorithm, the fractions of selected A- and B-candidates depend on the realizations of the random variables representing the quality (Y_i) and the estimated quality (X_i). As a result, these fractions (r_A and r_B) are random variables. For instance, if because of randomness the A-candidates are evaluated much worse than the B-candidates, then r_A will be 0 for the group-oblivious algorithm. Lemma 3.3.1 shows that when the population is large, these random fluctuations disappear. It shows that, when n is large, the performance of the various algorithms are simply characterized by r_A.

For a finite n, characterizing precisely the utility of an algorithm like group-oblivious is computationally difficult due to the correlations between the selection

of the different agents. Lemma 3.3.1 allows us to greatly simplify the study of the performance of the various algorithms because the function $\mathcal{U}$, defined in (3.4), depends only on one parameter r_A, and is simpler to characterize than the expectation over a finite number of candidates n.

3.3.5 Summary of main notation

We denote respectively by r_A^{obl}, $r_A^{\gamma\text{-obl}}$, r_A^{opt} and $r_A^{\gamma\text{-opt}}$ the asymptotic fraction of A-candidates that are selected for the group-oblivious, the γ-fair group-oblivious, the Bayesian-optimal and the γ-fair Bayesian-optimal algorithms. We also identify an important subcase of the γ-rule for $\gamma = 1$. In this case both the γ-fair group-oblivious algorithm and the γ-fair Bayesian-optimal algorithm select A-candidates at rate $r_A = \alpha$, so there is no difference between them. We name the corresponding selection algorithm as *demographic parity algorithm*.

We denote the expected performance of the introduced algorithms by

$$\mathcal{U}^{\text{obl}} = \mathcal{U}(r_A^{\text{obl}});\ \ \mathcal{U}^{\gamma\text{-obl}} = \mathcal{U}(r_A^{\gamma\text{-obl}});\ \ \mathcal{U}^{\text{opt}} = \mathcal{U}(r_A^{\text{opt}});\ \ \mathcal{U}^{\gamma\text{-opt}} = \mathcal{U}(r_A^{\gamma\text{-opt}});\ \ \mathcal{U}^{\text{dp}} = \mathcal{U}(r_A^{\text{dp}}).$$

We summarize the other notation in Table 3.1.

Table 3.1 – Summary of notation.

Y_i	latent quality of candidate i
X_i	estimated quality of candidate i
$\widetilde{Y}_i$	expected value of latent quality of candidate i given the estimate X_i
μ_G	expected value of latent quality Y_G
η_G^2	variance of latent quality Y_G
σ_G^2	variance of additive noise
$\hat{\sigma}_G^2$	variance of estimated quality X_G. It equals $\sigma_G^2 + \eta_G^2$.
$\tilde{\sigma}_G^2$	variance of expected quality $\widetilde{Y}_G$. It equals $\eta_G^4 / \left(\eta_G^2 + \sigma_G^2\right)$
r_G^{alg}	fraction of G-candidates that are selected by a given algorithm "alg"
θ_G^{alg}	threshold above which G-candidates are selected by the algorithm "alg"
ϕ, Φ, Φ^{-1}	PDF, CDF and quantile of the standard normal distribution $\mathcal{N}(0,1)$

3.4 Analysis of the General Model

In this section, we present the main technical results in the most general model. The results that we prove in this section are quite abstract; to make things more concrete and provide more intuitive results, we will instantiate this general model in important sub-cases in Section 3.5.

We start by showing why the two baseline algorithms lead to discrimination, in Theorem 3.4.1 for the group-oblivious and in Theorem 3.4.2 for the Bayesian-optimal

algorithm. Then, we specify in Theorem 3.4.3 conditions under which the γ-rule fairness mechanism increases the utility of selection compared to the unconstrained group-oblivious algorithm. Although it is clear that the γ-rule mechanism cannot increase the utility of the Bayesian-optimal algorithm (since it is an expected utility and the Bayesian-optimal algorithm maximizes it by definition), we prove in Theorem 3.4.4 that the ratio of the utilities of the unconstrained Bayesian-optimal and the γ-fair Bayesian-optimal algorithms is bounded.

3.4.1 Discrimination of baseline selection algorithms

Recall that we assume (without loss of generality) that group A is the high-variance group, that is $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$. Then, the distribution of X_A has longer tails compared to the distribution of X_B. Thus, if the selection size is small, A-candidates will be selected by the group-oblivious algorithm at higher rate compared to B-candidates because the probability to estimate an A-candidate as "outstanding" is higher than for B-candidates. In contrast, if the selection size is large, the chance of estimating an A-candidate as poor is larger than for B-candidates, in which case the group-oblivious algorithm selects a lower fraction of A-candidates. This can be formally stated as follows.

Theorem 3.4.1. *Assume without loss of generality that $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$. When using the group-oblivious selection algorithm, the selection rates for A- and B-candidates, r_A^{obl} and r_B^{obl}, satisfy:*

$$r_A^{\text{obl}} > r_B^{\text{obl}} \text{ if and only if } \alpha < \Phi\left(\frac{\Delta\mu - \Delta\beta}{\Delta\hat{\sigma}}\right),$$

where $\Delta\mu = \mu_A - \mu_B$, $\Delta\beta = \beta_A - \beta_B$ and $\Delta\hat{\sigma} = \hat{\sigma}_A - \hat{\sigma}_B$.

Proof Sketch. The group-oblivious algorithm sorts candidates by their estimated qualities X_i and takes the best αn by applying a group-independent threshold θ. The expression for the selection rates $r_G^{\text{obl}} = 1 - \Phi\left(\frac{\theta - \mu_G + \beta_G}{\hat{\sigma}_G}\right)$ and a simple rearrangement allows us to find such sizes of budget α for which selection rates for both groups become equal $r_A^{\text{obl}} = r_B^{\text{obl}}$. The result then follows from the corresponding properties of normal CDF and our assumption that $\hat{\sigma}_A > \hat{\sigma}_B$. A detailed proof is given in Section 3.8.1. □

The above result implies that, for a small selection budget, the group-oblivious algorithm will select high-variance candidates at a higher rate. Note that this result does not assume that this higher variance comes from the variance of the true quality (η_A^2 and η_B^2) or from the variance of the estimates (σ_A^2 and σ_B^2). It is only assumed that the variance of X_A, equal to $\hat{\sigma}_A^2 = \sigma_A^2 + \eta_A^2$, is larger than the one of X_B.

As we show below, nearly the opposite is true for the Bayesian-optimal algorithm: for a small selection budget, in the case of group-independent variance of the latent quality ($\eta_A^2 = \eta_B^2$), a Bayesian-optimal algorithm will select fewer candidates from the high-variance group. In the case where $\eta_A^2 \neq \eta_B^2$, though, which group is underrepresented will be determined by the variances $\tilde{\sigma}$ and not $\hat{\sigma}$, see our discussion below the

theorem. Note also that the specific budget threshold at which the transition happens is not the same as for the group-oblivious algorithm.

Theorem 3.4.2. *Assume that $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$. When using the Bayesian-optimal selection algorithm, the selection rates for A- and B-candidates, r_A^{opt} and r_B^{opt}, satisfy:*

$$r_A^{\text{opt}} < r_B^{\text{opt}} \text{ if and only if } \alpha < \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right),$$

where $\Delta\mu = \mu_A - \mu_B$, $\Delta\tilde{\sigma} = \tilde{\sigma}_A - \tilde{\sigma}_B$.

Proof Sketch. In the Bayesian-optimal algorithm, the candidates are sorted by their expected qualities $\widetilde{Y}$ and a group-independent threshold is applied to select the best αn candidates. The expression (3.2) for the expected quality $\widetilde{Y}_G$ allows us to compare the selection rates r_A^{opt} and r_B^{opt} for different groups A or B and to find such value of budget α for which $r_A^{\text{opt}} = r_B^{\text{opt}}$. Then using the fact that $\widetilde{Y}_G$ follows normal law and the relation between $\tilde{\sigma}_A$ and $\tilde{\sigma}_B$, we obtain our result. A complete proof can be found in Section 3.8.2. □

The result of Theorem 3.4.2 is consistent with the observation from (Phelps, 1972) in a simpler setting (without bias and with group-independent distribution of the latent quality Y): in the presence of differential variance, the candidates from the high-variance group will appear more similar to each other to the decision maker, hence the distribution of computed expected quality will have a longer tail for the low-variance group. As a consequence, for small enough selection budgets, candidates from the high-variance group will be selected at a lower rate.

Note that the result in Theorem 3.4.2 imposes a condition on the order between $\tilde{\sigma}_A^2 = \eta_A^4/\left(\eta_A^2 + \sigma_A^2\right)$, the variance of $\widetilde{Y}_A$, and $\tilde{\sigma}_B^2$; but it is not conditional on the relation between $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$, i.e., both $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$ and $\hat{\sigma}_A^2 \leq \hat{\sigma}_B^2$ are allowed. Hence the condition $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$ comes without loss of generality for this result. Note also that in the case where the variances of the true quality are the same for both groups ($\eta_A = \eta_B$), the two conditions from Theorems 3.4.1 and 3.4.2 are equivalent, that is $\tilde{\sigma}_A < \tilde{\sigma}_B$ if and only if $\hat{\sigma}_A > \hat{\sigma}_B$ (since it holds if and only if $\sigma_A > \sigma_B$). The main special cases that we consider in Section 3.5 (specifically those of Sections 3.5.1 and 3.5.2) are in this case (i.e., satisfy $\eta_A = \eta_B$).

3.4.2 The γ-rule mechanism can increase the utility of the group-oblivious algorithm

As we show in Theorem 3.4.1, the group-oblivious algorithm leads to overrepresentation of the high-variance group A, if the budget α is small. To mitigate this effect, the decision maker can use the γ-rule fairness mechanism introduced in Section 3.3.3.

In the next theorem, we provide a condition on budgets α for which using the γ-fair group-oblivious algorithm attains larger quality of selection compared to the unconstrained group-oblivious algorithm. The main message of this theorem is that

if A-candidates have larger variability of their estimate compared to B-candidates (i.e., $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$) and the variance of expected quality for A-candidates is smaller than for B-candidates (i.e., $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$), then the γ-fair group-oblivious algorithm leads to larger quality of selection compared to the group-oblivious algorithm for both small and large budgets α. As said earlier, when $\eta_A = \eta_B$, these conditions are always satisfied, up to switching the groups A and B. At the same time, there may exist a region of budgets α such that the γ-rule fairness mechanism harms the quality of selection compared to the group-oblivious algorithm.

Theorem 3.4.3. *Without loss of generality, assume that the estimates of quality for A-candidates has larger variance than for B-candidates $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$. Assume also that the variance of the expected quality is smaller for A-candidates than for B-candidates ($\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$), and let us define*

$$\alpha^{\min} = \min\left\{\Phi\left(\frac{\Delta\mu - \Delta\beta}{\Delta\hat{\sigma}}\right), \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)\right\}, \ \alpha^{\max} = \max\left\{\Phi\left(\frac{\Delta\mu - \Delta\beta}{\Delta\hat{\sigma}}\right), \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)\right\},$$

where $\Delta\mu = \mu_A - \mu_B$, $\Delta\beta = \beta_A - \beta_B$, $\Delta\hat{\sigma} = \hat{\sigma}_A - \hat{\sigma}_B$ and $\Delta\tilde{\sigma} = \tilde{\sigma}_A - \tilde{\sigma}_B$. We have:

(i) For any $\alpha \in (0, \alpha^{\min}) \cup (\alpha^{\max}, 1)$, the demographic parity algorithm strictly improves the selection quality compare to the group-oblivious algorithm and the γ-fair group-oblivious algorithm for $\gamma < 1$ weakly improves it:

$$\mathcal{U}^{\text{dp}} > \mathcal{U}^{\gamma\text{-obl}} \geq \mathcal{U}^{\text{obl}}.$$

(ii) If $\alpha^{\min} = \alpha^{\max}$, then for $\alpha = \alpha^{\min} = \alpha^{\max}$ one has $\mathcal{U}^{\text{dp}} = \mathcal{U}^{\gamma\text{-obl}} = \mathcal{U}^{\text{obl}}$.

(iii) Assume that $\alpha^{\min} \neq \alpha^{\max}$, then there exists $[\tilde{\alpha}^{\min}, \tilde{\alpha}^{\max}]$, where $\tilde{\alpha}^{\min} > \alpha^{\min}$ and $\tilde{\alpha}^{\max} < \alpha^{\max}$, such that for any $\alpha \in [\tilde{\alpha}^{\min}, \tilde{\alpha}^{\max}]$, the demographic parity algorithm strictly harms the selection quality compared to the group-oblivious algorithm and the γ-fair group-oblivious algorithm for $\gamma < 1$ weakly harms it:

$$\mathcal{U}^{\text{dp}} < \mathcal{U}^{\gamma\text{-obl}} \leq \mathcal{U}^{\text{obl}}.$$

Proof. We prove in Theorem 3.4.1 that if $\alpha < \Phi\left((\Delta\mu - \Delta\beta)/\Delta\hat{\sigma}\right)$, then the group-oblivious algorithm leads to overrepresentation of the high-variance group A. At the same time, the group-oblivious algorithm leads to underrepresentation of the group A if $\alpha > \Phi\left((\Delta\mu - \Delta\beta)/\Delta\hat{\sigma}\right)$. Similarly, we prove in Theorem 3.4.2 that if $\tilde{\sigma}_A < \tilde{\sigma}_B$, then for $\alpha < \Phi\left(\Delta\mu/\Delta\tilde{\sigma}\right)$, the Bayesian-optimal algorithm underrepresents the group A and for $\alpha > \Phi\left(\Delta\mu/\Delta\tilde{\sigma}\right)$ it overrepresents the group A.

Recall that for any value of α, the demographic parity algorithm requires that candidates from both groups, A and B, must be selected at equal rates, i.e., $r_A^{\text{dp}} = r_B^{\text{dp}} = \alpha$. It means that if $\alpha \in (0, \alpha^{\min}) \cup (\alpha^{\max}, 1)$, then the demographic parity algorithm will perform a selection such that either $r_A^{\text{obl}} < r_A^{\text{dp}} < r_A^{\text{opt}}$ or $r_A^{\text{opt}} < r_A^{\text{dp}} < r_A^{\text{obl}}$ (also $r_A^{\text{obl}} \leq r_A^{\gamma\text{-obl}} < r_A^{\text{opt}}$ or $r_A^{\text{opt}} < r_A^{\gamma\text{-obl}} \leq r_A^{\text{obl}}$ for $\gamma < 1$). In Section 3.8.3, we prove that the selection quality $\mathcal{U}$ is a concave function of r_A with a single maximum at $r_A = r_A^{\text{opt}}$. Hence, from this property we conclude that $\mathcal{U}^{\text{dp}} > \mathcal{U}^{\text{obl}}$ and $\mathcal{U}^{\gamma\text{-obl}} \geq \mathcal{U}^{\text{obl}}$ for $\gamma < 1$. Finally, (iii) is due to the fact that the utility $\mathcal{U}$ is a continuous and smooth function of r_A as we prove in Section 3.8.3. □

While the statement of Theorem 3.4.3 is somewhat complex due to its generality, in special cases (e.g., that of Section 3.5.1) we have $\alpha^{\min} = \alpha^{\max}$. This means that in the special case of Section 3.5.1, we are always in case (i) of Theorem 3.4.3: the γ-fair group-oblivious algorithm attains a larger utility than the corresponding baseline (or at worst an equal utility).

Note that the statement of Theorem 3.4.3 is under the assumption that $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$ *and* $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$. As discussed earlier, this assumption may not be without loss of generality if $\eta_A \neq \eta_B$. If it does not hold, then using the demographic parity algorithm could lead to a worse utility than the group-oblivious algorithm. Even in this case, however, the ratio $\mathcal{U}^{\text{obl}}/\mathcal{U}^{\text{dp}}$ remains bounded. Indeed, we can write $\mathcal{U}^{\text{obl}}/\mathcal{U}^{\text{dp}} \leq \mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ and the ratio $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ itself is upper-bounded as we show in the next section (Theorem 3.4.4).

3.4.3 Bounds on the decrease of utility due to imposing γ-rule on the Bayesian-optimal algorithm

By definition, the Bayesian-optimal algorithm maximizes the utility of the selection which means that imposing a γ-rule cannot increase the expected utility of the selection—in most cases it decreases it. In this section, however, we obtain a bound on the ratio of utilities for Bayesian-optimal and γ-fair Bayesian-optimal algorithms. This is stated in the following theorem:

Theorem 3.4.4. *Assume that* $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$ *and that* $\mu_A, \mu_B \geq 0$*, then for any budget* α *the ratio* $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\gamma\text{-opt}}$ *satisfies the following bound:*

$$1 \leq \frac{\mathcal{U}^{\text{opt}}}{\mathcal{U}^{\gamma\text{-opt}}} \leq 1 + \begin{cases} -\frac{\alpha}{p_A + p_B/\gamma} \cdot g(\mu_G, \tilde{\sigma}_G, p_G, \alpha), & \text{if } \alpha \leq \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right) \\ \left(1 - \frac{\alpha}{p_A + p_B\gamma}\right) \cdot g(\mu_G, \tilde{\sigma}_G, p_G, \alpha), & \text{if } \alpha > \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right) \end{cases}$$

where $\Delta\mu = \mu_A - \mu_B$, $\Delta\tilde{\sigma} = \tilde{\sigma}_A - \tilde{\sigma}_B$ *and* $g(\mu_G, \tilde{\sigma}_G, p_G, \alpha) = \frac{p_A}{\alpha} \frac{\Delta\mu + \Phi^{-1}(1-\alpha)\Delta\tilde{\sigma}}{\sum_G p_G \mu_G + \frac{\phi(\Phi^{-1}(1-\alpha))}{\alpha} \sum_G p_G \tilde{\sigma}_G}$.

Proof Sketch. The first inequality is due to the fact that the utility function $\mathcal{U}(r_A)$ is strictly concave and that it attains its maximum at $r_A = r_A^{\text{opt}}$ as we show in Section 3.8.3.

To prove the second inequality, we need a few preparatory steps. First, using the result of Theorem 3.4.2, we obtain that for the budgets $\alpha < \Phi\left(\Delta\mu/\Delta\tilde{\sigma}\right)$, we have $r_A^{\text{opt}} \leq r_A^{\gamma\text{-opt}} < r_A^{\text{dp}}$. Using the concavity of $\mathcal{U}$ and the mean value theorem from real analysis, we obtain: $\frac{\mathcal{U}(r_A^{\text{opt}}) - \mathcal{U}(r_A^{\text{dp}})}{r_A^{\text{opt}} - r_A^{\text{dp}}} \geq \mathcal{U}'(r_A = r_A^{\text{dp}}) \implies \mathcal{U}(r_A^{\text{opt}}) - \mathcal{U}(r_A^{\text{dp}}) \leq -\alpha \cdot \mathcal{U}'(r_A^{\text{dp}})$. After, we divide both parts by $\mathcal{U}(r_A = r_A^{\text{dp}})$. The expressions for $\mathcal{U}'(r_A = r_A^{\text{dp}})$ and $\mathcal{U}(r_A = r_A^{\text{dp}})$ can be written explicitly using the equation derived in Section 3.8.3. A complete proof is given in Section 3.8.4. □

The expression in Theorem 3.4.4 is general but complex due to the large number of model parameters. It can be simplified as we tighten up some of the assumptions (see Section 3.5). There are also interesting behaviors to observe for some values of the parameters. First, if the size of group A becomes small (i.e., $p_A \to 0$), we observe

that the function g converges to 0, hence the upper bound converges to 1. This is expected, since the introduction of the γ-rule mechanism will affect the selection in a tiny amount due to a small number of A-candidates. Second, as the selection budget decreases (i.e., $\alpha \to 0$), we can show using L'Hôpital's rule that the upper bound in this limit converges to $1 - \frac{p_A \Delta\tilde{\sigma}}{\sum p_G \tilde{\sigma}_G}$ for $\gamma = 1$. In other words, the difference in the expected values of qualities $\Delta\mu$ does not play any role. This is also quite natural, since for tiny selection budgets α, the competition is among the candidates with very large values of quality which is due to the variance of the distribution of latent quality but not their mean values.

3.5 Notable Special Cases of the General Model

The results in Section 3.4 might be difficult to interpret without considering some specific cases. In this section, we decompose the effects of different factors by tightening up some of the assumptions of our model while keeping the others in place. We consider the following important special cases: In Section 3.5.1 we assume that there is no bias in the estimation of quality and that the quality distribution is group-independent. This is the model studied in (Emelianov et al., 2020), where the only quantity that depends on the candidate's group is the noise variance (to isolate the differential variance effect). In Section 3.5.2, we assume that the quality distribution is group-independent but the estimates are biased. In Section 3.5.3 we assume unbiased estimates but let the quality be group-dependent. All these subcases allow us to greatly simplify the results of Theorem 3.4.3 and Theorem 3.4.4.

3.5.1 Group-independent latent quality and unbiased estimates

In this section, we assume that the underlying quality distribution is group-independent (this is the classical assumption in the literature, see for instance (Celis et al., 2020; Kleinberg and Raghavan, 2018)) and follows a normal law with mean μ and variance η^2. To isolate the effect of the variance, we also assume that quality estimates X_i are unbiased, i.e., $\beta_{G_i} = 0$. The main result of this section is that imposing a fairness constraint in this context *cannot* decrease the utility compared to using the unconstrained group-oblivious baseline. We also simplify the bound of Theorem 3.4.4 on the decrease of the utility of the Bayesian-optimal algorithm due to the γ-rule fairness mechanism.

First, the following corollary relates selection ratios for two baseline algorithms. It can be obtained directly from Theorems 3.4.1 and 3.4.2. (Recall that in this special case of group-independent quality distribution, we have $\sigma_A^2 > \sigma_B^2$ if and only if $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$, which is also if and only if $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$.)

Corollary 3.5.1 (Corollary of Theorems 3.4.1 and 3.4.2)**.** *Assume that the quality distribution is group-independent $Y_i \sim \mathcal{N}(\mu, \eta^2)$ and that the quality estimates X_i are unbiased $\beta_G = 0$, $\forall G \in \{A, B\}$. Assume without loss of generality that $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$. When using the group-oblivious selection algorithm and the Bayesian-optimal selection algorithm, the fractions r_G^{obl} and r_G^{opt} of selected candidates from each group satisfy:*

(i) $r_A^{\text{obl}} > r_B^{\text{obl}}$ *if and only if* $\alpha < 1/2$*;*

(ii) $r_A^{\text{opt}} < r_B^{\text{opt}}$ *if and only if* $\alpha < 1/2$*.*

Corollary 3.5.1 formalizes in simple terms, for the case of group-independent latent quality distributions and unbiased estimators, the discrimination that results from the two baseline algorithms. Notably, (i) states that for selection budgets below 1/2, the group-oblivious algorithm overrepresents the high-variance group. If the high-variance group is a minority, this is counter-intuitive. As noted in the introduction, however, these typically correspond to cases where the Bayesian-optimal baseline is more meaningful. Then, (ii) states that for small budgets, the Bayesian-optimal algorithm indeed underrepresents the high-variance group.

In Section 3.4 we specify a condition under which the γ-rule fairness mechanism is beneficial to the utility of the group-oblivious algorithm. In the special case of group-independent prior and unbiased estimator, the thresholds $\alpha^{\min}$ and $\alpha^{\max}$ defined in Theorem 3.4.3 coincide and are equal to 1/2. This implies the next theorem, which shows that for this case, the γ-rule fairness mechanism cannot decrease the average quality of a selection compared to the group-oblivious algorithm (without any condition on α).

Corollary 3.5.2 (Corollary of Theorem 3.4.3)**.** *Assume that the quality distribution is group-independent* $Y_i \sim \mathcal{N}(\mu, \eta^2)$ *and that the quality estimates* X_i *are unbiased* $\beta_G = 0$*,* $\forall G \in \{A, B\}$*. Let, without loss generality,* $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$*, then for any budget* $\alpha \neq 1/2$*, the demographic parity selection algorithm provides a larger utility than the* γ*-fair group-oblivious selection algorithm with* $\gamma < 1$*, which in turn provides a larger utility than the group-oblivious selection algorithm:*

$$\mathcal{U}^{\text{dp}} > \mathcal{U}^{\gamma\text{-obl}} \geq \mathcal{U}^{\text{obl}}.$$

The above inequality is an equality when $\alpha = 1/2$*.*

Proof. This result is a special case of Theorem 3.4.3. Since the distribution of quality is group-independent and there is no implicit bias, then the condition in Theorem 3.4.3 holds, and $\alpha^{\min}$ and $\alpha^{\max}$ coincide and become equal to 1/2. □

As for the general case, the γ-rule fairness mechanism cannot increase the selection quality of the Bayesian-optimal baseline. In Theorem 3.4.4, we obtained a bound on the decrease of utility. In the next result, we show how this result simplifies in the modeling assumptions of the current subsection. We provide the bound for $\alpha \leq 1/2$ as it is the most interesting setting. The one for $\alpha > 1/2$ can also be easily deduced.

Corollary 3.5.3 (Corollary of Theorem 3.4.4)**.** *Assume that quality distribution is group-independent* $Y_i \sim \mathcal{N}(\mu, \eta^2)$ *and quality estimates* X_i *are unbiased* $\beta_G = 0$*,* $\forall G \in \{A, B\}$*. Let, without loss of generality,* $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$ *and* $\mu \geq 0$*. Then for all* $\alpha \neq 1/2$*, the demographic parity selection algorithm provides a smaller utility than the* γ*-fair Bayesian-optimal selection algorithm with* $\gamma < 1$*, which in turns provides a smaller utility than the Bayesian-optimal*

selection algorithm. The utility ratio $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\gamma\text{-opt}}$ *for any budget* $\alpha \leq 1/2$ *has the following bound:*

$$1 \leq \frac{\mathcal{U}^{\text{opt}}}{\mathcal{U}^{\gamma\text{-opt}}} \leq 1 + g(\alpha) \cdot \frac{1}{p_A + (1-p_A)/\gamma} \cdot \frac{p_A(\nu - 1)}{p_A + (1-p_A)\nu},$$

where $g(\alpha) = \frac{\alpha\Phi^{-1}(1-\alpha)}{\phi\left(\Phi^{-1}(1-\alpha)\right)}$ *and* $\nu = \hat{\sigma}_A/\hat{\sigma}_B > 1$.

Proof. Direct from Theorem 3.4.4 when we set $\mu = \mu_A = \mu_B$ and $\eta^2 = \eta_A^2 = \eta_B^2$. □

For $\gamma = 1$, which is the case of demographic parity, we can further simplify the expression in Corollary 3.5.3. By using the fact that $g(\alpha)$ is decreasing with α and that $\lim_{\alpha\to 0} g(\alpha) = 1$, we can write

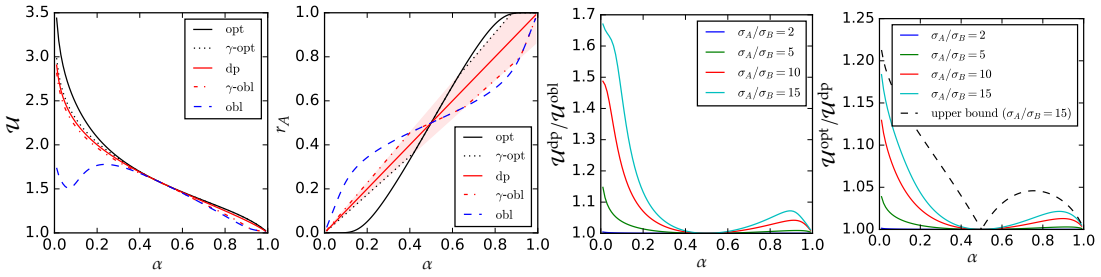
$$1 \leq \mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}} \leq 1 + \frac{p_A(\nu - 1)}{p_A + (1-p_A)\nu}.$$

Note that as ν tends to 1, meaning that there is no difference in variances between A and B group, the upper bound also tends to 1 and matches the lower bound. Most interestingly, we observe that the larger the difference in variances ν, the larger the upper bound. As ν tends to infinity, the upper bound tends to $1/(1-p_A)$. Hence, if, for instance, the high-variance group is the minority ($p_A < 1/2$), then the gap cannot be larger than 2.

Numerical illustrations

In Fig. 3.3, we show the obtained utilities $\mathcal{U}$, the selection fractions r_A and the gap values $\mathcal{U}^{\text{dp}}/\mathcal{U}^{\text{obl}}$ and $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ for different budgets α from 0.01 to 0.99. Fig. 3.3a illustrates the utilities corresponding to different selection algorithms. We observe that the utilities of the Bayesian-optimal and demographic parity selections decrease when α increases. This is expected because this graph represents the average quality of a selected candidate: the average quality decreases when the number of selected candidates increases. What is more surprising is that the behavior of the group-oblivious selection algorithm is not monotonous: the expected utility $\mathcal{U}$ increases when α goes from 0.1 to 0.3. In fact, when $\alpha < 0.1$, very few B-candidates are selected by the group-oblivious algorithm. When $\alpha \approx 0.1$–0.2, this algorithm selects a few good B-candidates which leads to an increased average performance.

In Fig. 3.3c we show the performance gap between group-oblivious and demographic parity selection algorithms for different values of σ_A and fixed $\sigma_B = 0.2$, $\eta = 1$. The values of σ_A are such that $\sigma_A/\sigma_B = k$, $k = 2, 5, 10, 15$. We see that the gap is in general larger when the selection size α is small. This is due to the fact that as the selection size increases, the selections by the group-oblivious and demographic parity algorithms become close. The performance gap is zero when $\alpha = 0.5$ because the selections are exactly the same (due to the symmetry of the underlying quality distribution), but it becomes positive again for larger values of α. In addition, the larger the differential variance ratio σ_A^2/σ_B^2, the larger the gain that demographic parity brings.



(a) Utility $\mathcal{U}$ (b) Selection rate r_A (c) Ratio $\mathcal{U}^{\text{dp}}/\mathcal{U}^{\text{obl}}$ (d) Ratio $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$

Figure 3.3 – **Utility $\mathcal{U}$, selection rate r_A and performance gaps** for different budgets α. The parameters are $\mu = 1$, $\eta = 1$, $\sigma_B = 0.2$, and $p_A = 0.4$; $\sigma_A = 3$ for panels (a,b).

Finally, in Fig. 3.3d we illustrate the performance gap between the Bayesian-optimal and the demographic parity selection algorithms for different values of σ_A and fixed $\sigma_B = 0.2$, $\eta = 1$. As in Fig. 3.3c, the values of σ_A are such that $\sigma_A/\sigma_B = k$, $k = 2, 5, 10, 15$. In addition, we also show the bound on the ratio $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ for different values of α and for fixed $k = 15$. We see that the upper bound developed in Theorem 3.4.4 is relatively tight for small values of α, but is quite loose when $\alpha \approx 0.5$.

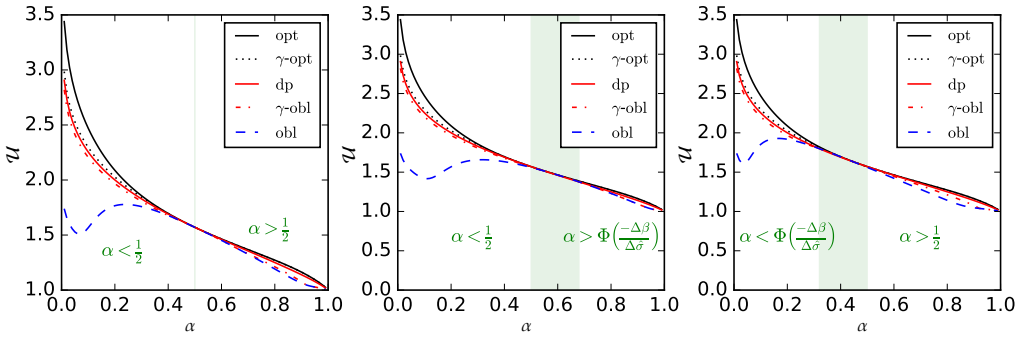
3.5.2 Group-independent latent quality distribution and biased estimates

In this section, we again assume that the underlying quality distribution is group-independent but we now assume that the estimates are both biased and with differential variance. Since the true quality distribution is group-independent, then $\mu = \mu_A = \mu_B$ and $\eta^2 = \eta_A^2 = \eta_B^2$. Recall that in this case, the conditions in Theorem 3.4.3 hold since under the assumption of $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$ which is w.l.o.g, the requirement $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$ is also satisfied. The expressions for the budgets $\alpha^{\min}$ and $\alpha^{\max}$ specified in Theorem 3.4.3 can also be simplified:

$$\alpha^{\min} = \min\left\{\Phi\left(\frac{-\Delta\beta}{\Delta\hat{\sigma}}\right), \frac{1}{2}\right\}, \alpha^{\max} = \max\left\{\Phi\left(\frac{-\Delta\beta}{\Delta\hat{\sigma}}\right), \frac{1}{2}\right\}.$$

We can get several insights from this simplification. First, if both groups are subject to the same amount of bias, $\beta_A = \beta_B$, then both $\alpha^{\min}$ and $\alpha^{\max}$ coincide, $\alpha^{\min} = \alpha^{\max} = 1/2$. Hence, according to Theorem 3.4.3, the γ-rule fairness mechanism in this case is beneficial to the utility of the group-oblivious algorithm for all budgets $\alpha \neq 1/2$. For $\alpha = 1/2$, both the γ-fair group-oblivious algorithm and the group-oblivious algorithm will perform the same $\mathcal{U}^{\gamma\text{-obl}} = \mathcal{U}^{\text{obl}}$ for all $\gamma > 0$. Hence, if the amount of bias is the same, the result is not different from the one when there is no bias at all (see Section 3.5.1), which is natural and expected. We illustrate this result in Fig. 3.4a which is the same as Fig. 3.3a.

Second, if the estimate for the high-variance group A has smaller bias than for the low-variance group B, i.e., $\Delta\beta = \beta_A - \beta_B < 0$, then the γ-fair mechanism will improve the utility of the group-oblivious algorithm for all $\alpha < 1/2$. It can be seen from the fact



(a) $\beta_A = \beta_B$ (b) $\beta_A < \beta_B$ (c) $\beta_A > \beta_B$

Figure 3.4 – **The quality of selection in the presence of bias and differential variance** for different budgets α. We assume that the quality distribution is group-independent, but A-candidates have larger variability of estimation compare to the B-candidates, i.e. $\sigma_A^2 > \sigma_B^2$. The quality distribution follows $\mathcal{N}(\mu = 1, \eta^2 = 1)$, the differential variance parameters are equal to $\sigma_A = 3$ and $\sigma_B = 0.2$. The bias parameters are equal to $\beta_A = 1, \beta_B = 1$ for 3.4a, $\beta_A = 0, \beta_B = 1$ for 3.4b and $\beta_A = 1, \beta_B = 0$ for 3.4c. The shaded green region indicates the case $\alpha \in [\alpha^{\min}, \alpha^{\max}]$, *i.e.*, when no increase of performance is guaranteed by Theorem 3.4.3.

that in this case $\Phi\left(\frac{-\Delta\beta}{\Delta\hat{\sigma}}\right) \geq 1/2$ which means that $\alpha^{\min} = 1/2$. This case is illustrated in Fig. 3.4b.

Perhaps counterintuitively, when implicit bias and implicit variance both affect the estimation, for some values of $\alpha \in (\alpha^{\min}, \alpha^{\max})$ specified in Theorem 3.4.3, the γ-fair group-oblivious algorithm will always perform worse than the group-oblivious algorithm. We observe the corresponding phenomenon on both Fig. 3.4b and Fig. 3.4c around the values of budgets $\alpha = 0.6$ and $\alpha = 0.4$, respectively.

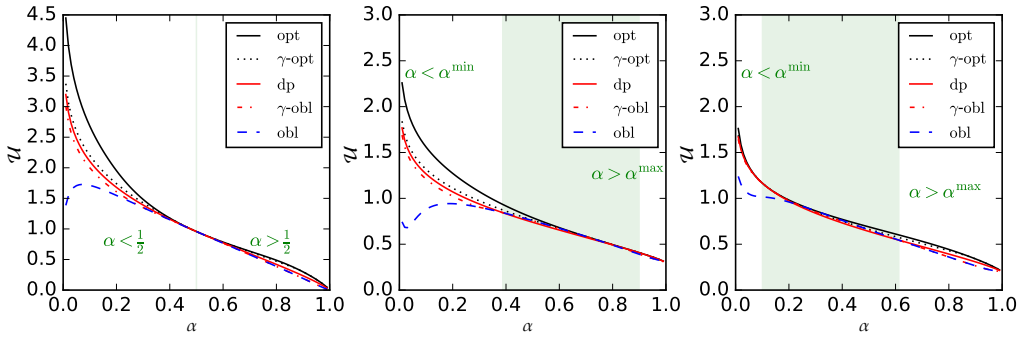
Finally, note that the Bayesian-optimal algorithm as well as the demographic parity (implicitly) remove the biases, hence, the results and discussion from Corollary 3.5.3 can also be applied in this section.

3.5.3 Group-dependent latent quality distribution and unbiased estimates

We now assume that there is *no bias* but that the underlying *quality distribution is group-dependent*. We can also distinguish different cases, when we isolate the effect of group-dependency of the distribution of quality by removing the implicit bias from our consideration. In this case, $\Delta\beta = 0$ and the budgets specified in Theorem 3.4.3 can be reformulated as follows:

$$\alpha^{\min} = \min\left\{\Phi\left(\frac{\Delta\mu}{\Delta\hat{\sigma}}\right), \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)\right\}, \quad \alpha^{\max} = \max\left\{\Phi\left(\frac{\Delta\mu}{\Delta\hat{\sigma}}\right), \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)\right\}.$$

We can draw several conclusions from this simplification. First, if both groups have equal means $\mu_A = \mu_B$ and if $\hat{\sigma}_A > \hat{\sigma}_B$, $\tilde{\sigma}_A < \tilde{\sigma}_B$, then the condition in Theorem 3.4.3



(a) $\mu_A = \mu_B$ (b) $\mu_A < \mu_B$ (c) $\mu_A > \mu_B$

Figure 3.5 – **The quality of selection in the presence of differential variance** for different budgets α. We assume that A-candidates have larger variability of their estimates compare to B-candidates $\hat{\sigma}_A > \hat{\sigma}_B$ as well as the relative amount of noise is larger for A-candidates than for B-candidates $\tilde{\sigma}_A < \tilde{\sigma}_B$. The implicit variance parameters are equal to $\sigma_A = 3$ and $\sigma_B = 1$. The distribution of quality is $\mathcal{N}(\mu_A = 0, \eta_A = 1)$ and $\mathcal{N}(\mu_B = 0, \eta_B = 2)$ for 3.5a, $\mathcal{N}(\mu_A = 0, \eta_A = 1)$ and $\mathcal{N}(\mu_B = 0.5, \eta_B = 1)$ for 3.5b and $\mathcal{N}(\mu_A = 0.5, \eta_A = 1)$ and $\mathcal{N}(\mu_A = 0, \eta_B = 1)$ for 3.5c. The shaded green region indicates the case $\alpha \in [\alpha^{\min}, \alpha^{\max}]$, *i.e.*, when no increase of performance is guaranteed by Theorem 3.4.3.

simplifies to $\alpha^{\min} = \alpha^{\max} = 1/2$, which is equivalent to the result in Corollary 3.5.2. Thus, in this case, the γ-rule mechanism improves the quality of group-oblivious selection for all budgets $\alpha \neq 1/2$. (If $\alpha = 1/2$, then $\mathcal{U}^{\text{obl}} = \mathcal{U}^{\gamma\text{-obl}}$ for all γ.) We illustrate this case in Fig. 3.5a and it is the same result as in Section 3.5.1. Second, if both groups have equal variances of quality $\eta_A^2 = \eta_B^2 = \eta^2$, then the condition $\tilde{\sigma}_A < \tilde{\sigma}_B$ from Theorem 3.4.3 holds automatically. We illustrate different cases of relations between μ_A and μ_B in Fig. 3.5b and Fig. 3.5c. Unfortunately, the bound on $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\gamma\text{-opt}}$ in Theorem 3.4.4 cannot be further simplified for the case of group-dependent quality distribution.

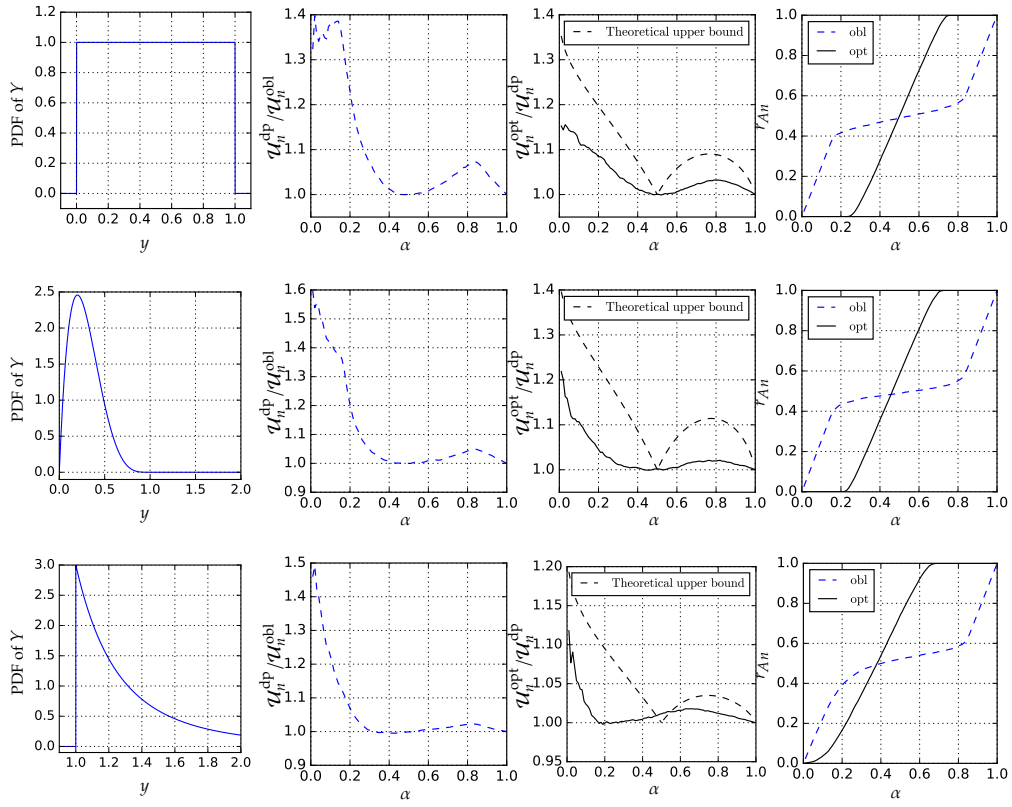
3.6 Experiments

In this section,[8] we challenge our theoretical results by using sets of data that do not satisfy our assumptions. We show in Section 3.6.1 that the results are qualitatively similar when the candidates' true quality comes from a non-normal distribution. We also observe a similar behavior when considering in Section 3.6.2 an artificial scenario that we construct using a real dataset coming from the national Indian exam data. We conclude in Section 3.6.3 with experiments that show that a case with $n = 50$ candidates behaves similarly as with $n = \infty$.

[8] All codes are available at: `https://gitlab.inria.fr/vemelian/differential-variance-code` .

3.6.1 Synthetic data with non-normal quality

Our assumption in the theoretical evaluation of Sections 3.4-3.5 was that qualities Y follow a normal distribution. In some cases, however, the quality distribution is quite different from normal and can be better modeled by a power law (Kleinberg and Raghavan, 2018), this for example the case for wealth, income or number of citations (Clauset et al., 2009), meaning that a minority possesses a large fraction of the aggregate quality. In this experiment, we vary the quality distribution and consider other distributions of quality Y: a Uniform distribution on $[0, 1]$, a Beta distribution with the shape parameter equal to 2 and the scale parameter equal to 5 or a Pareto distribution with a scale 1 and shape 3 (whose PDF $p_Y(y) = \frac{3}{y^4}$). We generate a single dataset of size $n = 10,000$. For this dataset we perform a group-oblivious, a demographic parity and a Bayesian-optimal selections. In Fig. 3.6, we report the sample utilities $\mathcal{U}_n$ and sample selection rates r_{An}. Note that in this section we consider no bias and group-independent quality distribution. Each line correspond to a different prior quality.



(a) Quality PDF (b) Ratio $\mathcal{U}^{dp}/\mathcal{U}^{obl}$ (c) Ratio $\mathcal{U}^{opt}/\mathcal{U}^{dp}$ (d) Selection rate r_A

Figure 3.6 – **Synthetic data** for different prior distributions of quality Y: Uniform on $[0, 1]$, Beta(2,5), and Pareto(1,3): Effects of fairness on utility $\mathcal{U}$ and selection rate r_A. The parameters are $p_A = 0.4$, $\sigma_A = 3$ and $\sigma_B = 0.2$. The number of candidates is fixed to $n = 10,000$.

In Fig. 3.6b, we show the performance gap between the group-oblivious and the demographic parity algorithms. We see that the demographic parity improves the utility of the group-oblivious algorithm in most of the cases and that the largest gap corresponds to the smallest budget α. Note that contrary to Corollary 3.5.2, the demographic parity does not always improve the utility of the group-oblivious algorithm. Yet, the loss due to the demographic parity is never larger than 0.1% while the gain can be up to 60%.

In Fig. 3.6c, the performance ratio for the Bayesian-optimal algorithm and the demographic parity algorithm is shown. As expected, the demographic parity harms the utility of the Bayesian-optimal algorithm for both small and large values of budget α. As the budget α increases, the performance gap decreases. To estimate the ratio between the Bayesian-optimal algorithm and the demographic parity, we plot also the value of upper bound from Theorem 3.5.3 that is calculated under an assumption that the quality distribution is normal. We observe that the bound is not tight, however, still dominates the values of $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ for most values of budget α.

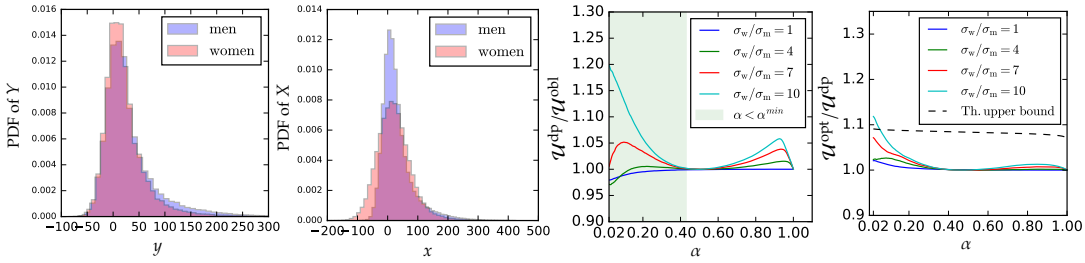
Finally, in Fig. 3.6d, we show how the selection fractions r_A^{obl} and r_A^{opt} depend on α. We see that for small budgets α, the group-oblivious algorithm tends to select more from group A, while for large budgets, the situation is opposite. In contrast, the Bayesian-optimal algorithm always selects A-candidates at lower rate if the selection budget α is small.

3.6.2 IIT-JEE scores dataset

In this section, we aim to consider a scenario in which the underlying quality distributions are non-normal and non-symmetric, and are group-dependent. To easily construct such a case, we create an artificial scenario by using a real dataset, the IIT-JEE dataset (*IIT-JEE dataset* 2019), with joint entrance exam results in India in 2009. These scores are used as an admission criterion to enter the high-rated universities. The dataset consists of $n = 384,977$ records. Every record has information about one student: its name, gender, grade for Mathematics, Physics, Chemistry and total grade. In the dataset, there are 98,028 women and 286,942 men. This dataset is the same as the one considered by Celis et al. (2020).

In order to construct a model of differential variance, we consider an artificial scenario where the field "grade" is the true latent quality Y of the candidates. The mean values and standard deviations of Y for the two groups are: $\mu_{\text{men}} = 30.8$, $\eta_{\text{men}} = 51.8$, $\mu_{\text{women}} = 21.2$, $\eta_{\text{women}} = 39.3$. We then suppose that an unbiased estimator X of the grade is observed. The standard deviation of estimation for male candidates is set to $\sigma_{\text{men}} = 10$. For the women group, which is the minority group, we consider different cases: $\sigma_{\text{women}} = k \cdot \sigma_{\text{men}}$, for $k = 1, 4, 7, 10$. The distribution of grades Y and observed values X for $k = 4$ are shown in Fig. 3.7a and 3.7b.

For the dataset we perform a group-oblivious (select best m), a demographic parity selection (select best m, but maintain the demographic parity condition $r_A = r_B$ up to one candidate) and a Bayesian-optimal selection. The selection size varies from 2% to



(a) Histogram of Y (b) Histogram of X (c) Ratio $\mathcal{U}^{dp}/\mathcal{U}^{obl}$ (d) Ratio $\mathcal{U}^{opt}/\mathcal{U}^{dp}$

Figure 3.7 – Distribution of Y and X given gender, and selection ratios for **IIT-JEE dataset** (*IIT-JEE dataset* 2019). Mean values and standards deviations of Y for two groups are: $\mu_m = 30.8$, $\eta_m = 51.8$, $\mu_w = 21.2$, $\eta_w = 39.3$. Added noise has standard deviation $\sigma_m = 10$ and $\sigma_w = k \cdot \sigma_m$; $k = 4$ in plot (b).

100% of total number of candidates, i.e., out of 384,977 students the decision maker selects 7,700 students or more. A selection rate of 2% was set by IIT in 2009 (Celis et al., 2020).

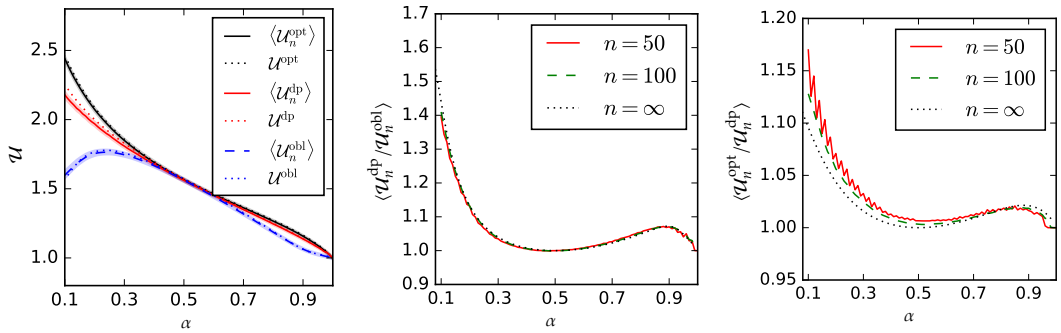
The results for the ratio of $\mathcal{U}^{dp}/\mathcal{U}^{obl}$ are given in Fig. 3.7c. We observe that for both small and large values of α, the demographic parity helps the utility of the group-oblivious algorithm, if the noise values of women evaluation σ_{women} are large, which agrees with the results from Theorem 3.4.3. We see that the gain can be up to around 20% if the selection size is small and up to 5% if the selection size is large. For the case where σ_{women} and σ_{men} are close, we observe no gain if the selection is large and we observe a minor loss in utility (around 2%) if the selection is small. This is due to the fact that in the dataset, there are more men with a high true latent quality Y, as seen in Fig. 3.7a. We also plot the region (for $k = 10$) from Theorem 3.4.3 in which the utility of the demographic parity algorithm should dominate the utility of the group-oblivious algorithm if the distribution of quality is a group-dependent normal.

Finally, on Fig. 3.7d we show the ratio $\mathcal{U}^{opt}/\mathcal{U}^{dp}$ for different values of $k = 1, 4, 7, 10$. In addition to these ratios, we also plot the bound from Theorem 3.4.4 for $k = 10$. We see that the bound is quite close to the actual value of $\mathcal{U}^{opt}/\mathcal{U}^{dp}$ for small α.

3.6.3 Accuracy of the approximation for small n

As discussed in Section 3.3, we cannot solve the problem with finite selection sizes exactly. Instead, we use an approximation that is exact as number of candidates n tends to infinity (Theorem 3.3.1). In this section, we question the accuracy of this approximation when the number n of candidates is relatively small. For our experiment, we generate datasets of different sizes $n = 50, 100$. For every size parameter n, we generate $10,000$ different datasets. For a population size n, we denote by $\langle \mathcal{U}_n \rangle$ the average quality of the selected candidates over our $10,000$ experiments. In each case, the true latent qualities Y are generated from a normal distribution $\mathcal{N}(1, 1)$.

In Fig. 3.8a we plot the average utilities $\langle \mathcal{U}_n \rangle$ for a population of $n = 100$, where



(a) Avg. utility $\langle \mathcal{U}_n \rangle$ ($n = 100$) (b) Ratio $\langle \mathcal{U}_n^{\text{dp}} / \mathcal{U}_n^{\text{obl}} \rangle$ (c) Ratio $\langle \mathcal{U}_n^{\text{opt}} / \mathcal{U}_n^{\text{dp}} \rangle$

Figure 3.8 – **Finite population size**: quality of selection and expected gain of the demographic parity over the group-oblivious algorithm. The quality distribution Y is $\mathcal{N}(\mu = 1, \eta^2 = 1)$ and the noise parameters are $\sigma_A = 3$, $\sigma_B = 0.2$. The number of experiments per set of parameters is $K = 10,000$. The shaded areas are the confidence intervals (corresponding to one standard deviation on the estimation of the empirical mean).

we select m individuals and where we vary m from 10 to 100. The shaded region corresponds to a confidence interval. We consider three selection algorithms (demographic parity, group-oblivious and Bayesian-optimal) and compare the performance for $n = 100$ with the limiting quantities $\mathcal{U}^{\text{dp}}$, $\mathcal{U}^{\text{obl}}$ and $\mathcal{U}^{\text{opt}}$. We observe that, even for $n = 100$, the average values of utility are close to the approximation. In Fig. 3.8b we compare the average ratio of performances $\langle \mathcal{U}_n^{\text{dp}} / \mathcal{U}_n^{\text{obl}} \rangle$ for different n. We observe that the approximation for $n = 50$ is a good prediction of the average gain provided by the use of demographic parity. Similarly, in Fig. 3.8c, we compare the average ratio of performances $\langle \mathcal{U}_n^{\text{opt}} / \mathcal{U}_n^{\text{dp}} \rangle$ for different n. Again, the curves for finite n are still quite close to the case where $n \to \infty$.

3.7 Conclusion and Discussion

In this chapter, we study a simple model of the selection problem that captures the phenomenon of differential variance, that is, the decision maker has estimates of the candidates' quality with different variances for different demographic groups. We distinguish two notable cases. In the first case, the decision maker does not have information about the estimate properties (variances and biases); as a result they use a group-oblivious algorithm. In the second case, every information about the distribution of quality is known, and the decision maker is Bayesian-optimal.

First, we show that both baseline algorithms (without any fairness constraint) lead to discrimination. Then we identify conditions under which, in the first case, the γ-rule fairness mechanism (a generalization of the $4/5$ rule) leads to a higher selection utility compared to using the group-oblivious baseline. In the second setting, the γ-rule

mechanism is harmful to the utility of Bayesian-optimal baseline but we prove that the utility decrease is bounded. Overall, our results contribute to a recent thread of works identifying cases in which, contrary to conventional wisdom, imposing fairness mechanisms does not come at the cost of utility (or even if it does, that the cost is bounded). Beyond fitting a particular application in detail, our results are useful in thinking about the impact of possible policies. For instance, they can help evaluate the effect of imposing a given fairness mechanism, or deciding whether or not to allow access to group information in a particular application.

Our theoretical results are obtained under the assumption that the true latent quality Y follows a normal law (to allow for analytical derivations). This assumption can be relaxed: we can plug into the model any distribution of latent quality (e.g., Pareto, uniform, etc.). We show numerically in Section 3.6 that it does not change the flavor of the main results. Extending these results theoretically is, however, challenging as in our proofs we operate with the expression for the conditional expectation of true latent quality given the noisy estimate. In a non-normal case, this conditional expectation cannot, in general, be expressed in closed form, which complicates the analysis.

Our modeling assumptions imply that a candidate's quality does not depend on the selection strategy used. If attaining a certain level of quality comes at a cost, then the interaction between decision makers and candidates may be seen as a game. It would be interesting to see how differential variance affects the incentives of candidates and how the γ-rule changes them in this game. We consider this game-theoretic formulation of the selection problem in Chapter 4.

Throughout this chapter, we have studied the effect of imposing demographic parity at the first stage on the final selection utility. However, a natural question to ask is, for example, what is the effect of imposing fairness at the first stage on the fairness of the second stage selection in case of a two-stage selection process? We leave a detailed investigation of this aspect of selection problems to Chapter 5.

3.8 Omitted Proofs

In this section we provide detailed proofs of the statements given before. Namely, these are proofs of Theorem 3.4.1, Theorem 3.4.2, Theorem 3.4.3 and Theorem 3.4.4.

3.8.1 Proof of Theorem 3.4.1

By our assumptions, the estimates of qualities for G-candidates follow a normal law with the mean $\mu_G - \beta_G$ and the variance $\hat{\sigma}_G^2 = \eta_G^2 + \sigma_G^2$. Recall that the selection rate r_G^{obl} for the group-oblivious algorithm is a probability for the G-candidate to have an estimated quality larger than a predefined group-independent threshold: $r_G^{\text{obl}} = \mathrm{P}(X \geq \theta^{\text{obl}} | G)$. Taking all that into account, the selection rate for G-candidates

can be expressed as:

$$r_G^{\text{obl}} = 1 - \Phi\left(\frac{\theta^{\text{obl}} - \mu_G + \beta_G}{\hat{\sigma}_G}\right).$$

This shows that the condition $r_A^{\text{obl}} > r_B^{\text{obl}}$ is equivalent to $\frac{\theta^{\text{obl}} - \mu_A + \beta_A}{\hat{\sigma}_A} < \frac{\theta^{\text{obl}} - \mu_B + \beta_B}{\hat{\sigma}_B}$, since Φ is an increasing function of its argument. Hence, by rearranging the terms we conclude that $\theta^{\text{obl}} > \frac{\mu_A\hat{\sigma}_B - \mu_B\hat{\sigma}_A}{\hat{\sigma}_B - \hat{\sigma}_A} + \frac{\beta_B\hat{\sigma}_A - \beta_A\hat{\sigma}_B}{\hat{\sigma}_B - \hat{\sigma}_A}$. By substituting the corresponding threshold to the expression for the selection rate r_G^{obl}, we end up with the expression for the values of budgets α for which $r_A^{\text{obl}} > r_B^{\text{obl}}$. The calculations show that for the budgets $\alpha < 1 - \Phi\left(\frac{(\mu_A - \mu_B) - (\beta_A - \beta_B)}{\hat{\sigma}_B - \hat{\sigma}_A}\right) = \Phi\left(\frac{\Delta\mu - \Delta\beta}{\Delta\hat{\sigma}}\right)$ using the group-oblivious algorithm leads to overrepresentation of a high-variance group A, where we use the notation $\Delta\mu = \mu_A - \mu_B$, $\Delta\beta = \beta_A - \beta_B$ and $\Delta\hat{\sigma} = \hat{\sigma}_A - \hat{\sigma}_B$.

3.8.2 Proof of Theorem 3.4.2

The Bayesian-optimal algorithm selects candidates for which the expected quality $\widetilde{Y}$ is larger than some group-independent but budget-dependent threshold $\tilde{\theta}$. Since $\widetilde{Y}_G$ follows a normal law with the mean μ_G and the variance $\tilde{\sigma}_G^2$, we can write that $r_G^{\text{opt}} = 1 - \Phi\left(\frac{\tilde{\theta} - \mu_G}{\tilde{\sigma}_G}\right)$. In the rest of the proof, without loss of generality we assume that $\tilde{\sigma}_A^2 < \tilde{\sigma}_B^2$, hence, we can calculate that

$$r_A^{\text{opt}} < r_B^{\text{opt}} \iff \frac{\tilde{\theta} - \mu_A}{\tilde{\sigma}_A} > \frac{\tilde{\theta} - \mu_B}{\tilde{\sigma}_B} \iff \tilde{\theta} > \frac{\mu_A\tilde{\sigma}_B - \mu_B\tilde{\sigma}_A}{\tilde{\sigma}_B - \tilde{\sigma}_A}.$$

By substituting the corresponding threshold to the expression for the selection rate r_G^{opt}, we end up with the expression for the values of budgets α for which $r_A^{\text{opt}} < r_B^{\text{opt}}$. The calculations show that this is for all budgets $\alpha < \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)$, where we use the notation $\Delta\mu = \mu_A - \mu_B$ and $\Delta\tilde{\sigma} = \tilde{\sigma}_A - \tilde{\sigma}_B$.

3.8.3 Properties of the utility $\mathcal{U}$ (Proof of Theorem 3.4.3)

In this section, we study the properties of the utility function $\mathcal{U}(r_A)$ independently from the selection algorithm used. We give the expression for the derivative of $\mathcal{U}$ as a function of r_A. This expression allows us to prove that the utility function $\mathcal{U}$ is strictly concave. This implies that as we have $r_A^{\text{obl}} \le r_A^{\gamma\text{-obl}} \le r_A^{\gamma\text{-opt}} \le r_A^{\text{opt}}$ or $r_A^{\text{obl}} \le r_A^{\gamma\text{-obl}} \le r_A^{\gamma\text{-opt}} \le r_A^{\text{opt}}$, one always has $\mathcal{U}(r_A^{\text{obl}}) \le \mathcal{U}(r_A^{\gamma\text{-obl}}) \le \mathcal{U}(r_A^{\gamma\text{-opt}}) \le \mathcal{U}(r_A^{\text{opt}})$, with strict inequalities whenever the above inequalities are strict.

Lemma 3.8.1. *Assume that the budget α is fixed.*

1. *The first derivative of the utility $\mathcal{U}(r_A)$ can be expressed as follows:*

$$\mathcal{U}'(r_A) = \frac{p_A}{\alpha}\left[\frac{(\theta_A + \beta_A)\eta_A^2 + \mu_A\sigma_A^2}{\eta_A^2 + \sigma_A^2} - \frac{(\theta_B + \beta_B)\eta_B^2 + \mu_B\sigma_B^2}{\eta_B^2 + \sigma_B^2}\right] \tag{3.5}$$

where θ_A, θ_B are such that $\mathrm{P}(X \geq \theta_A \mid G = A) = r_A$ and $\sum_{G\in\{A,B\}} \mathrm{P}(X \geq \theta_G \mid G)\cdot p_G = \alpha$.

2. *The utility $\mathcal{U}(r_A)$ is strictly concave.*

Proof. By definition of $\mathcal{U}$ in (3.4), the utility $\mathcal{U}$ equals $\mathcal{V}(\theta_A, \theta_B)$ where θ_A, θ_B are the unique thresholds such that $\mathrm{P}(X \geq \theta_A \mid G = A) = r_A$ and $\sum_{G\in\{A,B\}} \mathrm{P}(X \geq \theta_G \mid G) \cdot p_G = \alpha$. Using that $\widetilde{Y}_G$ and X_G are normally distributed these quantities can be expressed as:

$$\mathcal{V}(\theta_A, \theta_B) = \frac{1}{\alpha}\sum_G p_G \int_{\theta_G}^{\infty} dx \int_{-\infty}^{\infty} dy \left[y \cdot \frac{1}{\eta_G}\phi\left(\frac{y-\mu_G}{\eta_G}\right) \cdot \frac{1}{\sigma_G}\phi\left(\frac{x-y+\beta_G}{\sigma_G}\right)\right],$$
$$r_G(\theta_G) = \int_{\theta_G}^{\infty} dx \int_{-\infty}^{\infty} dy \left[\frac{1}{\eta_G}\phi\left(\frac{y-\mu_G}{\eta_G}\right) \cdot \frac{1}{\sigma_G}\phi\left(\frac{x-y+\beta_G}{\sigma_G}\right)\right].$$

Using the chain rule, we can write the first derivative of selection utility:

$$\frac{d\mathcal{U}}{dr_A} = \sum_G \frac{\partial \mathcal{V}}{\partial \theta_G}\frac{d\theta_G}{dr_A}. \tag{3.6}$$

From the budget constraint $p_A r_A + p_B r_B = \alpha$, by differentiating both parts by r_A, we obtain that $p_A \frac{dr_A}{dr_A} + p_B \frac{\partial r_B}{\partial \theta_B}\frac{d\theta_B}{dr_A} = 0$ which implies that $\frac{d\theta_B}{dr_A} = -\frac{p_A}{p_B}\frac{\partial \theta_B}{\partial r_B}$. Then, by substituting the obtained expression for $\frac{d\theta_B}{dr_A}$ into (3.6), we obtain that $\frac{d\mathcal{U}}{dr_A} = p_A\left(\frac{\partial \mathcal{V}}{\partial \theta_A}\frac{\partial \theta_A}{\partial r_A} - \frac{\partial \mathcal{V}}{\partial \theta_B}\frac{\partial \theta_B}{\partial r_B}\right)$. From this, the expression (3.5) follows directly.

We observe that the first derivative is linear in the selection thresholds θ_A and θ_B. Thus, as the selection rate r_A increases, the derivative $\mathcal{U}'_{r_A}$ decreases which means that the function $\mathcal{U}(r_A)$ is strictly concave. □

3.8.4 Proof of Theorem 3.4.4

Assume that $\alpha < \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)$. By Theorem 3.4.2, we have $r_A^{\mathrm{opt}} < r_A^{\mathrm{dp}}$. As we prove in Lemma 3.8.1, the utility function $\mathcal{U}$ is concave function of r_A. Using the concavity of $\mathcal{U}$ we have $\frac{\mathcal{U}(r_A^{\mathrm{opt}}) - \mathcal{U}(r_A^{\mathrm{dp}})}{r_A^{\mathrm{opt}} - r_A^{\mathrm{dp}}} \geq \mathcal{U}'(r_A = r_A^{\mathrm{dp}})$ which implies that $\mathcal{U}(r_A^{\mathrm{opt}}) - \mathcal{U}(r_A^{\mathrm{dp}}) \leq (0-\alpha)\cdot\mathcal{U}'(r_A^{\mathrm{dp}})$, where we use the fact that $r_A^{\mathrm{dp}} - r_A^{\mathrm{opt}} \leq \alpha$ for all budgets $\alpha < \Phi\left(\frac{\Delta\mu}{\Delta\tilde{\sigma}}\right)$.

By dividing both sides by $\mathcal{U}(r_A^{\mathrm{dp}})$, from the above inequality we obtain the following upper bound:

$$\frac{\mathcal{U}(r_A^{\mathrm{opt}})}{\mathcal{U}(r_A^{\mathrm{dp}})} \leq 1 - \alpha \cdot \frac{\mathcal{U}'(r_A^{\mathrm{dp}})}{\mathcal{U}(r_A^{\mathrm{dp}})},$$

The expression for $\mathcal{U}'(r_A^{\mathrm{dp}})$ can be written explicitly by using (3.5) and the fact that group-dependent thresholds for the demographic parity algorithm can be calculated

as $\theta_G^{\text{dp}} = \mu_G - \beta_G + \hat{\sigma}_G \Phi^{-1}(1-\alpha)$:

$$\mathcal{U}'(r_A^{\text{dp}}) = \frac{p_A}{\alpha}\left(\mu_A - \mu_B + \Phi^{-1}(1-\alpha)\left[\frac{\eta_A^2}{\sqrt{\sigma_A^2+\eta_A^2}} - \frac{\eta_B^2}{\sqrt{\sigma_B^2+\eta_B^2}}\right]\right) = \frac{p_A}{\alpha}\left(\Delta\mu + \Phi^{-1}(1-\alpha)\Delta\tilde{\sigma}\right).$$

The utility by the demographic parity algorithm can be calculated using the law of total expectation and the expected value of truncated normal distribution as follows:

$$\mathcal{U}(r_A^{\text{dp}}) = \sum_G p_G \mu_G + \frac{\phi\left(\Phi^{-1}(1-\alpha)\right)}{\alpha}\sum_G p_G \frac{\eta_G^2}{\sqrt{\sigma_G^2+\eta_G^2}} = \sum_G p_G \mu_G + \frac{\phi\left(\Phi^{-1}(1-\alpha)\right)}{\alpha}\sum_G p_G \tilde{\sigma}_G.$$

Hence, from the above inequality and the expressions for $\mathcal{U}(r_A^{\text{dp}})$ and $\mathcal{U}'(r_A^{\text{dp}})$, we can obtain the following upper bound on the ratio $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ for $\alpha < \Phi(\Delta\mu/\Delta\tilde{\sigma})$[9]:

$$\frac{\mathcal{U}(r_A^{\text{opt}})}{\mathcal{U}(r_A^{\text{dp}})} \leq 1 - \alpha \cdot \frac{p_A}{\alpha}\frac{\Delta\mu + \Phi^{-1}(1-\alpha)\Delta\tilde{\sigma}}{\sum_G p_G\mu_G + \frac{\phi\left(\Phi^{-1}(1-\alpha)\right)}{\alpha}\sum_G p_G\tilde{\sigma}_G}.$$

For the γ-fair Bayesian-optimal algorithm, the upper bound on $\mathcal{U}^{\text{opt}}/\mathcal{U}^{\text{dp}}$ can be calculated in a similar manner. The values of the selection rate difference for $\alpha < \Phi(\Delta\mu/\Delta\tilde{\sigma})$ can be upper bounded as $r_A^{\gamma\text{-opt}} - r_A^{\text{opt}} \leq \frac{\alpha}{p_A + p_B/\gamma}$, since the selection by the Bayesian-optimal algorithm lies either inside the γ-region $r_A \in \left[\frac{\alpha}{p_A+p_B/\gamma}, \frac{\alpha}{p_A+\gamma p_B}\right]$ or on its boundary. For $\alpha > \Phi(\Delta\mu/\Delta\tilde{\sigma})$, the difference can be upper bounded as $r_A^{\text{opt}} - r_A^{\gamma\text{-opt}} \leq 1 - \frac{\alpha}{p_A+\gamma p_B}$.

[9]Note that the case $\alpha > \Phi(\Delta\mu/\Delta\tilde{\sigma})$ is proven similarly, except that we use $r_A^{\text{opt}} - r_A^{\text{dp}} \leq 1 - \alpha$.

Chapter 4

Selection with Differential Variance in the Strategic Setting

This chapter is based on our publication (Emelianov et al., 2022b). To have a consistent notation across all chapters, we slightly modified the notations by changing W to Y, $\hat{W}$ to X, $\tilde{W}$ to $\tilde{Y}$, and x to r.

The code to generate all figures can be found at:
https://gitlab.inria.fr/vemelian/strategic-selection-code

Abstract To better understand discriminations and the effect of affirmative actions in selection problems (e.g., college admission or hiring), in Chapter 3 we proposed a model based on *differential variance*. This model assumes that the decision-maker has a noisy estimate of each candidate's quality and puts forward the difference in the noise variances between different demographic groups as a key factor to explain discrimination. The literature on differential variance, however, does not consider the strategic behavior of candidates who can react to the selection procedure to improve their outcome, which is well-known to happen in many domains.

In this chapter, we study how the strategic aspect affects fairness in selection problems. We propose to model selection problems with strategic candidates as a contest game: A population of rational candidates compete by *choosing* an effort level to increase their quality. They incur a cost-of-effort but get a (random) quality whose expectation equals the chosen effort. A Bayesian decision-maker observes a noisy estimate of the quality of each candidate (with differential variance) and selects the fraction α of best candidates based on their posterior expected quality; each selected candidate receives a reward S. We characterize the (unique) equilibrium of this game in the different parameters' regimes, both when the decision-maker is unconstrained and when they are constrained to respect the fairness notion of demographic parity. Our results reveal important impacts of the strategic behavior on the discrimination observed at equilibrium and allow us to understand the effect of imposing demographic

parity in this context. In particular, we find that, in many cases, the results contrast with the non-strategic setting. We also find that, when the cost-of-effort depends on the demographic group (which is reasonable in many cases), then it entirely governs the observed discrimination (i.e., the noise becomes a second-order effect that does not have any impact on discrimination). Finally we find that imposing demographic parity can sometimes increase the quality of the selection at equilibrium; which surprisingly contrasts with the optimality of the Bayesian decision-maker in the non-strategic case. Our results give a new perspective on fairness in selection problems, relevant in many domains where strategic behavior is a reality.

4.1 Introduction

Recent literature on fairness in selection problems analyzed the problem using models based on two key ingredients to explain discrimination. Kleinberg and Raghavan (2018) model selection problems with *implicit bias* (see also (Celis et al., 2021, 2020; Mehrotra et al., 2022)), that is, where the decision-maker implicitly under-evaluates the quality of the candidates from disadvantaged demographic groups. On the other hand, in Chapter 3, following ideas from the economics literature on *statistical discrimination* (see Section 4.2), we assume that the decision maker's estimate of the candidates quality is unbiased but has a higher variance for some demographic groups (a phenomenon terms *implicit* or *differential variance*). With both types of models, the authors study the discrimination that comes out of baseline decision-makers, and the impact of imposing fairness mechanisms such as the Rooney rule or the four-fifths rule.

In the above literature, the characteristics of the candidates used for selection (in particular, their qualities) are assumed to be fixed and exogenous—i.e., they do not depend on the selection procedure. In practice, however, candidates (i.e., individuals) involved in selection problems can *adapt* to the selection procedure in order to increase the chances of a positive outcome. Such a *strategic behavior* is observed in many domains (Patro et al., 2022; Woolley, 2017). A recent thread of literature on *strategic classification* is devoted to modeling and analyzing the impact of this strategic behavior on classification problem (Braverman and Garg, 2020; Dong et al., 2018; Hardt et al., 2016a; Kleinberg and Raghavan, 2019; Miller et al., 2020; Milli et al., 2019; Tsirtsis and Gomez Rodriguez, 2020; Zhang et al., 2019). The selection problem, however, is fundamentally different from a classification problem in that the number of positive predictions is constrained—this will in particular lead to competition between individuals, see below. Moreover, this thread of literature did not investigate discrimination issues. This leaves open the key question, which is our focus in this chapter: *How does the strategic behavior of candidates affect discrimination and the impact of fairness mechanisms in selection problems?*

The selection problem with strategic candidates modeled as a contest game In this chapter, we propose to model the selection problem with strategic candidates as a contest game (that is, roughly, as a game where candidates compete for a reward). We

consider a population of candidates. Each candidate chooses an effort level m that they exert to improve their quality, at a quadratic cost $Cm^2/2$ (where C is a constant coefficient). Each candidate has a latent quality Y drawn randomly whose expected value is equal to their selected effort m. A Bayesian decision-maker observes a noisy estimate X of the quality Y and selects a fraction α of the best candidates based on the posterior expected quality $\widetilde{Y} = \mathrm{E}(Y|X)$. All selected candidates receive a reward of size S, which is a quantitative measure of the benefit that the selection brings to a candidate (e.g., a job position or an education).

In our model, we assume that the population of candidates is divided in two groups: the high-noise (H) and the low-noise (L) group (which refers to the noise in the estimate X of the candidates quality). This group-dependent noise represents a form of information inequality common across different demographic groups: decision-makers are less familiar with candidates from certain groups (the H-candidates), such that they are less able to precisely estimate the qualities of these candidates. This phenomenon is called *differential variance* (see Chapter 3). It is at the basis of the economic theory of statistical discrimination and it was first described by (Phelps, 1972) to explain the racial inequality of wages. In addition to group-dependent noise, we also consider a group-dependent cost coefficient (i.e., $C_H \neq C_L$). The group-dependent cost coefficient models socioeconomic inequalities; e.g., for students from low-income families it is typically harder to reach a desired level of quality due to costly preparatory courses.

Overview of our results Our model of a selection problem with strategic candidates defines a (population) game. We first show that this game has a unique Nash equilibrium. Then we focus primarily on characterizing the equilibrium in the regime of large rewards S, which corresponds roughly to high-stake selection problems. In this regime, **we show, *at equilibrium*, the following discrimination resulting from our model**:

(i) If the cost coefficient is group-independent ($C_H = C_L$), then the high-noise candidates make a greater effort in average than the low-noise candidates. The latter are underrepresented[1] in the selection, which is counterintuitive and is in contrast with the results in the non-strategic setting studied in Chapter 3.

(ii) If the cost coefficient is group-dependent ($C_H \neq C_L$), then the noise level does not affect the outcome of the game. The cost-advantaged candidates (those for whom the cost coefficient is smaller) make a greater average effort compared to the cost-disadvantaged candidates, and the latter are underrepresented, irrespective of the noise. The noise is a second order effect compared to the cost difference, which totally dominates. This offers a potential explanation as to why low-noise

[1] "Underrepresented" in our paper means "having less representation in the selection than its share in the candidates' population". This is a classical definition in the algorithmic fairness literature where the notion of demographic parity would mean that the two groups are equally represented.

candidates are often not underrepresented in practice: this is because the cost of effort for the high-noise candidates is usually large.

In both cases stated above, one of the groups of candidates is always underrepresented. A potential remedy for this is the so-called *demographic parity* mechanism, which imposes that the decision-maker selects candidates of all demographic groups at equal rates. Next, **we characterize the equilibrium when the Bayesian decision-maker is constrained by the demographic parity mechanism** (still in the regime of large rewards S):

(i) We show that the demographic parity mechanism tends to equalize the average effort of the two groups compared to the unconstrained decision-maker: in most cases, the previously underrepresented group makes a greater average effort and the previously overrepresented group makes a lower average effort.

(ii) We characterize the change in the selection quality (or utility from the decision-maker perspective) from imposing demographic parity. Interestingly, we find that in some cases, the selection quality can improve compared to the unconstrained selection. This is surprising since, in the non-strategic setting, the unconstrained Bayesian decision-maker is optimal (see Chapter 3). Our result shows that it is no longer true in the strategic setting as in certain cases imposing a fairness mechanism can improve the selection quality, even against a Bayesian baseline. In other cases, we bound the degradation of quality that can result from imposing demographic parity.

For the case of small rewards S, **we get further analytical results on the equilibrium characterization**. We find that the results are different from that of the case of large S, and are similar to the ones obtained in the non-strategic setting (see Chapter 3): the high-noise candidates are always underrepresented if the selection size α is small enough. This indicates that, if the reward is small, then the impact of the strategic aspect (on discrimination results) is not important. We perform numerical experiments to illustrate the case of intermediate rewards S and how it matches the case of small and large S in their respective regimes. Finally, **we study the convergence of different dynamics to the Nash equilibrium** (namely the best response and fictitious play dynamics). We observe that the trajectories of the best response dynamics converge to limit cycles that have a higher average effort (for both groups) than at equilibrium, whereas the fictitious play dynamics seems to converge to equilibrium in our empirical results.

Implications of our results Our results show that it is crucial to take into account the strategic nature of candidates involved in selection problem in order to understand discrimination and to predict the effect of imposing a fairness mechanism. They also show that discrimination in selection problems is a somewhat nuanced issue: it depends not only on the strategic aspect of the candidates but also on the range of assumed rewards and on the cost of effort of the different demographic groups. This

means that a policy-maker, when considering whether to impose a fairness mechanism in a particular application, should evaluate the population of candidates (their costs of effort, etc.). It is worth noting, also, that we presented our results for a Bayesian decision-maker who computes a posterior estimate of the candidates quality (with knowledge of the group-dependent distributions). Our results are technically easy to extend to a group-oblivious decision-maker sorting the candidates by quality estimate X irrespective of their group (another standard baseline); but in that case the high-noise and low-noise groups are reversed (we discuss that in Section 4.7). Hence, a policy-maker would also need to evaluate the baseline decision-maker they face.

Outline The rest of the chapter is organized as follows. In Section 4.2, we present the related literature. In Section 4.3, we formulate the problem in a game-theoretic framework. In Section 4.4, we show that, for the case of large rewards S, the selection with strategic candidates leads to underrepresentation of one of the two groups of candidates. In Section 4.5, we study how demographic parity affects the incentives of candidates and the expected quality of the selection for large S. In Section 4.6, we complement our theoretical results by studying the convergence to equilibria and the case of small and intermediate rewards S. We conclude with a discussion in Section 4.7. We provide all omitted proofs in Section 4.8.

4.2 Related Work

Statistical discrimination The theory of statistical discrimination, initiated by Phelps (1972) and Arrow (1973), considers the uncertainty of information about individual characteristics to explain racial/gender inequality in decision-making. Phelps (1972) develops a model where each individual possesses a latent quality drawn from a fixed group-independent distribution. A Bayesian decision-maker observes a noisy estimate of individual's quality, where the noise is symmetric and zero-mean but has a group-dependent variance. The decision-maker assigns a wage equal to the expected posterior quality. This model is used to explain racial inequalities in wages. Lundberg and Startz (1983) extend Phelps' model to a strategic setting by assuming two groups of workers that choose the values of effort according to a group-independent quadratic costs. The effort induces a quality that is assigned randomly but equal to the effort in expectation. The authors show that, in the equilibrium, the high-noise candidates make lower effort and are payed less on average compared to the low-noise candidates; but that if the decision-maker is restricted to not use the group information for wage assignment, then the effort is equal for both groups. In our work, we use a similar model but in the context of selection problems where a fraction of candidates receives a (fixed) reward rather than assigning a variable wage to all candidates as in (Lundberg and Startz, 1983; Phelps, 1972). We also extend the model of (Lundberg and Startz, 1983) to have group-dependent cost-of-effort, which as we shall see has a crucial effect on discrimination. Finally, we consider a different fairness mechanism, namely demographic parity.

Many statistical discrimination models (see e.g., (Aigner and Cain, 1977; Arrow, 1973; Coate and Loury, 1993) and a survey in (Fang and Moro, 2011)) assume that individuals of different groups have identical *a priori* characteristics (e.g., cost-of-effort), but that the decision-maker uses their group-dependent belief when there is imperfect information to assess the performance of individuals in a group. In some cases, the discriminating beliefs (stereotypes) of decision-makers lead to equilibria in which these stereotypes are fulfilled. In contrast, we consider group-dependent cost-of-effort and group-dependent noise variance. We prove that the game in our model attains a unique equilibrium, and that discrimination occurs due to the group-dependent characteristics—i.e., if they become group-independent in our model, discrimination disappears.

Selection problems in the non-strategic settings Recent literature investigate statistical discrimination in selection problems, under the name of *differential variance*. In (Emelianov et al., 2022b), we show that differential variance, in the case of Bayesian decision-maker, leads to the underrepresentation of the high-noise candidates (for selection fractions below 0.5). We also study how quota-based fairness mechanisms (the 80%-rule and the demographic parity) affect the fairness-utility tradeoff. Garg et al. (2021) study a similar setting but where the performance of candidates is measured by *multiple* independent and unbiased estimates (from tests). The authors study how affirmative actions and access to testing affect the disparity level and the quality of the selected cohort. Some works also study the selection problem under implicit bias rather than differential variance—i.e., the decision-maker has a non-noisy but biased estimate of the candidates' qualities. Kleinberg and Raghavan (2018) show that this type of bias naturally leads to underrepresentation of the disadvantaged group, and they study how a fairness mechanism called the Rooney rule affects the selection quality. This work is extended by Celis et al. (2021, 2020) and Mehrotra et al. (2022). Our work complements those studies by assuming *strategic candidates* who can respond to a policy chosen by a Bayesian decision-maker. Similarly to (Garg et al., 2021) and (Emelianov et al., 2022b), we assume that the quality estimates are affected by differential variance, but we do not model bias (Kleinberg and Raghavan, 2018) as we assume a Bayesian decision-maker who can correct for the bias. The main difference is that we assume that the quality distribution is not fixed but chosen by candidates to maximize their payoff which equals to the expected reward minus the cost-of-effort.

Fairness in contests A classical model of contest is given by Lazear and Rosen (1981): individuals make a costly effort m that induces a noisy quality $Y = f(m) + \varepsilon$, where f is an increasing function of the effort; the player having the largest quality wins the prize S. Fairness in contest games was studied from different perspectives in economics and computer science literature (Fu and Wu, 2019). Schotter and Weigelt (1992) assume a two-player contest with a cost-advantaged and a cost-disadvantaged player—each pays a quadratic cost-of-effort but with different coefficients. They show that, at equilibrium, the cost-disadvantaged player makes a lower effort. Then they

show that an affirmative action (adding a bonus to the score of the cost-disadvantaged player) leads all players to make lower efforts but increases the winning probability of the cost-disadvantaged player. Our model shares similarities with that of Schotter and Weigelt (1992) (in particular group-dependent quadratic costs), but also has important differences: we consider an infinite population of candidates and we include group-dependent noise. Our results are also different as we analyze the dependence on S and we study another fairness mechanism (demographic parity)—in particular we find that it increases the effort of previously underrepresented group and sometimes even of both groups.

Strategic ranking Following the strategic classification literature (see previous section), Liu et al. (2021) study the *strategic ranking* problem. They assume that each individual has a latent ranking and makes some costly effort to affect it. The effort induces a score $g(\text{effort}) \cdot f(\text{latent fixed rank})$, for some fixed strictly increasing functions f and g, which in turn results in a post-effort ranking. Liu et al. (2021) study how different ranking reward functions (with and without randomization) of a fixed selection capacity c affect the characteristics of individuals and of the selection at equilibrium (average welfare and scores). Finally, they consider a setting with two groups of candidates that differ in a multiplicative parameter $\gamma > 1$ affecting the score and study the welfare gap for groups as a function of c. In our model, we do not assume any pre-effort ranking—the score of an individual is purely determined by the effort and by a group-dependent noise. Our model is simpler compared to (Liu et al., 2021) and is designed to capture the effects of the group-dependent cost-of-effort and noise in selection problems. This allows us to state concrete results involving the parameters of interest (the cost coefficient and noise variance). We also consider how a fairness mechanism (demographic parity) affects the group representations and the quality of selection at equilibrium.

4.3 The Model

4.3.1 Candidates and decision-maker

Candidates model We assume a non-atomic game with a unit mass of candidates indexed by $i \in [0,1]$. There are two groups of candidates: H and L. The letter $G \in \{H, L\}$ will denote any of these two groups, and the proportion of candidates from group G is $p_G \in (0,1)$. Each candidate i has a quality that depends on the *effort* m_i that this candidate makes. In college admission or in hiring, the effort m_i of a candidate i can be interpreted as candidate's achievements. It might, for instance, represent the number of courses followed by a student, the quality of number of degrees obtained, etc. We assume that a candidate i that chooses to make an effort $m_i \geq 0$ has a quality Y_i that is normally distributed with mean m_i and variance η^2:

$$Y_i \sim \mathcal{N}(m_i, \eta^2).$$

Making an effort m costs a candidate from group G a quadratic cost $C_G m^2/2$. The population-dependent cost coefficient C_G can model socioeconomic factors like the income of the parents or the country of origin; these factors might make harder for some candidates than others to make the same effort m.

Decision-maker A decision-maker wants to select the candidates having the largest qualities. To do so, the decision-maker observes a noisy estimate X_i of the quality Y_i of each candidate i:

$$X_i = Y_i + \sigma_{G_i} \cdot \varepsilon_i,$$

where ε_i is a zero-mean centered random variable from the standard normal distribution $\mathcal{N}(0,1)$; the noise variance $\sigma^2_{G_i}$ is assumed group-dependent.[2] The quality estimate X_i is a noisy but unbiased measurement of the quality Y_i of a candidate i, e.g., an interview result. The group-dependent variance of noise σ^2_G models the information inequality: if interviewers are more familiar with candidates of some demographic groups, they are more confident in their evaluation compared to that of candidates of other groups. Without loss of generality, we assume that $\sigma^2_H > \sigma^2_L$. We, thus, refer to H-candidates as *high-noise* candidates, and we refer to L-candidates as *low-noise* candidates.

We assume that the decision-maker knows[3] the effort m of all candidates, as well as the variances η^2 and σ^2_G, and selects a proportion $\alpha \in (0,1)$ of the candidates. The decision-maker aims at maximizing the expected quality of selected candidates and, therefore, selects the α proportion having the largest expected quality $\widetilde{Y}$. Using the property of conditional expectation for bivariate normal random variables, we can write the expectation of quality Y_i given X_i as

$$\widetilde{Y}_i = \mathrm{E}(Y_i|X_i) = X_i \rho^2_{G_i} + (1 - \rho^2_{G_i}) m_i, \tag{4.1}$$

where $\rho_{G_i} = \eta/\sqrt{\eta^2 + \sigma^2_{G_i}} \in [0,1]$ is the correlation coefficient between the quality Y_i and its estimate X_i. Since $\widetilde{Y}_i$ is a linear function of X_i, and X_i is distributed normally, the expected quality $\widetilde{Y}_i$ follows a normal distribution with mean m_i and variance $\tilde{\sigma}^2_{G_i} = \eta^4/\left(\sigma^2_{G_i} + \eta^2\right)$. Note that the larger the value of noise $\sigma^2_{G_i}$, the more the values of $\widetilde{Y}_i$ are concentrated around its mean value m_i (the smaller the variance $\tilde{\sigma}^2_{G_i}$). From (4.1), we observe that the decision-maker puts a higher weight on the effort m for the high-noise candidates compared to the low-noise candidates: for the same level of effort, the high-noise candidates will be seen by the decision-maker as having less variability of expected quality compared to that of the low-noise candidates.

[2]Note that, for simplicity, we assume that the variance of the quality η^2 does not depend on the candidate's group; Our results can be extended to the case of group-dependent η^2_G (see Section 4.7).

[3]Our results can be extended to the case where the effort m is not observable (see Section 4.7).

4.3.2 The population game

As the decision-maker selects the candidates having the largest expected quality, it will select all candidates whose expected quality $\widetilde{Y}$ is larger than some selection threshold θ. We denote by $r_G(m;\theta)$ the probability for a candidate of group G to be selected when their effort is m and the selection threshold is θ. It equals:[4]

$$r_G(m;\theta) = \mathrm{P}\left(\widetilde{Y} \geq \theta\right) = \Phi\left(\frac{m-\theta}{\tilde{\sigma}_G}\right), \tag{4.2}$$

where Φ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

We assume that each selected candidate receives a positive reward S, whereas the candidates who are not selected get a reward of 0. Hence, the *payoff* u_G of a candidate from population G is

$$u_G(m;\theta) = S \cdot r_G(m;\theta) - C_G \cdot m^2/2. \tag{4.3}$$

Given a selection threshold θ, each candidate strategically decides on the effort m that maximizes their expected payoff $u_G(m;\theta)$. Following the classical definition (Sandholm, 2010), we call it a *pure best response* and we denote the set of all *pure best response* strategies of a candidate as

$$b_G(\theta) = \{m \text{ such that } \forall m' \geq 0 : u_G(m;\theta) \geq u_G(m';\theta)\}.$$

We say that the best response is unique if the set of best responses $b_G(\theta)$ is reduced to a singleton. In such a case, by abuse of notation, we denote by $b_G(\theta)$ the unique best response.

Similarly, by $B_G(\theta)$ we denote the set of all *mixed best responses* of a candidate. It is the set of all probability distributions over the set of pure best responses $b_G(\theta)$.

Selection threshold For each population $G \in \{H, L\}$, we denote by μ_G the distribution of efforts used by this population G. It is a probability distribution[5] on $\mathbb{R}^+$. We denote by $\boldsymbol{\mu} = (\mu_H, \mu_L)$ the effort distributions of the two populations. We denote the cumulative distribution function of the expected quality $\widetilde{Y}$ induced by μ_G by F_{μ_G}, and the cumulative distribution function of expected quality of the total population by $F_{\boldsymbol{\mu}} = p_H F_{\mu_H} + p_L F_{\mu_L}$. The decision-maker selects the α-best candidates. Hence, it will select all candidate whose expected quality $\widetilde{Y}$ is above the $(1-\alpha)$-quantile of the distribution $F_{\boldsymbol{\mu}}$, that we denote by $\theta(\boldsymbol{\mu}) = F_{\boldsymbol{\mu}}^{-1}(1-\alpha)$.

[4] In the rest of the paper, to simplify the notation, we will drop the index i.

[5] If all candidates of the population G make the same effort, we say that μ_G is a *pure strategy*. Otherwise, μ_G is a *mixed strategy*. Formally, μ_G is a pure strategy if there exists m_0, such that $\mu_G(m) = \delta(m - m_0)$ where δ is the Dirac delta function.

The game The above definitions describe a population game between the candidates. We denote the game by $\mathcal{G}^{\text{un}}$, where the superscript "un" emphasizes that the decision-maker is *un*constrained,[6] i.e., it selects the best α candidates based on the expected quality $\widetilde{Y}$. Note that the game $\mathcal{G}^{\text{un}}$ is parameterized by the reward size S, the variance of quality η^2, the noise variances σ_G^2, the cost coefficients C_G, and the selection size α.

4.3.3 Existence and uniqueness of the Nash equilibrium

We use the standard definition of Nash equilibrium for populations games (Sandholm, 2010):

Definition 4.3.1 (Nash equilibrium)**.** *A pair of effort distributions $\boldsymbol{\mu} = (\mu_H, \mu_L)$ is called an equilibrium of the game $\mathcal{G}^{\text{un}}$ if for all populations $G \in \{H, L\}$, the support of μ_G is included in the set of best responses $b_G(\theta(\boldsymbol{\mu}))$, where $\theta(\boldsymbol{\mu}) = F_{\boldsymbol{\mu}}^{-1}(1-\alpha)$.*

A Nash equilibrium is a situation where no candidate has an incentive to change its decision: if the population plays $\boldsymbol{\mu}$, then the selection threshold will be $\theta(\boldsymbol{\mu})$. Hence, a candidate of a group G does not have an incentive to play a strategy that is not in $b(\theta(\boldsymbol{\mu}))$. As the support of μ_G is included in $b(\theta(\boldsymbol{\mu}))$, this implies that all candidates cannot obtain a higher payoff by unilaterally changing their decision.

In the rest of the chapter, we will study the property of the Nash equilibrium of the game, which exists and is unique as guaranteed by the following theorem.

Theorem 4.3.2. *The game $\mathcal{G}^{\text{un}}$ has a unique Nash equilibrium, that we denote by $\boldsymbol{\mu}^{\text{un}} = (\mu_H^{\text{un}}, \mu_L^{\text{un}})$.*

We will also denote by $\theta^{\text{un}} = F_{\boldsymbol{\mu}^{\text{un}}}^{-1}(1-\alpha)$ the selection threshold at equilibrium. Note that the equilibrium and the selection threshold depend on the game parameters, including the reward S. In the following, we will study the properties of the equilibrium as S grows. For simplicity of exposition, we omit the dependency on S and write $\boldsymbol{\mu}^{\text{un}}$ or θ^{un} instead of $\boldsymbol{\mu}^{\text{un}}(S)$ or $\theta^{\text{un}}(S)$.

We provide a detailed proof of Theorem 4.3.2 in Section 4.8.4 whose main ingredient is to show that there is a one-to-one mapping between the equilibria of the game and the fixed points of a multi-valued function T:

$$T(\theta) = \{F_{\boldsymbol{\mu}}^{-1}(1-\alpha) : \mu_G \in B_G(\theta)\}. \tag{4.4}$$

This simplifies the problem, as now we need to study the fixed points of a function of a single variable. By studying the first derivative of T, we show that the function T has a unique fixed point and, hence, the defined game has also a unique equilibrium.

4.3.4 Summary of main notation

We summarize the main notation in Table 4.1. Recall that, without loss of generality, we assume that $\sigma_H^2 > \sigma_L^2$ (which also implies that $\tilde{\sigma}_H^2 < \tilde{\sigma}_L^2$). We, thus, refer to H-candidates as *high-noise*, and we refer to L-candidates as *low-noise*. To refer to the group

[6] In Section 4.5, we study a decision-maker faced with a fairness constraint (called demographic parity).

with higher/lower cost-of-effort, we use the names *cost-disadvantaged/cost-advantaged*. If, for example, $C_H > C_L$, then we say that H-candidates are *cost-disadvantaged* and that L-candidates are *cost-advantaged*.

Table 4.1 – Summary of notation.

$Y_i \sim \mathcal{N}(m_i, \eta^2)$	quality of candidate i
$X_i = Y_i + \varepsilon_i \sigma_{G_i}$	quality estimate of candidate i ($\varepsilon_i \sim \mathcal{N}(0,1)$)
$\widetilde{Y}_i = \mathrm{E}(Y_i \mid X_i)$	expected quality of candidate i
μ_G	distribution of effort for the population G
$\boldsymbol{\mu} = (\mu_H, \mu_L)$	distribution of effort for the total population
F_{μ_G}	CDF of the expected quality $\widetilde{Y}$ for the population G
$F_{\boldsymbol{\mu}} = p_H F_{\mu_H} + p_L F_{\mu_L}$	CDF of the expected quality $\widetilde{Y}$ for the total population
$\alpha \in (0,1)$	selection size
$\theta(\boldsymbol{\mu}) = F_{\boldsymbol{\mu}}^{-1}(1-\alpha)$	selection threshold
p_G	proportion of G-candidates in the total population
$r_G(m;\theta)$	selection probability for a G-candidate with effort m given θ

4.4 Equilibrium Characterization and Resulting Discrimination

In this section, we characterize the equilibrium of the game for large[7] rewards S. We consider the case of large S as it models the competition in selection procedures with high stakes, e.g., hiring CEOs or college admission to high-ranked schools.

4.4.1 Properties of the best response

We start with the characterization of the best response $b_G(\theta)$ of a candidate. In Lemma 4.4.1 (whose proof is deferred to Section 4.8.2), we show that, when S is large, the best response $b_G(\theta)$ is unique for all thresholds θ, except one, that we call a *dropout threshold* θ_G^{d}. We show that the value of the best response $b_G(\theta)$ increases with the threshold θ until the latter reaches θ_G^{d}; it then drops down and decreases when $\theta \geq \theta_G^{\mathrm{d}}$. This means that when the selection threshold θ is too high, the candidates lose incentives to make any effort. To emphasize the dependency of the dropout threshold on the reward S, we write $\theta_G^{\mathrm{d}}(S)$.

Lemma 4.4.1 (Best reponse for large S)**.** *There exists S_0 such that for all rewards $S \geq S_0$, there exists a threshold $\theta_G^{\mathrm{d}}(S)$, called the dropout threshold, such:*

(i) When the selection threshold is $\theta = \theta_G^{\mathrm{d}}(S)$, there are two pure best response strategies:

[7] In Section 4.6.2, we show theoretical and numerical results for small and intermediate rewards S.

$b_G(\theta_G^{\mathrm{d}}(S)) = \left\{ b_G^{\min}\left(\theta_G^{\mathrm{d}}(S)\right), b_G^{\max}\left(\theta_G^{\mathrm{d}}(S)\right)\right\}$. *They satisfy:*

$$\lim_{S\to\infty} b_G^{\min}\left(\theta_G^{\mathrm{d}}(S)\right) = 0, \qquad \lim_{S\to\infty} \frac{b_G^{\max}\left(\theta_G^{\mathrm{d}}(S)\right)}{\theta_G^{\mathrm{d}}(S)} = 1.$$

(ii) For all $\theta \neq \theta_G^{\mathrm{d}}(S)$, the pure best response $b_G(\theta)$ is unique. Moreover, for any $\gamma \in (0,1)$, we have

$$\lim_{S\to\infty} b_G\left(\theta_G^{\mathrm{d}}(S)/\gamma\right) = 0 \text{ and } \lim_{S\to\infty} \frac{b_G\left(\gamma\theta_G^{\mathrm{d}}(S)\right)}{\theta_G^{\mathrm{d}}(S)} = \gamma.$$

Using the notation for asymptotic equivalence, we can write the statement of the theorem in a simpler form. For example, from the second part of (ii), we can infer that $b_G(\gamma\theta_G^{\mathrm{d}}) \sim \gamma\theta_G^{\mathrm{d}}$ for any $\gamma \in (0,1)$. Therefore, the above lemma shows that for a given candidate, the best response $b_G(\theta)$ increases essentially linearly up to the dropout threshold θ_G^{d}, and then drops to 0 when the selection threshold is too high, i.e., $\theta > \theta_G^{\mathrm{d}}$. We will illustrate this later in Fig. 4.1.

Note that the dropout threshold θ_G^{d} is group-dependent. As we will see later in Section 4.4.3, the most represented group at equilibrium will be the one having the largest dropout threshold. Hence, in Lemma 4.4.2, we show the asymptotic behavior of the dropout threshold as a function of S. As we expect, the dropout threshold increases with the reward S, and decreases with the cost coefficient C_G. A less intuitive property is its relation with the noise variance σ_G^2: we show that the dropout threshold increases as the noise variance σ_G^2 increases. The proof of Lemma 4.4.2 can be found in Section 4.8.3.

Lemma 4.4.2 (Dropout threshold for large S)**.** *Let $\theta_G^{\mathrm{d}}(S)$ be the dropout threshold.*

(i) For any C_G and G, the dropout threshold can be expressed as

$$\theta_G^{\mathrm{d}}(S) = \sqrt{2S/C_G}\,(1 + o(1)) \text{ as } S \to \infty.$$

(ii) If $C_H = C_L$ and $\sigma_H^2 > \sigma_L^2$, then there exists S_0 such that for all $S \geq S_0$,

$$\theta_H^{\mathrm{d}}(S) > \theta_L^{\mathrm{d}}(S).$$

From (i), we conclude that the dependency of the dropout threshold θ^{d} on the cost coefficients C_G is of higher order than the dependency on the noise variance σ_G^2. In other words, when $C_H > C_L$, we have $\theta_H^{\mathrm{d}}(S) < \theta_L^{\mathrm{d}}(S)$ for large enough S, regardless of the noise variance σ_G^2. When both groups have equal cost coefficients ($C_H = C_L$), the noise variance matters, and $\theta_H^{\mathrm{d}}(S) > \theta_L^{\mathrm{d}}(S)$ when $\sigma_H^2 > \sigma_L^2$ (or, equivalently, when $\tilde{\sigma}_H^2 < \tilde{\sigma}_L^2$).

4.4.2 Equilibrium strategy

As we show in the previous section, the dropout threshold θ_G^{d} is an important characteristics of candidates' best response: for thresholds θ smaller than the dropout threshold θ_G^{d}, the candidates make an effort proportional to the threshold whereas for thresholds larger than the dropout, the candidates make nearly zero efforts. In Theorem 4.4.3 given below, we prove that, for large rewards, the selection threshold at equilibrium is equal to the dropout threshold of one of the two groups. Note that this theorem is not stated in terms of the groups H and L but in terms of groups G_1 and G_2, where G_1 is the group that has the largest dropout threshold, which holds if $C_1 < C_2$ or ($C_1 = C_2$ and $\sigma_1^2 > \sigma_2^2$). Hence, the population G_1 can correspond to the population L if $C_H > C_L$ (it corresponds, otherwise, to the population H as we assumed that $\sigma_H^2 > \sigma_L^2$).

Theorem 4.4.3 (Equilibrium strategies)**.** *Fix $\alpha \in (0,1)$ and two populations of candidates, G_1 and G_2, such that ($C_1 < C_2$) or ($C_1 = C_2$ and $\sigma_1^2 > \sigma_2^2$). Then, there exists a reward S_0 such that for $S \geq S_0$:*

(i) If $\alpha \leq p_1$, then the equilibrium threshold $\theta^{\mathrm{un}}(S)$ is $\theta_1^{\mathrm{d}}(S)$. In this case:

- *The G_1-candidates play a mixed strategy that consists in playing $b_1^{\max}$ with probability τ_1 and $b_1^{\min}$, with probability $1-\tau_1$, where $\lim_{S\to\infty} \tau_1 = \alpha/p_1$.*
- *The G_2-candidates play the pure strategy $b_2(\theta_1^{\mathrm{d}}(S))$ that satisfies*

$$\lim_{S\to\infty} b_2(\theta_1^{\mathrm{d}}(S)) = 0.$$

(ii) If $\alpha > p_1$, then the equilibrium threshold $\theta^{\mathrm{un}}(S)$ is $\theta_2^{\mathrm{d}}(S)$. In this case:

- *The G_1-candidates play the pure strategy $b_1(\theta_2^{\mathrm{d}}(S))$ that satisfies*

$$\lim_{S\to\infty} b_1(\theta_2^{\mathrm{d}}(S))/\theta_2^{\mathrm{d}}(S) = 1.$$

- *The G_2-candidates play a mixed strategy that consists in playing $b_2^{\max}$ with probability τ_2 and $b_2^{\min}$ with probability $1-\tau_2$, where $\lim_{S\to\infty} \tau_2 = (\alpha - p_1)/p_2$.*

Proof Sketch. To simplify notation, we omit the dependence on S for all variables.

Let us prove that the dropout threshold θ_1^{d} is the fixed point of T for the case (i). We specify the efforts made in response to θ_1^{d}, and we show that they lead to the same selection threshold θ_1^{d}, i.e., θ_1^{d} is the fixed point of T. The case (ii) can be proven similarly; we provide the complete proof in Section 4.8.5.

We can show that, as $S \to \infty$, we have $r_G^{\max} := r_G(b_G^{\max}(\theta_G^{\mathrm{d}}); \theta_G^{\mathrm{d}}) \xrightarrow{S\to\infty} 1$ and $r_G^{\min} := r_G(b_G^{\min}(\theta_G^{\mathrm{d}}); \theta_G^{\mathrm{d}}) \xrightarrow{S\to\infty} 0$ for all $G \in \{G_1, G_2\}$.

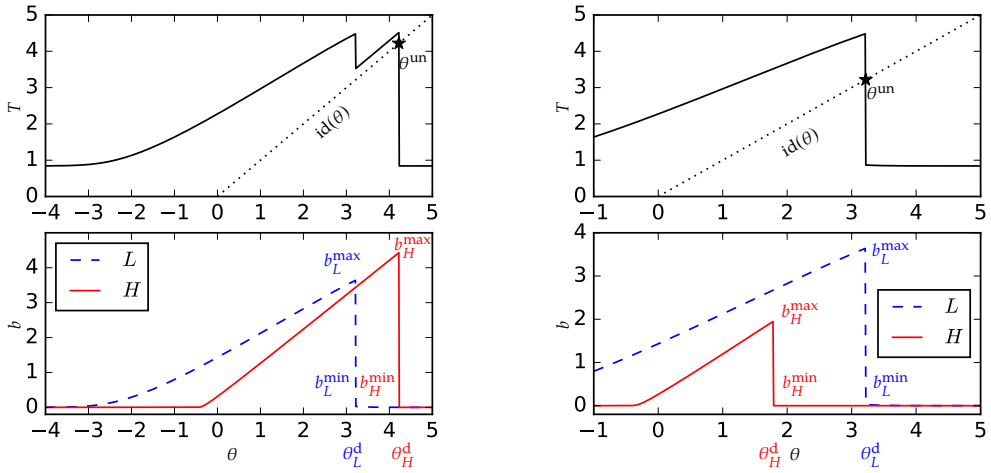
If $\alpha \leq p_1$, then, assume that G_1-candidates randomize their strategy by playing $b_1^{\max}$ with the probability τ_1 and $b_1^{\min}$ with the probability $1-\tau_1$. The G_2-candidates play the deterministic strategy $b_2(\theta_1^{\mathrm{d}}) \xrightarrow{S\to\infty} 0$. The probability τ_1 can be found from

the budget constraint: $\alpha = p_1(\tau_1 r_1^{\max} + (1-\tau_1) r_1^{\min}) + p_2 r_2(\theta_1^{\mathrm{d}}, b_2(\theta_1^{\mathrm{d}})) \xrightarrow{S\to\infty} p_1\tau_1$, so $\tau_1 \xrightarrow{S\to\infty} \alpha/p_1$.

The defined strategies satisfy the budget constraint, so θ_1^{d} is the fixed point of T and, hence, the defined distribution of effort is the equilibrium of the game $\mathcal{G}^{\mathrm{un}}$. □

Fig. 4.1 illustrates the results of Theorem 4.4.3. In Fig. 4.1a, we show the case of group-dependent noise but group-independent cost coefficient ($C_H = C_L$). In this case, the H-candidates have a higher dropout threshold compared to the L-candidates, hence, for $\theta_L^{\mathrm{d}} < \theta \le \theta_H^{\mathrm{d}}$, the H-candidates make a non-zero effort while the L-candidates make a nearly-zero effort. In our illustration, the selection size $\alpha = 0.1$ is smaller than the proportion of H-candidates $p_H = 0.5$. We can verify that for $\theta = \theta_H^{\mathrm{d}}$, if a proportion α/p_H of H-candidates plays $b_H^{\max}$, and the rest of H-candidates plays $b_H^{\min}$, then such a strategy satisfies the budget constraint. Hence, θ_H^{d} is the fixed point of the function T, so $\theta^{\mathrm{un}} = \theta_H^{\mathrm{d}}$.

In Fig. 4.1b, we illustrate the case of group-dependent cost coefficient ($C_H > C_L$). In this case, the L-candidates have a higher dropout threshold compared to the H-candidates, hence, for $\theta_H^{\mathrm{d}} < \theta \le \theta_L^{\mathrm{d}}$, the L-candidates make a non-zero effort while H-candidates make a nearly-zero effort. For the purpose of illustration, we again assume that the selection size $\alpha = 0.1$ is smaller than the proportion of L-candidates ($p_L = 0.5$). Similarly to the previous case, we can verify that for $\theta = \theta_L^{\mathrm{d}}$, if a proportion α/p_L of L-candidates plays $b_L^{\max}$, and the rest of L-candidates plays $b_L^{\min}$, then this strategy satisfies the budget constraint, so $\theta^{\mathrm{un}} = \theta_L^{\mathrm{d}}$.



(a) $C_H/C_L = 1$ (b) $C_H/C_L = 5$

Figure 4.1 – **Best response functions and the Nash equilibrium** for $S = 10$, $C_L = 1$, $\tilde{\sigma}_H = 0.1$ and $\tilde{\sigma}_L = 1$, $p_H = 0.5$ and $\alpha = 0.1$. Both figures illustrate the case (i) of Theorem 4.4.3. The dotted line on the uppermost panels is the identity function $\mathrm{id}(\theta) = \theta$.

4.4.3 Discrimination due to the group-dependent variance and cost

In Theorem 4.4.3, we show that the equilibrium distribution of effort $\boldsymbol{\mu}^{\text{un}}$ depends on the relation between the dropout threshold θ_G^{d} for different populations, H and L. We now study how the noise variance σ_G^2 and the cost coefficient C_G affect the representation of groups in the selection at equilibrium. Overall, we show that the group-dependent noise variance and the group-dependent cost coefficient can both lead to underrepresentation of some group of candidates at equilibrium.

For a given population G, we denote by $\bar{\mu}_G^{\text{un}}$ the *average effort* that candidates from group G exert at equilibrium, and we denote by $\bar{r}_G^{\text{un}}$ the *selection rate* which is the probability for a randomly chosen candidate from group G to be selected. Since $r_G(m;\theta)$ is the probability for a G-candidate to be selected when exerting an effort m (see (4.2)), these quantities satisfy:

$$\bar{\mu}_G^{\text{un}} = \int_{m\geq 0} m\, d\mu_G^{\text{un}}(m); \qquad \bar{r}_G^{\text{un}} = \int_{m\geq 0} r_G(m;\theta^{\text{un}})\, d\mu_G^{\text{un}}(m).$$

We say that the group H (or L) is *underrepresented* if $\bar{r}_H < \bar{r}_L$ (or $\bar{r}_L < \bar{r}_H$). Note that when we say "underrepresented', it means that a demographic group has less representation in the selection than its share in the population of all candidates. This definition is not conditioned on the assumption that the mean of the qualities Y is the same for both groups, but this assumption does make the notion of demographic parity more obviously appealing since there is no a priori distinction in the average quality between the groups. Nevertheless, we do not claim that demographic parity would be justified only under this assumption.

In Theorem 4.4.4 below, we show that if the cost coefficient C_G is group-independent ($C_H = C_L$), then the high-noise H-candidates make a larger effort compared to that of the low-noise L-candidates, and the latter are underrepresented. However, if the cost coefficient is group-dependent ($C_H \neq C_L$), we show that the noise variance does not play a role if the reward S is large enough: the cost-advantaged candidates make larger effort compared to that of the cost-disadvantaged candidates, and as a result, the latter are underrepresented. This theorem also shows that, as S goes to infinity, the ratios of efforts and selection rates can grow unbounded. For example, in the case (i) of Theorem 4.4.4, if there are enough H-candidates to fill the selection budget α, then L-candidates will asymptotically not be selected when S goes to infinity. We will see in Section 4.6.2, that for moderate values of reward S, the L-candidates still have some representation in the selection, but that can be very small.

Theorem 4.4.4 (Discrimination for large rewards S)**.** *Let μ_H^{un} and μ_L^{un}, r_H^{un} and r_L^{un} be the equilibrium effort distributions and selection rates of the game $\mathcal{G}^{\text{un}}$.*

(i) If $C_H < C_L$ or if $C_H = C_L$ and $\sigma_H^2 > \sigma_L^2$, then there exists S_0 such that for all $S \geq S_0$ the high-noise H-candidates make greater effort compared to the low-noise L-candidates, and the latter are underrepresented:

$$\lim_{S\to\infty} \frac{\bar{\mu}_L^{\text{un}}}{\bar{\mu}_H^{\text{un}}} = \lim_{S\to\infty} \frac{\bar{r}_L^{\text{un}}}{\bar{r}_H^{\text{un}}} = \begin{cases} 0 & \text{if } p_H \geq \alpha, \\ \frac{\alpha - p_H}{1-p_H} < 1 & \text{if } p_H < \alpha. \end{cases}$$

(ii) If $C_H > C_L$, then there exists S_0 such that for all $S \geq S_0$ the cost-advantaged L-candidates make a greater effort compared to the cost-disadvantaged H-candidates, and the latter are underrepresented:

$$\lim_{S\to\infty} \frac{\bar{\mu}_H^{\text{un}}}{\bar{\mu}_L^{\text{un}}} = \lim_{S\to\infty} \frac{\bar{r}_H^{\text{un}}}{\bar{r}_L^{\text{un}}} = \begin{cases} 0 & \text{if } p_L \geq \alpha, \\ \frac{\alpha - p_L}{1 - p_L} < 1 & \text{if } p_L < \alpha. \end{cases}$$

Proof. To prove (i), we compute the average effort and the selection rate given the equilibrium mixed strategies defined in Theorem 4.4.3. We consider the case of $\alpha \leq p_H$ and the case of $\alpha > p_H$ separately.

If $\alpha \leq p_H$, then $\bar{\mu}_H^{\text{un}} = b_H^{\max}\tau_H + b_H^{\min}(1-\tau_H) = \alpha/p_H \cdot \sqrt{2S/C_H}(1+o(1))$, $S \to \infty$ and $\bar{\mu}_L^{\text{un}} = b_L(\theta_H^{\text{d}}) \xrightarrow{S\to\infty} 0$, so $\bar{\mu}_L^{\text{un}}/\bar{\mu}_H^{\text{un}} \xrightarrow{S\to\infty} 0$. Similarly, the expected selection rates for H- and L-candidates:

$$\bar{r}_H^{\text{un}} = r_H^{\max}\tau_H + r_H^{\min}(1-\tau_H) \xrightarrow{S\to\infty} \alpha/p_H,$$
$$\bar{r}_L^{\text{un}} = r_L(\theta_H^{\text{d}}, b_L(\theta_H^{\text{d}})) \xrightarrow{S\to\infty} 0,$$

which implies that $\bar{r}_L^{\text{un}}/\bar{r}_H^{\text{un}} \xrightarrow{S\to\infty} 0$.

If $\alpha > p_H$, then we, again, consider the strategies in the equilibrium:

$$\bar{\mu}_H^{\text{un}} = b_H(\theta_L^{\text{d}}) = \sqrt{2S/C_L}(1+o(1)),\ S \to \infty,$$
$$\bar{\mu}_L^{\text{un}} = b_L^{\max}\tau_L + b_L^{\min}(1-\tau_L) = (\alpha - p_H)/p_L \cdot \sqrt{2S/C_L}(1+o(1)),\ S \to \infty.$$

Similarly, for the selection rates, we have:

$$\bar{r}_H^{\text{un}} = r_H(\theta_L^{\text{d}}, b_H(\theta_L^{\text{d}})) \xrightarrow{S\to\infty} 1,$$
$$\bar{r}_L^{\text{un}} = r_L^{\max}\tau_L + r_L^{\min}(1-\tau_L) \xrightarrow{S\to\infty} (\alpha - p_H)/p_L.$$

The proof of (ii) is identical to the proof of (i) so we omit it here. □

4.5 Effects of the Demographic Parity Mechanism on the Selection

In Theorem 4.4.4, we show that the equilibrium in the game $\mathcal{G}^{\text{un}}$ leads to underrepresentation of one of the groups of candidates: for the group-independent cost coefficient ($C_H = C_L$), the low-noise candidates are underrepresented; for the group-dependent cost coefficient ($C_H \neq C_L$), the cost-disadvantaged candidates are underrepresented.

To reduce the inequality of representation, colleges (and employers) sometimes perform *affirmative actions*. This can take the form of quotas for low-income or minority groups or any forms of promotions. Among those affirmative actions are ones that make sure that the selection rates for the groups are close, meaning that the proportions of both groups in the selection should be close to proportions of the groups in the total population. The strongest condition among those is the *demographic parity* (see (Barocas et al., 2019)) which requires that selection rates for both populations must be equal: $\bar{r}_H = \bar{r}_L$. This implies, given the budget constraint $\bar{r}_H p_H + \bar{r}_L p_L = \alpha$, that $\bar{r}_H = \bar{r}_L = \alpha$.

4.5.1 The game with the demographic parity mechanism

The demographic parity mechanism removes the competition among the groups of candidates, H and L, as from each group G a proportion $\alpha \in (0,1)$ must be accepted. Hence, the game with the demographic parity mechanism, that we denote by $\mathcal{G}^{\text{dp}}$, can be represented as two independent games, $\mathcal{G}_H^{\text{un}}$ and $\mathcal{G}_L^{\text{un}}$, each defined for a single group, H or L. Note that for this game, the selection thresholds will be different for the two groups: for a group G, the decision-maker will select all candidates whose expected quality is greater or equal than $\theta(\mu_G) = F_{\mu_G}^{-1}(1-\alpha)$, which corresponds to the α best fraction of this group—it is the quantile of the distribution of expected qualities induced by μ_G and not the one induced by $\boldsymbol{\mu}$ as in the unconstrained case.

The equilibrium of the game $\mathcal{G}^{\text{dp}}$ where the Bayesian decision-maker has a demographic parity constraint is defined as follows.

Definition 4.5.1 (Nash equilibrium of the game $\mathcal{G}^{\text{dp}}$)**.** *A pair of effort distributions $\boldsymbol{\mu} = (\mu_H, \mu_L)$ is an equilibrium of the game with the demographic parity constraint $\mathcal{G}^{\text{dp}}$ if for both groups $G \in \{H, L\}$:*

$$\mu_G \in B_G(\theta(\mu_G)),$$

where $\theta(\mu_G) = F_{\mu_G}^{-1}(1-\alpha)$.

Mimicking the unconstrained case, we denote by $\boldsymbol{\mu}^{\text{dp}} = (\mu_H^{\text{dp}}, \mu_L^{\text{dp}})$ the equilibrium of this game, and by $\theta_G^{\text{dp}} = F_{\mu_G^{\text{dp}}}^{-1}(1-\alpha)$ the group-dependent equilibrium selection threshold. The superscript "dp" emphasizes that the decision-maker is *demographic parity*-constrained.

Since the game $\mathcal{G}^{\text{dp}}$ can be represented as two separate games, $\mathcal{G}_H^{\text{un}}$ and $\mathcal{G}_L^{\text{un}}$, that has a unique Nash equilibrium according to Theorem 4.3.2, the game $\mathcal{G}^{\text{dp}}$ also has a unique Nash equilibrium.

4.5.2 Efforts induced by the demographic parity mechanism

In the previous section, we showed that, for large rewards S, it is possible that only one of the two groups makes a positive effort while the other group considers that the game is not worth playing because of a too unfair competition. The situation is radically different with demographic parity mechanism as each candidate competes with similar candidates. As we show below, the demographic parity mechanism pushes the previously underrepresented group to make more effort than before. Moreover, it can also push the overrepresented group to make more effort.

In the first theorem below, we characterize the equilibrium strategy of the game with the demographic parity mechanism $\mathcal{G}^{\text{dp}}$. This result is a direct corollary of Theorem 4.4.3 as we consider two separated games with an unconstrained decision-maker.

Theorem 4.5.2 (Equilibrium strategy with the demographic parity mechanism)**.** *There exists S_0 such that for all $S \geq S_0$, the equilibrium of the game with the demographic parity*

mechanism $\mathcal{G}^{\text{dp}}$ *is a pair of distributions* $\boldsymbol{\mu}^{\text{dp}} = (\mu_H^{\text{dp}}, \mu_L^{\text{dp}})$ *where for each group* $G \in \{H, L\}$, μ_G^{dp} *consists in playing* $b_G^{\max}$ *with probability* τ_G *and* $b_G^{\min}$ *with probability* $1 - \tau_G$, *where* $\lim_{S\to\infty} \tau_G = \alpha$.

Proof. According to the demographic parity mechanism, the selection rate per each group G must be equal to $\bar{r}_G = \alpha$. Hence, this theorem can be seen as a special case of Theorem 4.4.3 but with a single group of mass 1 out of which we need to select the best $\alpha \in (0, 1)$. By applying directly the result of Theorem 4.4.3, we show that the proposed strategy is the equilibrium strategy. □

In Corollary 4.5.3 given below, we compare the efforts made by two groups at equilibrium. We show that, if the cost coefficient is group-independent, then the demographic parity mechanism equalizes the effort as S grows (together with the selection rates as $\bar{r}_H^{\text{dp}} = \bar{r}_L^{\text{dp}}$ by definition). For group-dependent cost coefficient, the cost-disadvantaged H-candidates make lower average effort compared to that of L-candidates yet the average effort ratio is bounded by $\sqrt{C_H/C_L}$. This is in contrast to Section 4.4.3 where we show that the average effort ratio can be unbounded in the case of the unconstrained decision-maker.

Corollary 4.5.3 (Equilibrium effort ratio in $\mathcal{G}^{\text{dp}}$)**.** *Let* μ_H^{dp} *and* μ_L^{dp} *be the equilibrium effort distributions in the game with the demographic parity mechanism* $\mathcal{G}^{\text{dp}}$. *The average efforts of both populations* $G \in \{H, L\}$ *satisfy:*

$$\lim_{S\to\infty} \frac{\bar{\mu}_H^{\text{dp}}}{\bar{\mu}_L^{\text{dp}}} = \sqrt{\frac{C_L}{C_H}}.$$

In particular, if $C_H = C_L$, *then the average efforts of both populations grow at equal rates. If* $C_H \neq C_L$, *then the cost-disadvantaged candidates make a lower average effort compared to that of the cost-advantage candidates.*

Proof. Using the equilibrium strategies of the game with the demographic parity mechanism $\mathcal{G}^{\text{dp}}$ from Theorem 4.5.2, we show that:

$$\lim_{S\to\infty} \frac{\bar{\mu}_H^{\text{dp}}}{\bar{\mu}_L^{\text{dp}}} = \lim_{S\to\infty} \frac{\tau_H b_H^{\max} + (1-\tau_H) b_H^{\min}}{\tau_L b_L^{\max} + (1-\tau_L) b_L^{\min}} = \sqrt{\frac{C_L}{C_H}}.$$

□

By reducing the competition between groups, affirmative action policies are often criticized because they might encourage individuals to make less effort, which reduces the overall quality of the selected candidates. We show below that, in fact, the demographic parity mechanism always encourages the previously underrepresented group to make a larger effort than in the unconstrained case. For the previously overrepresented group, the situation varies: in many situations, the candidates from the overrepresented group will make a lower effort than in the unconstrained case, but when $\sqrt{C_L/C_H} < \alpha$ and $\alpha > p_L$, we show that they make a higher effort compared

to that of the unconstrained case. Note that the last result may seem rather counterintuitive as demographic parity reduces the competition between groups. Later, in Section 4.5.3, we show the implications of this result on the average quality of the selected candidates.

Corollary 4.5.4 (Equilibrium effort ratio in $\mathcal{G}^{\text{un}}$ vs. $\mathcal{G}^{\text{dp}}$)**.** *Let μ_H^{un} and μ_L^{un} be the equilibrium effort distributions in the unconstrained game $\mathcal{G}^{\text{un}}$. Similarly, let μ_H^{dp} and μ_L^{dp} be the equilibrium effort distributions in the game with the demographic parity constraint $\mathcal{G}^{\text{dp}}$.*

(i) If $C_H = C_L$ and $\sigma_H^2 > \sigma_L^2$, then

$$\lim_{S\to\infty} \frac{\bar{\mu}_H^{\text{un}}}{\bar{\mu}_H^{\text{dp}}} = \begin{cases} 1/p_H & \text{if } \alpha \le p_H, \\ 1/\alpha & \text{if } \alpha > p_H, \end{cases} \text{ and } \lim_{S\to\infty} \frac{\bar{\mu}_L^{\text{un}}}{\bar{\mu}_L^{\text{dp}}} = \begin{cases} 0 & \text{if } \alpha \le p_H, \\ \frac{\alpha - p_H}{\alpha - \alpha p_H} & \text{if } \alpha > p_H. \end{cases}$$

(ii) If $C_H > C_L$, then

$$\lim_{S\to\infty} \frac{\bar{\mu}_H^{\text{un}}}{\bar{\mu}_H^{\text{dp}}} = \begin{cases} 0 & \text{if } \alpha \le p_L, \\ \frac{\alpha - p_L}{\alpha - \alpha p_L} & \text{if } \alpha > p_L, \end{cases} \text{ and } \lim_{S\to\infty} \frac{\bar{\mu}_L^{\text{un}}}{\bar{\mu}_L^{\text{dp}}} = \begin{cases} 1/p_L & \text{if } \alpha \le p_L, \\ \sqrt{\frac{C_L}{C_H}}\frac{1}{\alpha} & \text{if } \alpha > p_L. \end{cases}$$

Proof Sketch. Similarly, as in the proof of Corollary 4.5.3, we calculate the limits by using the equilibrium strategies found in Theorem 4.4.3 and in Theorem 4.5.2. The complete proof is given in Section 4.8.6. □

4.5.3 Selection quality with and without the demographic parity mechanism

We now show the implication of the previous result on how the demographic parity mechanism affects the selection quality. In a non-strategic setting (see e.g., (Emelianov et al., 2022b)), the unconstrained decision maker is optimal in expectation. We show here that this no longer holds in the strategic setting: there exist scenarios under which *the Bayesian decision-maker is not optimal* in terms of the expected quality of selection. In such settings, the demographic parity constraint leads to a more qualified cohort. This counterintuitive phenomenon is due to the fact that demographic parity induces less competition between groups but a more fair competition within each group (compared to the unconstrained case).

We denote by $\mathcal{U}^{\text{un}}$ and by $\mathcal{U}^{\text{dp}}$ the expected quality at equilibrium of the selected candidates for the unconstrained and the demographic parity constrained games. In the theorem below, we characterize the ratio of the equilibrium cohort qualities by the unconstrained decision-maker from Section 4.4, and the demographic parity constrained decision-maker that we study in this section. For group-independent cost coefficient ($C_H = C_L$), we show that the quality ratio $\mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}}$ tends to 1 in the limit of large S. For group-dependent cost coefficient ($C_H \neq C_L$), we show that the quality ratio can be smaller than one—the Bayesian decision-maker is not optimal if the candidates are strategic, and the demographic parity mechanism can lead to a better-qualified cohort.

Theorem 4.5.5 (Selection quality ratio for $\mathcal{G}^{\text{un}}$ and $\mathcal{G}^{\text{dp}}$)**.** *Let $\mathcal{U}^{\text{un}}$ and $\mathcal{U}^{\text{dp}}$ be the expected qualities of selection at equilibrium of the game $\mathcal{G}^{\text{un}}$ and of the game $\mathcal{G}^{\text{dp}}$, respectively.*

(i) If $C_H = C_L$, then the ratio of the expected quality given by the unconstrained and the demographic parity constrained decision-makers tends to one with S:

$$\lim_{S\to\infty} \frac{\mathcal{U}^{\text{un}}}{\mathcal{U}^{\text{dp}}} = 1.$$

(ii) If $C_H > C_L$, then the ratio of the expected quality given by the unconstrained and the demographic parity constrained decision-makers can be smaller than one. Formally, for $c = \sqrt{C_L/C_H}$:

$$\lim_{S\to\infty} \frac{\mathcal{U}^{\text{un}}}{\mathcal{U}^{\text{dp}}} = \begin{cases} \frac{1}{cp_H+p_L} > 1 & \text{if } p_L \geq \alpha, \\ \frac{c}{cp_H+p_L} < 1 & \text{if } p_L < \alpha. \end{cases}$$

Proof Sketch. First, we prove that as $S \to \infty$, the expected quality of the selected cohort grows at equal rate with the expected effort: $\mathrm{E}(Y_G \cdot [\widetilde{Y}_G \geq \theta]) \sim \bar{\mu}_G$. Hence, using the equilibrium strategies from Theorem 4.4.3 and Theorem 4.5.2, we can estimate the ratio $\mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}}$ in the limit of $S \to \infty$. The complete proof is given in Section 4.8.7. □

We emphasize that the condition under which the demographic parity mechanism improves the average selection equality is when the selection rate α is larger than the size of the cost-advantaged group. The improvement of selection quality due to the demographic parity mechanism is explained by the fact that, without the demographic parity constraint, the advantaged minority has no incentives to make a large effort because the competition includes a lot of cost-disadvantaged candidates. Once the competition is among candidates of each separate populations, the cost-advantaged candidates have to compete with other cost-advantaged candidates, so they have to make a larger effort to be selected.

The demographic parity can decrease the average selection quality when $\alpha \leq p_L$ and when both groups have different cost coefficients C_G. In this case, if the low-noise L-candidates are the majority, then the ratio of quality $\mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}}$ cannot be larger than 2 as S goes to infinity, regardless of the cost coefficients.

4.6 Complementary Results

In the previous sections, we studied the properties of the Nash equilibrium of the game for large rewards S and showed that they can lead to discriminations. In this section, we complement this theoretical analysis by studying two (essentially independent) problems: First, does the population converge to the equilibrium? Second, what happens when the reward S is not large?

4.6.1 Convergence to the Nash equilibrium

To answer the first question, we perform a series of numerical experiments in which the decisions are made repeatedly. At a given time t, the candidates consider past data to make a strategic decision. This could represent, for instance, the case of college admission where candidates consider the distribution of grades from previous years; in this example, each decision epoch is a different year.

We study two population dynamics: *best response* and *fictitious play*.

- For the *best response* dynamics, at each of the discrete times $t = 1, 2, \dots, T$, the candidates observe the strategy played at the previous time step $\boldsymbol{\mu}^{(t-1)}$ and play a best response to it:

$$m_G^{(t)} \in b_G(\theta(\boldsymbol{\mu}^{(t-1)})).$$

- For the *fictitious play* dynamics, at each of the discrete times $t = 1, 2, \dots, T$, the candidates observe the whole history of plays and assume that the distribution of efforts is the empirical distribution of effort from time 1 to T. Candidates then play a best response to it:

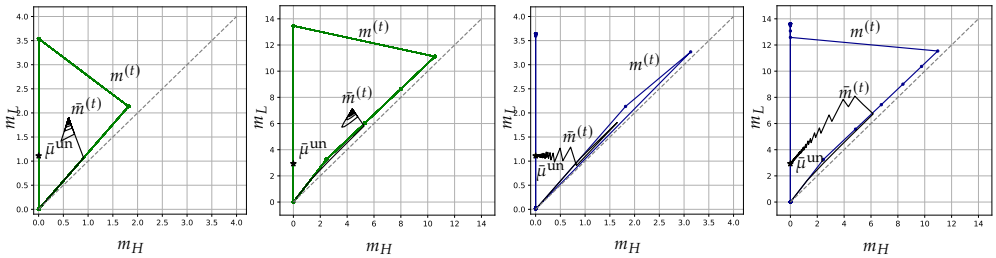
$$m_G^{(t)} \in b_G(\theta(\hat{\boldsymbol{\mu}}^{(t)})),$$

where $\theta(\hat{\boldsymbol{\mu}}^{(t)}) = F^{-1}_{\hat{\mu}^{(t)}}(1-\alpha)$ and $\hat{\boldsymbol{\mu}}^{(t)} = \sum_{s=1}^{t-1} \frac{1}{t-1} \boldsymbol{\mu}^{(s)}$.

We numerically evaluate these two policies and report the results in Fig. 4.2. For the best response dynamics, we observe that $\boldsymbol{m}^{(t)} = (m_H^{(t)}, m_L^{(t)})$ converges to a limit cycle for any starting point. This is because when S is large,[8] the best response map is not continuous (recall Fig. 4.1). The period of the limit cycle increases with the reward size S but the behavior is similar for all S: starting from $(0, 0)$, the candidates from both populations increase the effort as time increases. Then, the competition becomes too high and one of the populations drops out, i.e., make almost zero effort. After, the competition is only among the candidates of a single population until it becomes too difficult and all candidates drop out and return to the initial state. The cycle ends here, and the new cycle starts. In Fig. 4.2a and 4.2b, we also plot the average trajectory $\bar{\boldsymbol{m}}^{(t)} = (\bar{m}_H^{(t)}, \bar{m}_L^{(t)})$, where $\bar{m}_G^{(t)} = \frac{1}{t-1} \sum_{s=1}^{t-1} m_G^{(s)}$. We observe that the average over the trajectory seems to converge, yet the average effort over the trajectory is significantly larger than that of the average equilibrium effort for both groups.

The case of fictitious play dynamics is different, and it is depicted in Fig. 4.2c and 4.2d. In this case, the empirical distribution of efforts *does* converge to the Nash equilibrium. This is illustrated on the figure by the fact that the empirical average of effort converges to the average value of effort of the Nash equilibrium: $\bar{m}_G^{(t)} \xrightarrow{t\to\infty} \bar{\mu}_G^{\text{un}}$ for both groups $G \in \{H, L\}$. Note that this is not a pointwise convergence but rather a convergence to a cycle: at equilibrium, the strategy played by the L-candidates converges to a cycle on the values $b_L^{\min}$ and $b_L^{\max}$.

[8] For $S < \frac{1}{2} C_G \tilde{\sigma}_G^2 / \phi(1)$ we can show that the function T is a contraction mapping, so any trajectory of the best response dynamics converges to the Nash equilibrium.



(a) BR, $S = 10$ (b) BR, $S = 100$ (c) FP, $S = 10$ (d) FP, $S = 100$

Figure 4.2 – **Best response (BR) and fictitious play (FP) dynamics** for different rewards S. The parameters of simulations are $T = 500$, $C_H = 1.5$, $C_L = 1$, $p_H = p_L = 0.5$, $\alpha = 0.1$, $\tilde{\sigma}_H = 0.6$ and $\tilde{\sigma}_L = 1$.

4.6.2 Selection problems with small and intermediate rewards

We start with the case of small rewards $S < C_G\tilde{\sigma}_G^2/\phi(1)$ for which the results are quite different from the ones obtained for large S in Section 4.4. We show that the cost ratio C_L/C_H, as well as the expected quality variance ratio $\tilde{\sigma}_L^2/\tilde{\sigma}_H^2$, play significant roles in the outcome of the game. If the cost for the H-candidates is too high compared to that of the L-candidates, then the H-candidates make lower effort for all selection sizes α. Otherwise, if the cost for H-candidates is comparable to that of L-candidates (e.g., $C_H = C_L$) or it is lower, then for both small and large selection sizes α, the high-noise H-candidates make less effort compared to that of low-noise L-candidates. *For both cases, there exists a parameter dependent value of α_0 such that for all values of $\alpha \leq \alpha_0$, the high-noise H-candidates are underrepresented.* Interestingly, in (ii) of Theorem 4.6.1, we also observe that it is possible that H-candidates make a larger effort than L-candidates, yet they are selected at a lower rate, $\bar{r}_H^{\text{un}} < \bar{r}_L^{\text{un}}$.

Overall, the high-level interpretation of the below result is that *for small enough rewards S and for small enough values of α, the high-noise H-candidates are always underrepresented*. This result is in contrast with the results for large S studied in Section 4.4, and it is similar to the result in the non-strategic setting studied in Emelianov et al., 2022b; Garg et al., 2021.

Theorem 4.6.1 (Discrimination for small rewards S)**.** *Assume that $S < C_G\tilde{\sigma}_G^2/\phi(1)$ for all $G \in \{H, L\}$. Let μ_H^{un} and μ_L^{un}, r_H^{un} and r_L^{un} be the equilibrium efforts and selection rates of the game $\mathcal{G}^{\text{un}}$. Denote $K_\mu := \sqrt{\frac{-2\ln(C_H\tilde{\sigma}_H/(C_L\tilde{\sigma}_L))}{1/\tilde{\sigma}_H^2 - 1/\tilde{\sigma}_L^2}}$ and $K_r := \sqrt{\mathcal{W}\left(\frac{S^2\left(\frac{1}{C_H\tilde{\sigma}_H} - \frac{1}{C_L\tilde{\sigma}_L}\right)^2}{2\pi(\tilde{\sigma}_L - \tilde{\sigma}_H)^2}\right)}$, where $\mathcal{W}$ is the Lambert function defined as the inverse to the function $f(\lambda) = \lambda e^\lambda$ and Φ^c is the complementary cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$.*

(i) If $C_H\tilde{\sigma}_H > C_L\tilde{\sigma}_L$, then $\bar{\mu}_H^{\text{un}} < \bar{\mu}_L^{\text{un}}$ for all $\alpha \in (0, 1)$, and $\bar{r}_H^{\text{un}} < \bar{r}_L^{\text{un}}$ if and only if $\alpha < \Phi^c(-K_x)$.

(ii) If $C_H\tilde{\sigma}_H \leq C_L\tilde{\sigma}_L$, *then*

$$\bar{\mu}_H^{\text{un}} < \bar{\mu}_L^{\text{un}} \iff \alpha \in \left(0, \sum_{G\in\{A,B\}} p_G\Phi^c\left(K_\mu/\tilde{\sigma}_G\right)\right) \cup \left(\sum_{G\in\{A,B\}} p_G\Phi^c\left(-K_\mu/\tilde{\sigma}_G\right), 1\right),$$
$$\bar{r}_H^{\text{un}} < \bar{r}_L^{\text{un}} \iff \alpha < \Phi^c\left(K_x\right).$$

Proof Sketch. We show that for $S < C_G\tilde{\sigma}_G^2/\phi(1)$ the best response in pure strategies is unique, hence the equilbirium effort distribution is a singleton $\mu_G^{\text{un}} = \delta(m - m_G^{\text{un}})$. Note that the first-order condition on a maximum of the payoff function u_G is also a sufficient condition; it can be written:

$$\frac{S}{\tilde{\sigma}_G}\phi\left(\frac{m_G^{\text{un}} - \theta^{\text{un}}}{\tilde{\sigma}_G}\right) - C_G m_G^{\text{un}} = 0 \iff m_G^{\text{un}} = \frac{S}{C_G\tilde{\sigma}_G}\phi\left(\frac{m_G^{\text{un}} - \theta^{\text{un}}}{\tilde{\sigma}_G}\right).$$

Since we aim to find a value of α when $m_H^{\text{un}} = m_L^{\text{un}}$, we equate the right-hand sides of the above equation for two groups, and, by solving this equation, we obtain the values of $\theta^{\text{un}} - m_G^{\text{un}}$. By substituting this expression to the budget constraint, we derive the value of α at which $m_H^{\text{un}} = m_L^{\text{un}}$:

$$\alpha = \sum_{G\in\{H,L\}} p_G\Phi^c\left(\frac{\theta^{\text{un}} - m_G^{\text{un}}}{\tilde{\sigma}_G}\right).$$

The proof for r_G^{un} is similar to that of m_G^{un}. A complete proof is given in Section 4.8.8. □

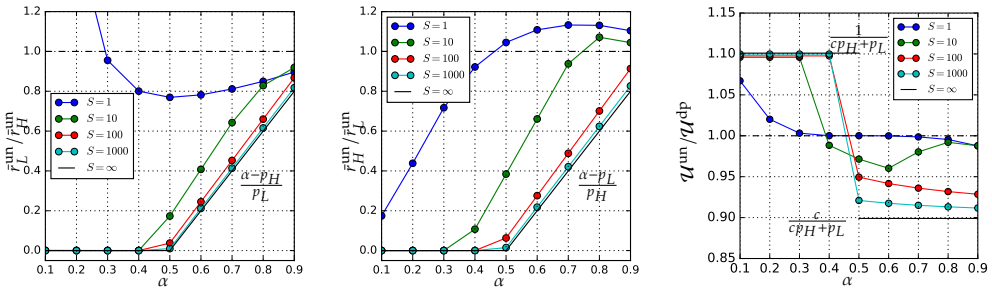
Intermediate rewards To conclude our analysis, we fill the gap between our theoretical results for the cases of small and large rewards S using numerical simulations.[9] We perform our numerical simulations for the values of reward $S = 1, 10, 100, 1000$. The simulation result for $S = 1$ is studied theoretically in the first part of this section, as $S = 1$ satisfies the condition $S < C_G\tilde{\sigma}_G^2/\phi(1)$. The result for $S = \infty$, studied in Section 4.4 and Section 4.5, is represented in Fig. 4.3 using a black solid line.

We plot the ratio of $\bar{r}_L^{\text{un}}/\bar{r}_H^{\text{un}}$ for the case of group-independent cost coefficient in Fig. 4.3a, the ratios of $\bar{r}_H^{\text{un}}/\bar{r}_L^{\text{un}}$ and $\mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}}$ for the case of group-dependent cost coefficient in Fig. 4.3b and Fig. 4.3c. Overall, we observe a relatively smooth transition between the two regimes of S in all figures. In addition, the behavior for $S = 100$ and for $S = 1000$ is quite close to the behavior for $S = \infty$.

4.7 Conclusion and Discussion

In this chapter, we propose a simple model of selection with strategic candidates who are faced with group-dependent cost-of-effort and group-dependent noise variance. We characterize the resulting discrimination at equilibrium as well as the impact of removing it through the demographic parity mechanism that mandates equal representation across groups. Note that, in the context of our strategic model, demographic

[9] The code can be found at https://gitlab.inria.fr/vemelian/strategic-selection-code



(a) $\bar{r}_L^{\text{un}}/\bar{r}_H^{\text{un}}$ ($C_H/C_L = 1$) (b) $\bar{r}_H^{\text{un}}/\bar{r}_L^{\text{un}}$ ($C_H/C_L = 1.5$) (c) $\mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}}$ ($C_H/C_L = 1.5$)

Figure 4.3 – Characterization of the equilibrium for different rewards S. The parameters of simulations are $C_L = 1$, $p_L = 0.5$, $\tilde{\sigma}_H = 0.6$ and $\tilde{\sigma}_L = 1$.

parity is not the only fairness notion that might make sense and one could be tempted to consider, for instance, a meritocratic notion of fairness (where the representation would be commensurate with the effort). It remains important, however, to understand the impact of imposing demographic parity as it is one of the most commonly used fairness notions.

Throughout the chapter, we made several simplifying assumptions, often to make the results easier to state and understand and to better isolate the effect of strategic behavior. Our work can be extended, however, in multiple ways:

Group-dependent variance of quality We assumed that the variance of latent qualities is group-independent, i.e., $\eta_H^2 = \eta_L^2 = \eta^2$. This assumption can easily be removed, as all the results can equivalently be stated for $\tilde{\sigma}_H^2$ and $\tilde{\sigma}_L^2$ even with group-dependent variance. We make this assumption only for simplicity of exposition since it implies that $\tilde{\sigma}_H^2 < \tilde{\sigma}_L^2$ if and only if $\sigma_H^2 > \sigma_L^2$.

Unobservable effort In our model, we assume that the effort m is observable to the decision-maker. If the effort m is not observable, we argue that the decision-maker performs the selection based on the noisy estimate X rather than the posterior expectation $\widetilde{Y}$ (which corresponds to the group-oblivious decision-maker in the terminology of Chapter 3). All the results from this chapter still hold but we need to replace the variance of the expected qualities $\tilde{\sigma}_G^2$ by the variance of the estimate $\eta^2 + \sigma_G^2$. Note that most of the statements will be reversed as $\tilde{\sigma}_H^2 < \tilde{\sigma}_L^2$ if and only if $\sigma_H^2 > \sigma_L^2$. In this case, the high-noise group has a higher variance of estimate, and the low-noise group has a lower variance of estimate. Hence, when the low-noise group is underrepresented in our model, the low-noise group is overrepresented in the model with unobservable effort.

More than two groups of candidates. In our model, we assume two groups of candidates, yet the results can be extended to more than two groups (e.g., in the proof of uniqueness of the equilibrium we do not rely on the fact that the number of

populations is two). This additional dimension would simply add more interactions between the groups and complicate the statement of the results. For instance, in the case of multiple groups where two of them are subject to equal cost coefficients, we expect that the sorting will be with respect to the noise for these two groups, and with respect to the cost coefficients for the rest of the groups.

Monomial cost function and non-Gaussian noise. In our model, we assume a quadratic cost and Gaussian noise. We believe that these results will not change much if we assume other symmetric noise and monomial cost functions, i.e., $C_G m^{d_G}$. In this case, we expect that the best response could be characterized by a dropout threshold, as in our model. In addition, the power d_G of the monomial would be another important feature of the model: if it is group-dependent, then we expect that the dropout threshold will grow as $(S/C_G)^{1/d_G}$ and the candidates with higher d_G will drop out earlier, independently on the relations of C_G and $\tilde{\sigma}_G^2$.

Other models of candidate's utility In our model, we assume rational risk-neutral candidates faced with different costs of efforts, and we also assume a risk-neutral Bayesian decision-maker. In non-strategic settings (see Chapter 3), such a decision-maker is proven to be Bayesian-optimal, yet we prove that it is not optimal in the strategic setting. In our model, we do not consider other barriers for candidates as, for example, self-selection. To model self-selection, we can assume that there is a minimal threshold θ_G^{self} that each candidate should pass. The outcome of the equilibrium would then depend on the relation between θ_G^{self} and the dropout threshold θ_G^{d}. We can also easily consider a risk-averse (or risk loving) decision maker: in the case of an exponential utility function, this will lead to an additive bias β_G proportional to the variance $\tilde{\sigma}_G^2$, i.e., $\widetilde{Y}_i \sim \mathcal{N}(m_{G_i} + \beta_{G_i}, \tilde{\sigma}_{G_i}^2)$.

4.8 Omitted Proofs

4.8.1 Properties of the best response

Recall that the payoff function of an individual with effort m and given the selection threshold θ is

$$u(m;\theta) = S \cdot \Phi\left(\frac{m-\theta}{\tilde{\sigma}}\right) - \frac{1}{2}Cm^2,$$

where Φ is the CDF of the standard normal distribution.

Denoting ϕ the PDF of the standard normal distribution, the first two derivatives of u with respect to m are:

$$\frac{\partial u}{\partial m}(m;\theta) = \frac{S}{\tilde{\sigma}}\phi\left(\frac{\theta-m}{\tilde{\sigma}}\right) - Cm,$$
$$\frac{\partial^2 u}{(\partial m)^2}(m;\theta) = \frac{S}{\tilde{\sigma}}\phi\left(\frac{\theta-m}{\tilde{\sigma}}\right)\frac{\theta-m}{\tilde{\sigma}^2} - C.$$

The payoff function u is defined on $[0, \infty)$, and it is continuous and continuously differentiable. Moreover, $\frac{\partial u}{\partial m}(m = 0) > 0$ and $u(m; \theta) \xrightarrow{m\to\infty} -\infty$. Hence, all local maxima of u must satisfy the first-order condition (FOC) $\frac{\partial u}{\partial m}(m; \theta) = 0$ and the second-order condition (SOC) $\frac{\partial^2 u}{\partial m^2} \leq 0$. The maximum of the payoff function is attained in one of the local maxima.

We start by a first lemma.

Lemma 4.8.1 (Maxima of u)**.** *Fix S, C and $\tilde{\sigma}$.*

(i) If $S < C\tilde{\sigma}^2/\phi(1)$, then there is a unique global maximum of $u(m; \theta)$ for all θ.

(ii) If $S \geq C\tilde{\sigma}^2/\phi(1)$, then there exists a unique $\theta^{\mathrm{d}}(S)$ such that for $\theta = \theta^{\mathrm{d}}(S)$ there are two global maxima of $u(m; \theta)$, and for $\theta \neq \theta^{\mathrm{d}}$, there is a unique global maximum of $u(m; \theta)$.

Proof. Let us denote by $z = (m - \theta)/\tilde{\sigma}$ and let $v(z) = \frac{S}{\tilde{\sigma}}\phi(z) - C\tilde{\sigma}z$ and $w(z) := -\frac{S}{\tilde{\sigma}}z\phi(z) - C\tilde{\sigma} = dv(z)/dz$. The first and second derivatives of u can be expressed as a function of v and w:

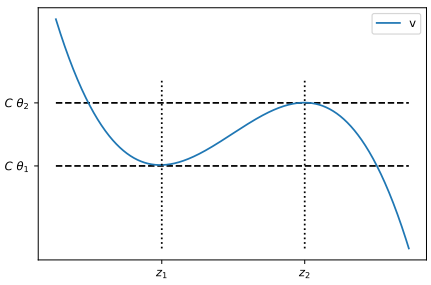
$$\frac{\partial u}{\partial m}(m; \theta) = v(z) - C\theta,$$
$$\frac{\partial^2 u}{(\partial m)^2}(m; \theta) = \frac{1}{\tilde{\sigma}}w(z).$$

The function $z \cdot \phi(z)$ has the global maximum at $z = 1$ which is equal to $\phi(1)$, and the global minimum at $z = -1$ which is equal to $-\phi(1)$. Hence, two cases are possible:

(i) If $\phi(1) < C\tilde{\sigma}^2/S$, then $w(z) < 0$ for all $z \neq 1$. Hence, v is a strictly decreasing function (since $w < 0$ in this case), so the FOC gives a unique solution m which is a global maximum of u.

(ii) If $\phi(1) \geq C\tilde{\sigma}^2/S$, then the equation $w(z) = 0$ has two real solutions, denoted by $z_1 \leq z_2$. They can be found explicitly, i.e., $z_{1,2} = -\sqrt{-\mathcal{W}_{1,2}\left(-\frac{2\pi C^2\tilde{\sigma}^4}{S^2}\right)}$ where $\mathcal{W}$ is the Lambert function defined as the inverse to the function $f(y) = ye^y$. Note also that $z_1 \xrightarrow{S\to\infty} -\infty$ and $z_2 \xrightarrow{S\to\infty} 0$.

We consider the latter case (ii) in details. We can verify, that the function v is a decreasing function for $z \in (-\infty, z_1) \cup (z_2, \infty)$, and it is an increasing function for $z \in (z_1, z_2)$. This shows that the function $v(z)$ has the same shape as the curve on the right. As a result, the FOC condition $v(z) - C\theta = 0$ can have at most three solutions depending on the value of θ.
Indeed, let $\theta_1 = v(z_1)/C$ and $\theta_2 = v(z_2)/C$.



- For all $\theta \notin (\theta_1, \theta_2)$, the FOC gives a unique solution which is a global maximum of $u(m;\theta)$.
- For all $\theta \in (\theta_1, \theta_2)$, we have that $v(z_1) \leq C\theta$ and $v(z_2) \geq C\theta$ which guarantees three solutions to FOC which we denote by $m_1 \leq m_2 \leq m_3$.
- For $\theta = \theta_1$,we have $u(m_1(\theta);\theta) = u(m_2(\theta);\theta)$, and for $\theta = \theta_2$ we have that $u(m_2(\theta);\theta) = u(m_3(\theta);\theta)$. For all $\theta \in (\theta_1, \theta_2)$, the m_2 is a local minimum, m_1 and m_3 are local maxima.

Consider the two continuous, differentiable and non-negative functions $\Delta_{12}(\theta) = u(m_1(\theta);\theta) - u(m_2(\theta);\theta)$ and $\Delta_{32}(\theta) = u(m_3(\theta);\theta) - u(m_2(\theta);\theta)$. We can see that $\Delta_{12}(\theta_1) = 0$ and $\Delta_{32}(\theta_2) = 0$.

We can also verify that $\frac{\partial u(m(\theta),\theta)}{\partial \theta} = Cm\left(\frac{\partial m}{\partial \theta} - 1\right) - Cm\frac{\partial m}{\partial \theta} = -Cm(\theta) < 0$. Therefore, the function Δ_{12} is increasing , since $\frac{d\Delta_{12}}{d\theta} = C(m_2(\theta) - m_1(\theta)) > 0$, whereas the function Δ_{32} is decreasing since $\frac{d\Delta_{32}}{d\theta} = C(m_2(\theta) - m_3(\theta)) < 0$. Hence, there is a unique $\theta^{\mathrm{d}} \in (\theta_1, \theta_2)$, such that $\Delta_{12}(\theta^{\mathrm{d}}) = \Delta_{23}(\theta^{\mathrm{d}})$ which is equivalent to $u(m_1(\theta^{\mathrm{d}}), \theta^{\mathrm{d}}) = u(m_3(\theta^{\mathrm{d}}), \theta^{\mathrm{d}})$.

□

4.8.2 Proof of Lemma 4.4.1

We start with the proof of the case (i). According to Lemma 4.8.1, there are two pure best response values, $b^{\max}$ and $b^{\min}$, that correspond to the dropout threshold θ^{d}. Following the definition of the expected payoff $u(m;\theta)$, we can show the following upper bound on the values of pure best responses $b(\theta^{\mathrm{d}}) \in \{b^{\max}(\theta^{\mathrm{d}}), b^{\min}(\theta^{\mathrm{d}})\}$:

$$0 < u(b(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = S \cdot r(b(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) - \frac{Cb^2(\theta^{\mathrm{d}})}{2} \leq S - \frac{Cb^2(\theta^{\mathrm{d}})}{2}.$$

This implies that $b(\theta^{\mathrm{d}}) \leq \sqrt{2S/C}$.

Second, we show that $|\theta^{\mathrm{d}} - b(\theta^{\mathrm{d}})|$ is not bounded as $S \to \infty$. By assuming that $|\theta^{\mathrm{d}} - b(\theta^{\mathrm{d}})| < \varepsilon$ for all S, we end up with the following contradictory inequality that must hold for any value of S:

$$\phi\left(\frac{\varepsilon}{\tilde{\sigma}}\right) < \phi\left(\frac{\theta^{\mathrm{d}} - b(\theta^{\mathrm{d}})}{\tilde{\sigma}}\right) = \frac{C\tilde{\sigma}}{S} \cdot b(\theta^{\mathrm{d}}) \leq \tilde{\sigma}\sqrt{2C/S} \xrightarrow{S\to\infty} 0,$$

where ϕ is the PDF of the standard normal distribution $\mathcal{N}(0,1)$.

Since the value of $|\theta^{\mathrm{d}} - b(\theta^{\mathrm{d}})|$ is not bounded, by studying the first and the second derivatives of the payoff function (as in Lemma 4.8.1), we can show that $(b^{\max}(\theta^{\mathrm{d}}) - \theta^{\mathrm{d}}) \xrightarrow{S\to\infty} +\infty$ and $(b^{\min}(\theta^{\mathrm{d}}) - \theta^{\mathrm{d}}) \xrightarrow{S\to\infty} -\infty$. Hence, the selection rates, corresponding to the $b^{\max}$ and $b^{\min}$ converge to:

$$\lim_{S\to\infty} r(b^{\max}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = \lim_{S\to\infty} \Phi\left(\frac{b^{\max}(\theta^{\mathrm{d}}) - \theta^{\mathrm{d}}}{\tilde{\sigma}}\right) = 1,$$

$$\lim_{S\to\infty} r(b^{\min}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = \lim_{S\to\infty} \Phi\left(\frac{b^{\min}(\theta^{\mathrm{d}})-\theta^{\mathrm{d}}}{\tilde{\sigma}}\right) = 0,$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$.

Asymptotic behavior of $b^{\min}$ ***and*** $b^{\max}$**.** Using the definition of the dropout threshold θ^{d} and the definition for $b^{\min}$, $b^{\max}$, we can write:

$$0 < u(b^{\max}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = u(b^{\min}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) \iff$$
$$0 < r(b^{\max}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) - \frac{C(b^{\max}(\theta^{\mathrm{d}}))^2}{2S} = r(b^{\min}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) - \frac{C(b^{\min}(\theta^{\mathrm{d}}))^2}{2S} < r(b^{\min}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}).$$

As $\lim_{S\to\infty} r(b^{\min}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = 0$ and $\lim_{S\to\infty} r(b^{\max}(\theta^{\mathrm{d}});\theta^{\mathrm{d}}) = 1$, it implies that

$$\lim_{S\to\infty} \frac{C(b^{\max}(\theta^{\mathrm{d}}))^2}{2S} = 1 \text{ and } \lim_{S\to\infty} \frac{C(b^{\min}(\theta^{\mathrm{d}}))^2}{2S} = 0.$$

Hence, we show the asymptotic behavior of the pure strategy best response at θ^{d} for $S \to \infty$:

$$b^{\max}(\theta^{\mathrm{d}}(S)) = \sqrt{2S/C}(1+o(1)) \text{ and } b^{\min}(\theta^{\mathrm{d}}(S)) = o(\sqrt{2S/C}).$$

Asymptotic behavior of θ^{d}**.** Finally, we consider the asymptotic behavior of the dropout threshold θ^{d}. The dropout threshold $\theta^{\mathrm{d}}(S)$ is unbounded: if we assume the opposite, then $b^{\max}$ would be $o(1)$, $S \to \infty$ since $b^{\max}$ must satisfy the FOC:

$$b^{\max}(\theta^{\mathrm{d}}) = \frac{S}{C\tilde{\sigma}}\phi\left(\frac{\theta^{\mathrm{d}}-b^{\max}(\theta^{\mathrm{d}})}{\tilde{\sigma}}\right).$$

Hence, consider the following limit which we find using l'Hôpital rule and the properties of $b^{\min}$ and $b^{\max}$ proven above:

$$\lim_{S\to\infty} \frac{\theta^{\mathrm{d}}(S)}{\sqrt{2S/C}} = \lim_{S\to\infty} \frac{\frac{b^{\max}(\theta^{\mathrm{d}}(S))+b^{\min}(\theta^{\mathrm{d}}(S))}{2S}}{\frac{1}{2}S^{-1/2}\sqrt{2/C}} = \lim_{S\to\infty} \frac{b^{\max}(\theta^{\mathrm{d}}(S))+b^{\min}(\theta^{\mathrm{d}}(S))}{\sqrt{2S/C}} = 1.$$

Refined asymptotic behavior of $b^{\min}$**.** Since $b^{\min}(\theta^{\mathrm{d}}(S)) = o(\sqrt{2S/C})$ and $\theta^{\mathrm{d}}(S) = \sqrt{2S/C}(1+o(1))$ as $S \to \infty$, then, using the first-order condition, we show

$$b^{\min}(\theta^{\mathrm{d}}(S)) = \frac{S}{C\tilde{\sigma}}\phi\left(\frac{\theta^{\mathrm{d}}(S)-b^{\min}(\theta^{\mathrm{d}}(S))}{\tilde{\sigma}}\right) = \frac{S}{C\tilde{\sigma}}\phi\left(\frac{\sqrt{2S/C}(1+o(1))}{\tilde{\sigma}}\right) \xrightarrow{S\to\infty} 0.$$

Now, we are ready to proof the case (ii). According to Lemma 4.8.1, the pure best response is unique. Consider two cases:

- If $\theta(S) = \gamma\theta^{\mathrm{d}}(S)$, then $\lim_{S\to\infty}\left(b(\gamma\theta^{\mathrm{d}}_G(S)) - \gamma\theta^{\mathrm{d}}_G(S)\right) = +\infty$. Consider the following limit that we calculate using l'Hôpital rule:

$$\lim_{S\to\infty} \frac{b(\gamma\theta^{\mathrm{d}}(S))}{\gamma\theta^{\mathrm{d}}(S)} = \lim_{S\to\infty} \frac{db/d\theta \cdot d\theta/dS}{d\theta/dS} = \lim_{S\to\infty} \frac{db}{d\theta} = \lim_{S\to\infty} \frac{b(b-\theta)}{b(b-\theta)+\tilde{\sigma}^2} = 1.$$

- If $\theta(S) = \theta^{\mathrm{d}}(S)/\gamma$, then $\lim_{S\to\infty}\left(b(\theta^{\mathrm{d}}(S)/\gamma) - \theta^{\mathrm{d}}(S)/\gamma\right) = -\infty$. Since the payoff function u is non-negative, we have:

$$u \geq 0 \iff Cb(\theta^{\mathrm{d}}(S)/\gamma)^2/(2S) \leq \Phi\left(\frac{b(\theta^{\mathrm{d}}(S)/\gamma) - \theta^{\mathrm{d}}(S)/\gamma}{\tilde{\sigma}}\right) \xrightarrow{S\to\infty} 0.$$

  Hence, $b(\theta^{\mathrm{d}}(S)/\gamma) = o(\sqrt{2S/C})$. Given that the best response must satisfy the FOC:

$$b(\theta^{\mathrm{d}}(S)/\gamma) = \frac{S}{C\tilde{\sigma}}\phi\left(\frac{\theta^{\mathrm{d}}(S)/\gamma - b(\theta^{\mathrm{d}}(S)/\gamma)}{\tilde{\sigma}}\right) = \frac{S}{C\tilde{\sigma}}\phi\left(\frac{\sqrt{2S/C}\frac{1}{\gamma}(1+o(1))}{\tilde{\sigma}}\right) = o(1).$$

4.8.3 Proof of Lemma 4.4.2

Case (i). The asymptotic behavior of the dropout threshold is given in Section 4.8.2. We recall the proof here. Consider the following limit which we find using l'Hôpital rule and the properties of $b^{\min}$ and $b^{\max}$ proven in Lemma 4.4.1:

$$\lim_{S\to\infty}\frac{\theta^{\mathrm{d}}(S)}{\sqrt{2S/C}} = \lim_{S\to\infty}\frac{\frac{b^{\max}(\theta^{\mathrm{d}}(S))+b^{\min}(\theta^{\mathrm{d}}(S))}{2S}}{\frac{1}{2}S^{-1/2}\sqrt{2/C}} = \lim_{S\to\infty}\frac{b^{\max}(\theta^{\mathrm{d}}(S)) + b^{\min}(\theta^{\mathrm{d}}(S))}{\sqrt{2S/C}} = 1.$$

Case (ii). To prove that the dropout threshold $\theta^{\mathrm{d}}(S;\tilde{\sigma})$ is decreasing with $\tilde{\sigma}$, we differentiate with respect to $\tilde{\sigma}$ the condition on the equal payoffs for $b^{\min}$ and $b^{\max}$ strategies, and we obtain:

$$\frac{\partial\theta^{\mathrm{d}}}{\partial\tilde{\sigma}} = -\frac{b^{\max}(\theta^{\mathrm{d}}(S)) + b^{\min}(\theta^{\mathrm{d}}(S)) - \theta^{\mathrm{d}}(S)}{\tilde{\sigma}} < 0 \text{ for large enough } S.$$

4.8.4 Proof of Theorem 4.3.2

Let us denote by T the set-valued function

$$T(\theta) = \{F^{-1}_{\boldsymbol{\mu}}(1-\alpha) : \mu_G \in B_G(\theta)\},$$

where $\boldsymbol{\mu} = (\mu_H, \mu_L)$ and $F_{\boldsymbol{\mu}}$ is the CDF of $\widetilde{Y}$ induced by $\boldsymbol{\mu}$: $F_{\boldsymbol{\mu}} = p_H F_{\mu_H} + p_L F_{\mu_L}$.

Lemma 4.8.2. *There is a one-to-one correspondence between the fixed points of T and the equilibria $\boldsymbol{\mu}^{\mathrm{un}}$ of the game $\mathcal{G}^{\mathrm{un}}$.*

Proof. If $\boldsymbol{\mu}^{\mathrm{un}}$ is an equilibrium of the game $\mathcal{G}^{\mathrm{un}}$, then there is a unique θ^{un} such that $F^{-1}_{\boldsymbol{\mu}^{\mathrm{un}}}(1-\alpha) = \theta^{\mathrm{un}}$ since $F_{\boldsymbol{\mu}^{\mathrm{un}}}(\theta)$ is a monotone function. This θ^{un} is a fixed point of T since $\mu^{\mathrm{un}}_G \in B_G(\theta^{\mathrm{un}})$ by definition of $\boldsymbol{\mu}^{\mathrm{un}}$.

On the other hand, if θ^{un} is a fixed point of T, then there is a unique $\boldsymbol{\mu}^{\mathrm{un}} = (\mu^{\mathrm{un}}_H, \mu^{\mathrm{un}}_L)$ such that $\mu^{\mathrm{un}}_G \in B_G(\theta^{\mathrm{un}})$:

1. If $\theta \neq \theta^{\mathrm{d}}_G$ for all $G \in \{H, L\}$, then the pure best responses b_H and b_L are unique (see Lemma 4.8.1 in Section 4.8.1), so $\boldsymbol{\mu}^{\mathrm{un}}$ is unique.

2. If, without loss of generality, $\theta = \theta_H^{\mathrm{d}}$, then there is a unique pure best response b_L, and two pure response values $b_H^{\max}$, $b_H^{\min}$ (see Lemma 4.8.1 in Section 4.8.1). The function $F(\tau) = F_{\boldsymbol{\mu}}(\theta)$ is monotone in $\tau \in [0,1]$ which is a probability for $\mu_H = \tau\delta(m - b_H^{\max}) + (1-\tau)\delta(m - b_H^{\min})$. It is also an equilibrium of $\mathcal{G}^{\mathrm{un}}$ as $\theta = F_{\boldsymbol{\mu}}^{-1}(1-\alpha)$ by definition of θ. Hence, $\boldsymbol{\mu}^{\mathrm{un}} = (\mu_H, \mu_L)$ is an equilibrium of the game $\mathcal{G}^{\mathrm{un}}$ and it is unique.

□

By showing a one-to-one correspondence between equilibria of $\mathcal{G}^{\mathrm{un}}$ and the fixed points of T, it is left us to prove that the function $T(\theta)$ has a unique fixed point θ^{un}, i.e., a solution to $T(\theta^{\mathrm{un}}) = \theta^{\mathrm{un}}$ is unique. This will imply that the equilibrium of the game $\mathcal{G}^{\mathrm{un}}$ is unique. We use the following lemmas.

Lemma 4.8.3. *Assume that $m(\theta)$ satisfies FOC and SOC defined in Section 4.8.1.*

(i) If $m > \theta$, then $0 < \frac{dm}{d\theta} < 1$.

(ii) If $m \leq \theta$, then $\frac{dm}{d\theta} \leq 0$.

Proof. First, we derive the expression for the first derivative. We differentiate the both sides of FOC defined in Section 4.8.1:

$$\frac{dm}{d\theta} = \frac{S}{C\tilde{\sigma}}\left(\frac{m-\theta}{\tilde{\sigma}}\right)\phi\left(\frac{\theta-m}{\tilde{\sigma}}\right)\left(\frac{1-\frac{dm}{d\theta}}{\tilde{\sigma}}\right) = \frac{m-\theta}{\tilde{\sigma}^2}m\left(1-\frac{dm}{d\theta}\right) \implies \frac{dm}{d\theta} = \frac{m(m-\theta)}{m(m-\theta)+\tilde{\sigma}^2}.$$

Second, from the SOC defined in Section 4.8.1, we find that $m(m-\theta) + \tilde{\sigma}^2 > 0$. Hence, the sign of the derivative $dm/d\theta$ is determined only be the sign of its numerator $m(m-\theta)$. Note that the value of m is strictly positive as it satisfies the FOC, so $dm/d\theta > 0$ if and only if $m > \theta$. □

Lemma 4.8.4. *For all S and $\theta \neq \theta_G^{\mathrm{d}}(S)$, the total derivative $dT/d\theta$ can be expressed as:*

$$\frac{dT}{d\theta} = \frac{\sum_G p_G \frac{1}{\tilde{\sigma}_G}\phi\left(\frac{T-b_G}{\tilde{\sigma}_G}\right)\cdot\frac{b_G(b_G-\theta)}{b_G(b_G-\theta)+\tilde{\sigma}_G^2}}{\sum_G p_G \frac{1}{\tilde{\sigma}_G}\phi\left(\frac{T-b_G}{\tilde{\sigma}_G}\right)} < 1.$$

Proof. Since T is an implicit function of θ, we use the chain rule to find the total derivative:

$$\frac{dT}{d\theta} = \sum_G \frac{\partial T}{\partial b_G}\frac{db_G}{d\theta}.$$

By differentiating both sides of the budget constraint, we can find the partial derivative $\partial T/\partial m_G$:

$$\frac{\partial}{\partial b_G}\left(\sum_G p_G r_G(b_G;T)\right) = 0 \iff \frac{\partial T}{\partial b_G} = \frac{p_G\frac{1}{\tilde{\sigma}_G}\phi\left(\frac{T-b_G}{\tilde{\sigma}_G}\right)}{\sum_G p_G\frac{1}{\tilde{\sigma}_G}\phi\left(\frac{T-b_G}{\tilde{\sigma}_G}\right)}.$$

Finally, using the fact that $\partial T/\partial b_G > 0$, where $\sum_G \partial T/\partial b_G = 1$ and Lemma 4.8.3, we show that $dT/d\theta < 1$.

□

Existence First, let us show that there is an interval $[\theta_0, \theta_1]$, such that for all $\theta \in \mathbb{R}$, we have that $\theta_0 \leq T(\theta) \leq \theta_1$:

- Since the best response $b_G(\theta) \geq 0$ for all $G \in \{H, L\}$, then for any fixed α, let θ_0 be the solution to the equation $\sum_G p_G \Phi^c \left(\frac{\theta_0 - 0}{\tilde{\sigma}_G}\right) = \alpha$. Hence, $T(\theta) \geq \theta_0$ for all $\theta \in \mathbb{R}$.
- Since the best response $b_G(\theta) \leq \sqrt{2S/C_G}$ for all $G \in \{H, L\}$, then for any fixed α, let θ_1 be the solution to the equation $\sum_G p_G \Phi^c \left(\frac{\theta_1 - \sqrt{2S/C_G}}{\tilde{\sigma}_G}\right) = \alpha$. Hence, $T(\theta) \leq \theta_1$ for all $\theta \in \mathbb{R}$.

Therefore, we can consider the function T on the interval $[\theta_0, \theta_1]$ which is compact and convex. The graph of the function T is closed and for all $\theta \in [\theta_0, \theta_1]$, we have that $T(\theta)$ is convex (for $\theta \neq \theta_G^{\mathrm{d}}$, the value of $T(\theta)$ is unique, and for $\theta = \theta_G^{\mathrm{d}}$, the value of $T(\theta)$ is an interval). Hence, using Kakutani fixed point theorem, we show that the fixed point exists.

Uniqueness We show that there exists a fixed point of the function T. For all $\theta \neq \theta_G^{\mathrm{d}}$, we have that $dT/d\theta < 1$ (Lemma 4.8.4), and we have that $\lim_{\theta \to \theta_G^{\mathrm{d}}+} T(\theta) \leq \lim_{\theta \to \theta_G^{\mathrm{d}}-} T(\theta)$ for all $G \in \{H, L\}$. Hence, there is a unique solution to the fixed-point problem $T(\theta) = \theta$, and, as a result, a unique equilibrium of the game $\mathcal{G}^{\mathrm{un}}$ due to Lemma 4.8.2.

4.8.5 Proof of Theorem 4.4.3

Case (i)

We prove that the dropout threshold $\theta_1^{\mathrm{d}}(S)$ is the fixed point of T, i.e., it corresponds to the equilibrium of the game $\mathcal{G}^{\mathrm{un}}$. As we show in the proof of Lemma 4.4.1, as $S \to \infty$, we have $r_G^{\max} := r(b^{\max}(\theta_G^{\mathrm{d}}); \theta_G^{\mathrm{d}}) \xrightarrow{S \to \infty} 1$ and $r_G^{\min} := r(b^{\min}(\theta_G^{\mathrm{d}}); \theta_G^{\mathrm{d}}) \xrightarrow{S \to \infty} 0$.

If $\alpha \leq p_1$, and given that the selection rate for the G_2-candidates at θ_1^{d} tends to zero with S, assume that G_1-candidates randomize their strategy by playing $b_1^{\max}$ with the probability τ_1 and $b_1^{\min}$ with the probability $1 - \tau_1$. The G_2-candidates play $b_2(\theta_1^{\mathrm{d}}) \xrightarrow{S \to \infty} 0$. The probability τ_1 can be found from the budget constraint: $\alpha = p_1(\tau_1 r_1^{\max} + (1 - \tau_1) r_1^{\min}) + p_1 r_2(\theta_1^{\mathrm{d}}, b_2(\theta_1^{\mathrm{d}})) \xrightarrow{S \to \infty} p_1 \tau_1$, so $\tau_1 \xrightarrow{S \to \infty} \alpha / p_1$. This strategy satisfies the budget constraint, so θ_1^{d} is the fixed point of T.

Case (ii)

We now prove that the dropout threshold θ_2^{d} is fixed point of T. If $\alpha > p_1$, then selecting all candidates from G_1 group would not be enough, and some G_2-candidates are needed to satisfy the selection rate α.

Given θ_2^{d}, the fraction of selected G_1-candidates would be equal to $r_1(\theta_2^{\mathrm{d}}, b_1(\theta_2^{\mathrm{d}})) \xrightarrow{S\to\infty} 1$. We claim the G_2-candidates will play $b_2^{\max}$ with probability τ_2, and $b_2^{\min}$ with probability $1-\tau_2$. The probability τ_2 can be found from the budget constraint:

$$\alpha = p_1 r_1(\theta_2^{\mathrm{d}}, b_1(\theta_2^{\mathrm{d}})) + p_2(r_2^{\max}\tau_2 + r_2^{\min}(1-\tau_2)) \xrightarrow{S\to\infty} p_1 + p_2\tau_2.$$

Hence, for $\tau_2 \xrightarrow{S\to\infty} \frac{\alpha-p_1}{p_2}$, the dropout threshold, θ_2^{d} is the fixed point of T.

4.8.6 Proof of Corollary 4.5.4

We use the expressions for equilibrium strategies from Theorem 4.4.3 and Theorem 4.5.2.

(i) If $C_H = C_L = C$ and $\sigma_H^2 > \sigma_L^2$:

$$\lim_{S\to\infty}\frac{\bar\mu_H^{\mathrm{un}}}{\bar\mu_H^{\mathrm{dp}}} = \lim_{S\to\infty}\begin{cases}\frac{\alpha/p_H\sqrt{2S/C}(1+o(1))}{\alpha\sqrt{2S/C}(1+o(1))} & \text{if } \alpha\le p_H\\ \frac{\sqrt{2S/C}(1+o(1))}{\alpha\sqrt{2S/C}(1+o(1))} & \text{if } \alpha > p_H\end{cases} = \begin{cases}1/p_H & \text{if } \alpha\le p_H,\\ 1/\alpha & \text{if } \alpha > p_H,\end{cases}$$

$$\lim_{S\to\infty}\frac{\bar\mu_L^{\mathrm{un}}}{\bar\mu_L^{\mathrm{dp}}} = \lim_{S\to\infty}\begin{cases}\frac{o(1)}{\alpha\sqrt{2S/C}(1+o(1))} & \text{if } \alpha\le p_H\\ \frac{\frac{\alpha-p_H}{p_L}\sqrt{2S/C}(1+o(1))}{\alpha\sqrt{2S/C}(1+o(1))} & \text{if } \alpha > p_H\end{cases} = \begin{cases}0 & \text{if } \alpha\le p_H,\\ \frac{\alpha-p_H}{\alpha-\alpha p_H} & \text{if } \alpha > p_H.\end{cases}$$

(ii) If $C_H > C_L$:

$$\lim_{S\to\infty}\frac{\bar\mu_H^{\mathrm{un}}}{\bar\mu_H^{\mathrm{dp}}} = \lim_{S\to\infty}\begin{cases}\frac{o(1)}{\alpha\sqrt{2S/C_H}(1+o(1))} & \text{if } \alpha\le p_L\\ \frac{(\alpha-p_L)/p_H\sqrt{2S/C_H}(1+o(1))}{\alpha\sqrt{2S/C_H}(1+o(1))} & \text{if } \alpha > p_L\end{cases} = \begin{cases}0 & \text{if } \alpha\le p_L,\\ \frac{\alpha-p_L}{\alpha-\alpha p_L} & \text{if } \alpha > p_L,\end{cases}$$

$$\lim_{S\to\infty}\frac{\bar\mu_L^{\mathrm{un}}}{\bar\mu_L^{\mathrm{dp}}} = \lim_{S\to\infty}\begin{cases}\frac{\alpha/p_L\sqrt{2S/C_L}(1+o(1))}{\alpha\sqrt{2S/C_L}(1+o(1))} & \text{if } \alpha\le p_L\\ \frac{\sqrt{2S/C_H}(1+o(1))}{\alpha\sqrt{2S/C_L}(1+o(1))} & \text{if } \alpha > p_L\end{cases} = \begin{cases}1/p_L & \text{if } \alpha\le p_L,\\ \sqrt{\frac{C_L}{C_H}}\frac{1}{\alpha} & \text{if } \alpha > p_L.\end{cases}$$

4.8.7 Proof of Theorem 4.5.5

First, using the formula for the expected value of the truncated normal distribution, we find:

$$\begin{aligned}\mathcal{U}^{\mathrm{un}}(m_G;\theta) &= \sum_G p_G\,\mathrm{E}(Y_G\cdot[\widetilde{Y}_G > \theta_G]) = \sum_G p_G\,\mathrm{E}(Y_G|\widetilde{Y}_G > \theta_G)\,\mathrm{P}(\widetilde{Y}_G \ge \theta_G)\\ &= \sum_G p_G\left[m_G\Phi\left(\frac{\theta_G - m_G}{\tilde\sigma_G}\right) + \tilde\sigma_G\phi\left(\frac{\theta_G - m_G}{\tilde\sigma_G}\right)\right].\end{aligned}$$

Second, using Lemma 4.4.1, we can show that for all $\gamma \in (0,1)$ and $\theta_G = \gamma \cdot \theta_G^{\mathrm{d}}$, we have:

$$b_G(\theta_G) \cdot \Phi\left(\frac{\theta_G - b_G(\theta_G)}{\tilde{\sigma}_G}\right) + \tilde{\sigma}_G \phi\left(\frac{\theta_G - b_G(\theta_G)}{\tilde{\sigma}_G}\right) = \theta_G(1+o(1))(1+o(1)) + \tilde{\sigma}_G \cdot o(1) = \theta_G(1+o(1))$$

and, for $\theta_G = \theta_G^{\mathrm{d}}/\gamma$, we have:

$$b_G(\theta_G) \cdot \Phi\left(\frac{\theta_G - b_G(\theta_G)}{\tilde{\sigma}_G}\right) + \tilde{\sigma}_G \phi\left(\frac{\theta_G - b_G(\theta_G)}{\tilde{\sigma}_G}\right) \xrightarrow{S\to\infty} 0.$$

For $\theta_G = \theta_G^{\mathrm{d}}$, we have:

$$b_G^{\max}(\theta_G) \cdot \Phi\left(\frac{\theta_G - b_G^{\max}(\theta_G)}{\tilde{\sigma}_G}\right) + \tilde{\sigma}_G \phi\left(\frac{\theta_G - b_G^{\max}(\theta_G)}{\tilde{\sigma}_G}\right) = \theta_G(1+o(1)),$$
$$b_G^{\min}(\theta_G) \cdot \Phi\left(\frac{\theta_G - b_G^{\min}(\theta_G)}{\tilde{\sigma}_G}\right) + \tilde{\sigma}_G \phi\left(\frac{\theta_G - b_G^{\min}(\theta_G)}{\tilde{\sigma}_G}\right) \xrightarrow{S\to\infty} 0.$$

For the game $\mathcal{G}^{\mathrm{dp}}$, we use the equilibrium strategy from Theorem 4.5.2, and can write that:

$$\begin{aligned}\mathcal{U}^{\mathrm{dp}} &= p_H(\alpha + o(1)) \cdot \theta_H^{\mathrm{d}}(S)(1+o(1)) + p_L \cdot (\alpha + o(1))\, \theta_L^{\mathrm{d}}(S)(1+o(1)) \\ &= \alpha p_H \cdot \theta_H^{\mathrm{d}}(S)(1+o(1)) + \alpha p_L \cdot \theta_L^{\mathrm{d}}(S)(1+o(1)), S \to \infty.\end{aligned}$$

For the game $\mathcal{G}^{\mathrm{un}}$, we use the equilibrium strategy from Theorem 4.4.3. We consider separately the case of group-independent and the case of group-dependent costs.

Group-independent cost

If $\alpha \le p_H$, then

$$\mathcal{U}^{\mathrm{un}} = p_H(\alpha/p_H + o(1)) \cdot \theta_H^{\mathrm{d}}(1+o(1)) + p_L \cdot o(1) = \alpha \cdot \sqrt{2S/C}(1+o(1)), S \to \infty.$$

If $\alpha > p_H$, then

$$\mathcal{U}^{\mathrm{un}} = p_H(1+o(1)) \cdot \theta_L^{\mathrm{d}}(1+o(1)) + p_L \cdot \left(\frac{\alpha - p_H}{p_L} + o(1)\right) \theta_L^{\mathrm{d}}(1+o(1)) = \alpha\sqrt{2S/C}(1+o(1)), S \to \infty,$$

where we use the assumption on equal costs $C_H = C_L$ and the result of Lemma 4.4.2 on the asymptotic behavior of the dropout threshold. Hence, we can calculate the limit:

$$\lim_{S\to\infty} \mathcal{U}^{\mathrm{dp}}/\mathcal{U}^{\mathrm{un}} = 1.$$

Group-dependent cost

If $\alpha \leq p_L$, then

$$\mathcal{U}^{\text{un}} = p_L(\alpha/p_L + o(1)) \cdot \theta_L^{\text{d}} \cdot (1 + o(1)) + p_H \cdot o(1) = \alpha \cdot \sqrt{2S/C_L} \cdot (1 + o(1)).$$

If $\alpha > p_L$, then

$$\begin{aligned}\mathcal{U}^{\text{un}} &= p_L(1 + o(1)) \cdot \theta_H^{\text{d}} \cdot (1 + o(1)) + p_H \cdot \left(\frac{\alpha - p_L}{p_H} + o(1)\right) \theta_H^{\text{d}}(1 + o(1)) \\ &= p_L(1 + o(1))\theta_H^{\text{d}} \cdot (1 + o(1)) + (\alpha - p_L)\theta_H^{\text{d}}(1 + o(1)).\end{aligned}$$

Hence, for $c := \sqrt{C_L/C_H}$, we can write that:

$$\lim_{S\to\infty} \mathcal{U}^{\text{un}}/\mathcal{U}^{\text{dp}} = \begin{cases} \lim_{S\to\infty} \frac{\alpha\theta_L^{\text{d}}(1+o(1))}{\alpha p_H \theta_H^{\text{d}}(1+o(1)) + \alpha p_L \theta_L^{\text{d}}(1+o(1))} = \frac{1}{cp_H + p_L} & \text{if } \alpha \leq p_L, \\ \lim_{S\to\infty} \frac{p_L\theta_H^{\text{d}}(1+o(1)) + (\alpha - p_L)\theta_H^{\text{d}}(1+o(1))}{\alpha p_H \theta_H^{\text{d}}(1+o(1)) + \alpha p_L \theta_L^{\text{d}}(1+o(1))} = \frac{c}{cp_H + p_L} & \text{if } \alpha > p_L, \end{cases}$$

where $\frac{c}{cp_H + p_L} < \frac{c}{cp_H + cp_L} = 1$.

4.8.8 Proof of Theorem 4.6.1

Equilibrium efforts Let us first find for which $\alpha \in (0, 1)$ we have the equality of effort at equilibrium, i.e., $m_H^{\text{un}} = m_L^{\text{un}}$. Since the equilibrium values are the best responses to θ^{un}, and the pure best response is unique (see Lemma 4.8.1), we can write:

$$\begin{aligned} m_H^{\text{un}} = m_L^{\text{un}} &\iff \frac{S}{C_H\tilde{\sigma}_H}\phi\left(\frac{\theta^{\text{un}} - m_H^{\text{un}}}{\tilde{\sigma}_H}\right) = \frac{S}{C_L\tilde{\sigma}_L}\phi\left(\frac{\theta^{\text{un}} - m_L^{\text{un}}}{\tilde{\sigma}_L}\right) \\ &\iff \theta^{\text{un}} - m_H^{\text{un}} = \pm\sqrt{-2\frac{\log(C_H\tilde{\sigma}_H) - \log(C_L\tilde{\sigma}_L)}{1/\tilde{\sigma}_H^2 - 1/\tilde{\sigma}_L^2}}. \end{aligned}$$

For $C_H\tilde{\sigma}_H > C_L\tilde{\sigma}_L$, there exist no real solution to the above equation. In this case, we can show that $m_H^{\text{un}} < m_L^{\text{un}}$ for all $\alpha \in (0, 1)$. Assume the opposite, i.e., $m_H^{\text{un}} > m_L^{\text{un}}$ for all θ, then $\phi((\theta - m_H^{\text{un}})/\tilde{\sigma}_H) < \phi((\theta - m_L^{\text{un}})/\tilde{\sigma}_L)$ and also $C_H\tilde{\sigma}_H > C_L\tilde{\sigma}_L$ which contradicts the initial assumption.

For $C_H\tilde{\sigma}_H \leq C_L\tilde{\sigma}_L$, there are two values of $\alpha_{m_H^{\text{un}}=m_L^{\text{un}}}$ which correspond to equal equilibrium efforts $m_H^{\text{un}} = m_L^{\text{un}}$:

$$\alpha^{(1,2)}_{m_H^{\text{un}}=m_L^{\text{un}}} = \sum_G p_G \Phi^c\left(\pm\frac{\sqrt{-2\frac{\log(C_H\tilde{\sigma}_H)-\log(C_L\tilde{\sigma}_L)}{1/\tilde{\sigma}_H^2-1/\tilde{\sigma}_L^2}}}{\tilde{\sigma}_G}\right).$$

We can verify that $db_H/d\alpha > db_L/d\alpha$ for $\alpha^{(1)}_{m_H^{\text{un}}=m_L^{\text{un}}}$ and $db_H/d\alpha < db_L/d\alpha$ for $\alpha^{(2)}_{m_H^{\text{un}}=m_L^{\text{un}}}$ which concludes the proof.

Equilibrium selection rates Let us find such $\alpha_{r_H=r_L}$ for which both groups are selected at equal rates, i.e., $r_H^{\text{un}} = r_L^{\text{un}}$. This is equivalent to:

$$\Phi^c\left(\frac{\theta^{\text{un}} - m_H^{\text{un}}}{\tilde{\sigma}_H}\right) = \Phi^c\left(\frac{\theta^{\text{un}} - m_L^{\text{un}}}{\tilde{\sigma}_L}\right) = \alpha_{r_H=r_L} \iff \theta^{\text{un}} = m_G^{\text{un}} + \tilde{\sigma}_G\Phi^{-1}(1-\alpha_{r_H=r_L}),\ \forall G \in \{H, L\}.$$

By definition of the best response, we can also write that:

$$m_G^{\text{un}} = \frac{S}{C_G\tilde{\sigma}_G}\phi\left(\frac{\theta^{\text{un}} - m_G^{\text{un}}}{\tilde{\sigma}_G}\right) = \frac{S}{C_G\tilde{\sigma}_G}\phi\left(\Phi^{-1}(1-\alpha_{r_H=r_L})\right),$$
$$\theta^{\text{un}} = \frac{S}{C_G\tilde{\sigma}_G}\phi\left(\Phi^{-1}(1-\alpha_{r_H=r_L})\right) + \tilde{\sigma}_G\Phi^{-1}(1-\alpha_{r_H=r_L}).$$

The equality for θ^{un} is possible only for such α, when

$$\frac{S}{C_H\tilde{\sigma}_H}\phi\left(\Phi^{-1}(1-\alpha_{r_H=r_L})\right) + \tilde{\sigma}_H\Phi^{-1}(1-\alpha_{r_H=r_L}) = \frac{S}{C_L\tilde{\sigma}_L}\phi\left(\Phi^{-1}(1-\alpha_{r_H=r_L})\right) + \tilde{\sigma}_L\Phi^{-1}(1-\alpha_{r_H=r_L}),$$

which is if and only if

$$S\left(\frac{1}{C_H\tilde{\sigma}_H} - \frac{1}{C_L\tilde{\sigma}_L}\right)\phi\left(\Phi^{-1}(1-\alpha_{r_H=r_L})\right) = \Phi^{-1}(1-\alpha_{r_H=r_L})(\tilde{\sigma}_L - \tilde{\sigma}_H).$$

We solve the corresponding equation by letting $z \coloneqq \Phi^{-1}(1-\alpha_{r_H=r_L})$ and solving with respect to z. After, $\alpha_{r_H=r_L} = 1 - \Phi(z)$.

Let $\xi = S\left(\frac{1}{C_H\tilde{\sigma}_H} - \frac{1}{C_L\tilde{\sigma}_L}\right)/(\tilde{\sigma}_L - \tilde{\sigma}_H)$. Then by definition of ϕ:

$$\xi\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} = z \iff z\cdot e^{\frac{z^2}{2}} = \xi\frac{1}{\sqrt{2\pi}}.$$

By squaring both sides of the above equation, we end up with the equation of type $y \cdot e^y = const$ which can be solved using the Lambert $\mathcal{W}$ function defined as the inverse to $f(y) = y \cdot e^y$. As a result, by solving for z^2 and taking the square root with the sign equal to the sign of ξ, we show:

$$\alpha_{r_H=r_L} = 1 - \Phi\left(\text{sgn}(\xi)\sqrt{\mathcal{W}\left(\frac{\xi^2}{2\pi}\right)}\right).$$

We can verify that the $dr_H/d\alpha > dr_L/d\alpha$ at $\alpha_{r_H=r_L}$, so for all $\alpha < \alpha_{r_H=r_L} = 1 - \Phi\left(\text{sgn}(\xi)\sqrt{\mathcal{W}\left(\frac{\xi^2}{2\pi}\right)}\right)$, we have that $r_H^{\text{un}} < r_L^{\text{un}}$.

Chapter 5

The Price of Local Fairness in Multistage Selection

This chapter is based on our publication (Emelianov et al., 2019).

The code to generate all figures can be found at:
https://github.com/vitaly-emelianov/multistage_fairness/

Abstract In this chapter, we study fairness in k-stage selection problems where additional features are observed at every stage. We first introduce two fairness notions, *local* (per stage) and *global* (final stage) fairness, that extend the classical fairness notions to the k-stage setting. We propose a simple model based on a probabilistic formulation and show that the locally and globally fair selections that maximize precision can be computed via a linear program. We then define the *price of local fairness* to measure the loss of precision induced by local constraints; and investigate theoretically and empirically this quantity. In particular, our experiments show that the price of local fairness is generally smaller when the sensitive attribute is observed at the first stage; but globally fair selections are more locally fair when the sensitive attribute is observed at the second stage—hence in both cases it is often possible to have a selection that has a small price of local fairness and is close to locally fair.

5.1 Introduction

The existing literature on fairness in selection problems typically considers one-shot decision processes whereby, from a set of features observed about an individual—one of them being a 'sensitive feature' based on which discrimination is defined—, one needs to decide whether or not to "select" him/her (where select can mean hire, grant a loan or parole, etc. depending on the context). The problem in this setting is how to learn a decision rule from past data that respects certain fairness constraints. In many applications, however, decisions are made in multiple stages. In hiring for

instance, a subset of candidates is first selected for interview based on resume (or high-level candidate's information) and a final selection is then made from the subset of interviewed candidates. In police practices, there are often multiple stages of decisions with increasingly high levels of investigation of the individuals not released at the previous stage; as for instance in the famous stop-question-and-frisk practice by the New-York City Police Department.

A distinctive specificity of the multistage setting, besides the fact that decisions are made in multiple stages, is that in many cases additional features get known at later stages for the subset of individuals selected at earlier stages, but one needs to make the early-stage selection without observing those features. This raises a number of new questions that are fundamental to fair multistage selection. First, given that there are multiple layers of decisions, *how should fairness be defined?* In particular, should it be defined at each individual stage, on the final decision, or otherwise? Second, given that one has to make decisions with only partial information at early stages, *how to make an optimal selection?* Finally, given that the sensitive feature can be observed at different stages, *is it better to observe the sensitive feature at earlier or later stages (for both fairness and utility)?* This last question intuitively relates to recurrent public debates such as "should gender identification be removed from CVs?".

In this chapter, we study the k-stage selection problem, in which there is a fixed limit (or budget) of candidates that can be selected at each stage (as is natural in the applications discussed). To tackle the questions above, we propose a simple model based on a probabilistic formulation in which we assume perfect knowledge of the joint distribution of features at all stages and of the conditional probability of being a desirable candidate conditioned on feature values. Based on this model, we are then able to make the following contributions.

We introduce two meaningful notions of fairness for the k-stage setting: *local fairness* (the selection is fair at each stage) and *global fairness* (only the final selection needs to be fair). These definitions extend classical group fairness notions for one-stage decision making (such as demographic parity or equal opportunity) to the multistage setting and they apply regardless of when the sensitive feature is observed (at first stage or later). We show that local fairness implies global fairness and we propose a linear formulation of the problem that allows us to compute the selection algorithm that maximizes precision while satisfying (local or global) fairness and per-stage budget constraints in expectation. As local fairness is a more restrictive condition, the precision of the optimal globally fair algorithm is naturally higher than for the locally fair algorithm. To capture this gap, we define the *price of local fairness* ($PoLF$) as the ratio of the two and prove a simple upper bound—showing that imposing local fairness cannot be arbitrarily bad. We also define the notion of violation of local fairness ($VoLF$) to capture how far from locally fair the optimal globally fair algorithm is.

Finally, we conduct a numerical study in a two-stage setting using three classical datasets. Our results show that the $PoLF$ can be large (up to 1.6 in some cases). This implies that in some cases, enforcing local fairness constraints can reduce the precision by 60% compared to a globally fair algorithm. The $VoLF$ is also sometimes large (up to

0.6 in our experiments), which means that imposing only a global fairness constraint can be highly unfair at intermediate stages. We finally compare what happens when the sensitive feature is observed at the first stage or at the second stage. We find that the $PoLF$ is generally higher when the sensitive feature is observed at the second stage; while conversely the $VoLF$ is generally higher when the sensitive feature is observed at the first stage. These results show that, in most cases, it is possible to get at least approximate fairness at each stage and precision close to globally-fair optimal together; either by imposing local fairness if the sensitive feature is observed at first stage (where $PoLF$ is small) or by hiding the sensitive feature at first stage and using a globally fair algorithm (which is close to locally fair since $VoLF$ is then small).

Overall, our results provide intuitive answers towards better understanding fairness in multistage selection. To that end, we intentionally used the simplest model that captures the main features of a multistage selection problem and how an optimal selection algorithm is affected by the fairness notion considered and the time at which the sensitive feature is observed—rather than using a more practical but complex model. We believe that it is a good abstraction to start with, but we elaborate further on our model's limitations in Section 5.7.

5.2 Related Work

As mentioned earlier, there have been many recent works on defining fairness and constructing algorithms that respect those definitions for the case of one-stage decision making (Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2012; Hardt et al., 2016b; Kilbertus et al., 2017; Kleinberg et al., 2017; Lipton et al., 2018; Pedreshi et al., 2008; Zafar et al., 2017b). In this work, we focus on two classical notions of fairness for the one-shot classification setting: demographic parity (or disparate impact) and equal opportunity (or disparate mistreatment) (Hardt et al., 2016b; Zafar et al., 2017b). There are also works on fairness in sequential learning (Heidari and Krause, 2018; Jabbari et al., 2017; Joseph et al., 2016; Valera et al., 2018). The model in those papers is to sequentially consider each individual and make decision for them, but there is no notion of refining selection through multiple stages by getting additional features.

Closer to our work, a few papers investigate multistage classification/selection without fairness considerations (Senator, 2005; Trapeznikov et al., 2012). Schumann et al. (2019) model the interview decisions in hiring as a multi-armed bandit problem and consider getting extra features at a cost for a subset of candidates, but they do not have fairness constraints: they propose an algorithm for their bandit problem and show that it leads to higher diversity than other algorithms.

To the best of our knowledge, our model is the first that proposes concrete fairness notions for multistage selection and algorithms to maximize utility under fairness constraints. The only other papers discussing fairness in the context of two-stage or composed decision making are (Bower et al., 2017; Dwork and Ilvento, 2019), but they do not model additional features becoming available at the second stage for the subselected individuals, which is the key element of our analysis.

5.3 Multistage Selection Framework

5.3.1 Basic Setting and Notation

Assume that there is a continuum of candidates[1] of a unit mass each described by d features, and consider the following k-stage selection process. At the first stage, we observe some of the features $x_1, \dots, x_{d_1}$ of all candidates where $d_1 < d$. We then select α_1 proportion of them that "pass" to the second stage. At the second stage, we observe some extra features of these α_1 candidates $x_{d_1+1}, \dots, x_{d_2}$ ($d_1 < d_2$) that were not known at the previous stage. Using the features of both stages, we do a selection, from the α_1 that passed the first stage of $\alpha_2 \leq \alpha_1$ candidates that pass to the next stage, and so on. At the last stage k, we observe all $d_k = d$ features of the α_{k-1} candidates and select α_k among those who passed the stage $k-1$.

We assume that each candidate is endowed with a *label* $y \in \{0, 1\}$, which encodes whether the candidate is "good" or "bad" according to the purpose of the selection, i.e., if $y = 1$ we would like to have this candidate in our final selection, if $y = 0$ we would prefer not. The label y is not known until the end and is therefore not available to make the selection.

We assume that the decision maker knows the joint distribution of features and the conditional probability that expresses the probability that the candidate is "good" given all its features. We will denote by $p_{x_1 \dots x_d} = \mathrm{P}(X_1 = x_1, \dots, X_d = x_d)$ the probability to observe a specific realization of features and by $p^{y=1}_{x_1 \dots x_d} = \mathrm{P}(Y = 1 | X_1 = x_1, \dots, X_d = x_d)$ the probability that a candidate is good ($y = 1$) given its features $x_1, \dots, x_d$.

5.3.2 Probabilistic Selection and Budget Constraints

In the following, we will consider a class of selection algorithms that perform a probabilistic selection of candidates. Such an algorithm takes as an input a list of probability values $p^{(i|i-1)}_{x_1 \dots x_{d_i}}$ for all stages $i \in \{1 \dots k\}$ and all possible combination of features. Then, for each candidate that passed stage $i-1$ and has features $(x_1 \dots x_{d_i})$, the algorithm selects this candidate for the next stage with probability $p^{(i|i-1)}_{x_1 \dots x_{d_i}}$, with the convention that everyone passes stage 0.

For each stage i, we define a binary predictor $\hat{Y}_i$ that is equal to 1 if the candidate is selected at stage i (by convention, $\hat{Y}_0 = 1$ for all candidates). We assume that, *on average*, the proportion of candidates that can be selected by the algorithm at stage i is at most α_i and exactly α_k for the last stage, with $1 \geq \alpha_1 \geq \cdots \geq \alpha_k$. We denote by $\boldsymbol{\alpha}_{-k} = (\alpha_1, \cdots, \alpha_{k-1})^T$ the selection sizes of the first $k-1$ stages.

5.3.3 Performance Metric

We measure the performance of a given selection algorithm in terms of precision. The precision is the fraction of the selected candidates that indeed were "good" for selection:

[1] We use the term candidates in a generic sense to refer to elements of the initial set that can be selected.

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \mathrm{P}(Y = 1 | \hat{Y}_k = 1),$$

where the denominator is the number of selected candidates.

The choice of precision may seem arbitrary but it is in fact a very natural metric when the size of the final selection is fixed as in our setting. Indeed, maximizing precision is then equivalent to maximizing most other meaningful metrics as formalized in the next theorem.

Theorem 5.3.1. *Assume that the selection size* $\mathrm{P}(\hat{Y}_k = 1)$ *is fixed (to* α_k*). Then maximization of precision is equivalent to maximization of true positive rate, true negative rate, accuracy and* f_1*-score; and to minimization of false positive rate and false negative rate.*

5.4 Fairness Notions in Multistage Setting

In this section, we propose new notions of fairness for the multistage selection problem. We assume that there exists, amongst all features that describe candidates, a sensitive feature $G \in \{A, B\}$ that indicates whether or not a candidate belongs to a sensitive group that should not be discriminated against.

The literature has introduced multiple definitions of fairness for the single-stage setting (and it is worth mentioning that in most of the cases those fairness criteria cannot be satisfied simultaneously (Chouldechova, 2017)). The most relevant notions in the context of selection problems are *demographic parity* (DP) and *equal opportunity* (EO). We first recall the definition of these fairness criteria in the traditional setting of single-stage selection. We then extend them to the multistage setting by showing that there are essentially two relevant notions of fairness: *local* and *global* fairness.

5.4.1 Classical Fairness Notions in Single-Stage

Let $\hat{Y}$ be a binary predictor that decides which candidates belong to the selection. The first fairness definition, widely known as *demographic parity*, states that the predictor $\hat{Y}$ is fair if it is statistically independent from the sensitive attribute $G \in \{A, B\}$.

Definition 5.4.1 (Demographic Parity, DP)**.** *The binary predictor* $\hat{Y}$ *satisfies DP with respect to* G *if* $\hat{Y}$ *and* G *are independent:*

$$\mathrm{P}(\hat{Y} = 1 | G = A) = \mathrm{P}(\hat{Y} = 1 | G = B). \tag{5.1}$$

DP does not take into account the actual label Y. Hardt et al. (2016b) and Zafar et al. (2017b) argue that DP is not the most relevant notion of fairness in cases where we have ground truth on the quality of the candidates (which is our case since we assume statistical knowledge of the probabilities of labels). In such cases, one might want to be fair among the candidates that are worth selecting, a metric called *equal opportunity* (Hardt et al., 2016b) (an equivalent notion called disparate mistreatment is proposed in (Zafar et al., 2017b)):

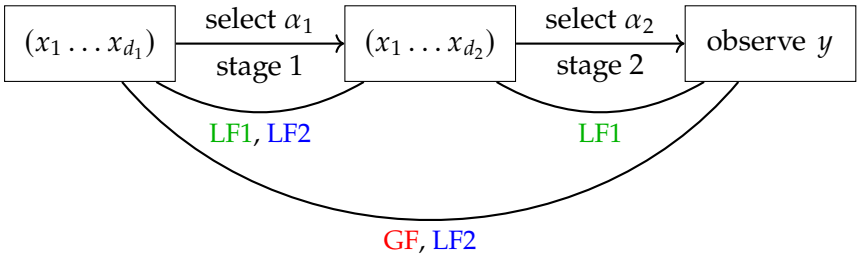


Figure 5.1 – Illustration of the different fairness definitions for a two-stage selection.

Definition 5.4.2 (Equal Opportunity, EO (Hardt et al., 2016b))**.** *The binary predictor $\hat{Y}$ satisfies EO with respect to G if $\hat{Y}$ and G are independent given that $Y = 1$:*

$$\mathrm{P}(\hat{Y} = 1 | Y = 1, G = A) = \mathrm{P}(\hat{Y} = 1 | Y = 1, G = B). \tag{5.2}$$

In the remainder, we systematically consider DP and EO.

5.4.2 Local and Global Fairness in Multistage

Existing fairness notions apply to single-stage selection, where we have only one binary predictor $\hat{Y}$. In the case of k-stage selection, we have k binary predictors $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_k)$. In this section, we develop different notions of fairness that extend existing notions to the k-stage selection setting.

We propose three definitions that we believe correspond to three reasonable notions of fairness. The high-level idea of each definition is depicted in Fig. 5.1. For the sake of brevity of exposition, we present the formal definitions for the demographic parity criterion, the translation to EO (or to any other fairness notion) being straightforward.

The first fairness notion, *local fairness 1 (LF1)*, imposes that the selection be fair at every stage with respect to the set of candidates that reached that stage. In other words the selection of each stage i is fair with respect to the population that "passed" stage $i - 1$.

Definition 5.4.3 (Local Fairness 1, LF1)**.** *A k-stage selection algorithm satisfies LF1 if (for the case of DP), $\forall i \in \{1, \cdots, k\}$:*

$$\mathrm{P}(\hat{Y}_i = 1 | \hat{Y}_{i-1} = 1, G = A) = \mathrm{P}(\hat{Y}_i = 1 | \hat{Y}_{i-1} = 1, G = B).$$

The second fairness notion that we propose, *local fairness 2 (LF2)*, prescribes that the selection should be fair at each stage with respect to the initial set of candidates.

Definition 5.4.4 (Local Fairness 2, LF2)**.** *A k-stage selection algorithm satisfies LF2 if (for the case of DP), $\forall i \in \{1, \cdots, k\}$:*

$$\mathrm{P}(\hat{Y}_i = 1 | G = A) = \mathrm{P}(\hat{Y}_i = 1 | G = B).$$

In the last definition, *global fairness (GF)*, we allow the predictor $\hat{Y}_i$ to be unfair at each stage before the last, but we require the final decision $\hat{Y}_k$ to be fair with respect to the initial set of candidates.

Definition 5.4.5 (Global Fairness, GF)**.** *A k-stage selection algorithm satisfies GF if (for the case of DP):*

$$\mathrm{P}(\hat{Y}_k = 1|G = A) = \mathrm{P}(\hat{Y}_k = 1|G = B).$$

Note that the above definitions can be adapted to EO by conditioning on $Y = 1$ in all formulas.

5.4.3 Equivalence between LF1 and LF2

In the following theorem, we show that both notions of local fairness, LF1 and LF2 are equivalent. Therefore in the rest of the chapter, we will simply name a multistage selection algorithm that satisfies LF1 (and thus LF2) as a being *locally fair* (LF). An algorithm satisfying the global fairness definition will be called *globally fair* (GF).

Theorem 5.4.6 (Relations between fairness notions)**.** *For both DP and EO:*

(i) A selection algorithm satisfies LF1 if and only if it satisfies LF2. We call such an algorithm locally fair (LF).

(ii) A locally fair selection algorithm is globally fair (GF).

5.5 Utility Maximization as a Linear Program

Our goal is to find the binary predictors $(\hat{Y}_1, \ldots, \hat{Y}_k)$ corresponding to stages from 1 to k, respectively, that maximize precision while respecting budget and fairness constraints:

$$\begin{aligned} \max_{\hat{Y}_1,\ldots,\hat{Y}_k} \quad & \mathrm{P}(Y = 1|\hat{Y}_k = 1) \\ & \mathrm{P}(\hat{Y}_i = 1) \le \alpha_i, \qquad i \le k-1 \\ & \mathrm{P}(\hat{Y}_k = 1) = \alpha_k \\ & f_j(\hat{Y}_1, \ldots, \hat{Y}_k) = 0, \quad j \le t \end{aligned} \tag{5.3}$$

where functions $f_j(\cdot)$ of the binary predictors correspond to the fairness constraints we impose. For instance, for a globally fair algorithm (DP) we have only one fairness constraint: $f(\hat{Y}_1, \ldots, \hat{Y}_k) = \mathrm{P}(\hat{Y}_k = 1|G = A) - \mathrm{P}(\hat{Y}_k = 1|G = B)$.

Using the assumption that the final stage size constraint is $\mathrm{P}(\hat{Y}_k = 1) = \alpha_k$ we can write the precision as follows:

$$\mathrm{P}(Y=1|\hat{Y}_k=1) = \frac{1}{\alpha_k} \sum_{x_1\ldots x_d} p^{y=1}_{x_1\ldots x_d} p_{x_1\ldots x_d} \prod_{j=1}^{k} p^{(j|j-1)}_{x_1\ldots x_{d_j}}. \tag{5.4}$$

Using the notation introduced in Section 5.3.2, the probability $\mathrm{P}(\hat{y}_i = 1)$ that candidate passes stage i is

$$\mathrm{P}(\hat{Y}_i = 1) = \sum_{x_1\ldots x_d} p_{x_1\ldots x_d} \prod_{j=1}^{i} p^{(j|j-1)}_{x_1\ldots x_{d_j}}. \tag{5.5}$$

Hence, the constraints on the selection size $\mathrm{P}(\hat{Y}_i = 1) \leq \alpha_i$ for $i < k$ and $\mathrm{P}(\hat{Y}_k = 1) = \alpha_k$ can be expressed using (5.5).

The fairness constraints can be developed in the same manner, e.g., for the globally fair case (DP):

$$f(\hat{Y}_1, \ldots, \hat{Y}_k) = \mathrm{P}(\hat{Y}_k = 1 | G = A) - \mathrm{P}(\hat{Y}_k = 1 | G = B),$$

where $\forall g \in \{A, B\}$,

$$\mathrm{P}(\hat{Y}_k = 1 | G = g) = \frac{\sum\limits_{x_i} \prod_{j=1}^{k} p^{(j|j-1)}_{x_1 \ldots x_{d_j}} \cdot p_{x_1 \ldots g \ldots x_d}}{\sum\limits_{x_i} p_{x_1 \ldots g \ldots x_d}}. \tag{5.6}$$

From (5.4), we see that the objective is not linear in the variables $p^{(j|j-1)}_{x_1 \ldots x_{d_j}}$ due to the product of probabilities. Similarly, we observe from (5.5) and (5.6) that the constraints are also not linear in these variables. However, we can show that by using the change of variables $\tilde{p}^{(i|i-1)}_{x_1 \ldots x_{d_i}} = \prod_{j=1}^{i} p^{(j|j-1)}_{x_1 \ldots x_{d_j}}$, it can be made linear. This shows that it is possible to compute the variables $p^{(j|j-1)}_{x_1 \ldots x_{d_j}}$ that maximize precision (5.3) using a linear program (LP) (see details in Section 5.8.1), which is key to applicability. It should be noted, however, that the number of variables in (LP) grows exponentially with the number of features.

To distinguish between the different notions of fairness, we will denote by $\mathcal{U}^{\mathrm{LF}}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ and $\mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ the value of the problem (LP)—i.e., the maximum utility—when the fairness constraints correspond to local and global fairness, respectively. Similarly, we will denote by $\mathcal{U}^{\mathrm{un}}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ the optimal precision value when no fairness constraint are imposed (we call it the *un*fair case).

5.5.1 Solution Properties wrt Budget Constraints

The selection sizes may be related to some budget or to some physical resources of our problem and are crucial parameters. As we show in the next theorem, the optimal utility values are monotonic and concave as functions of budget sizes $\alpha_1, \ldots, \alpha_{k-1}$. This property can be useful for budget optimization and is illustrated as well in Fig. 5.2.

Theorem 5.5.1 (Monotonicity and concavity). *For $\mathcal{U} \in \{\mathcal{U}^{\mathrm{LF}}, \mathcal{U}^{\mathrm{GF}}, \mathcal{U}^{\mathrm{un}}\}$ and any fairness constraints that can be expressed as linear homogeneous equations[2] (such as DP and EO), we have that $\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ is*

(i) non-decreasing and concave with respect to $\boldsymbol{\alpha}_{-k}$;

(ii) non-increasing with respect to α_k.

Note that $\mathcal{U}$ can be concave or convex or none of the two with respect to α_k, depending on the problem's parameters.

[2]See details in Lemma 5.8.2 in Section 5.8.1.

5.5.2 The Price of Local Fairness

We are now ready to define our central notion—the *price of local fairness*—that represents the price to pay for being fair at intermediate stages compared to a globally fair solution.

Definition 5.5.2 (Price of Local Fairness, $PoLF$)**.** *Let*

$$PoLF(\boldsymbol{\alpha}_{-k}, \alpha_k) = \frac{\mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k)}{\mathcal{U}^{\mathrm{LF}}(\boldsymbol{\alpha}_{-k}, \alpha_k)}.$$

It should be clear that the locally fair algorithm is more constrained than the globally fair. Thus, we have:

$$\mathcal{U}^{\mathrm{LF}}(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \mathcal{U}^{\mathrm{un}}(\boldsymbol{\alpha}_{-k}, \alpha_k).$$

This implies that the values of $PoLF(\boldsymbol{\alpha}_{-k}, \alpha_k)$ are always larger than or equal to 1. Using only the final selection size α_k, it is also possible to compute an upper bound as follows.

Theorem 5.5.3 ($PoLF$ bound)**.** *For all* $(\boldsymbol{\alpha}_{-k}, \alpha_k)$*, we have:*

$$1 \leq PoLF(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \min\left(\frac{1}{\alpha_k}, \frac{1}{\mathrm{P}(Y=1)}\right).$$

For instance, if the final stage selection size is $\alpha_k = 0.3$ (as in our numerical examples), the globally fair algorithm can outperform the locally fair one by a factor at most 3.33. While this bound is probably loose, we will see in our numerical example that the $PoLF$ can be as large as 1.6 on real data.

5.6 Empirical Analysis

In this section we implement[3] the optimization algorithms in order to capture tendencies on real datasets and to provide general insights. We consider the two-stage selection process, since it is the most easily interpretable. Thus, $\boldsymbol{\alpha}_{-k} = \alpha_1$ and $\alpha_k = \alpha_2$. In our experiments we use three datasets: `Adult` (Dua and Graff, 2017), `COMPAS` (Larson et al., 2016) and `German Credit Data` (Dua and Graff, 2017). We adapt these datasets to our two stage fair selection problem by leaving 6 features, binarizing them (see details in Section 5.9) and artificially separating in two stages. We estimate the statistics $p_{x_1 \ldots x_d}$ and $p^{y=1}_{x_1 \ldots x_d}$ from data. We then use a linear solver for the linear program (LP) that gives us the optimal utility $\mathcal{U}(\alpha_1, \alpha_2)$ for the fair and unfair cases.

5.6.1 Analysis of the Price of Local Fairness

We consider three different scenarios: (i) the sensitive attribute G is observed at the first stage; (ii) at the second stage; (iii) never used in the selection process. We distinguish

[3] All codes are available at https://github.com/vitaly-emelianov/multistage_fairness/

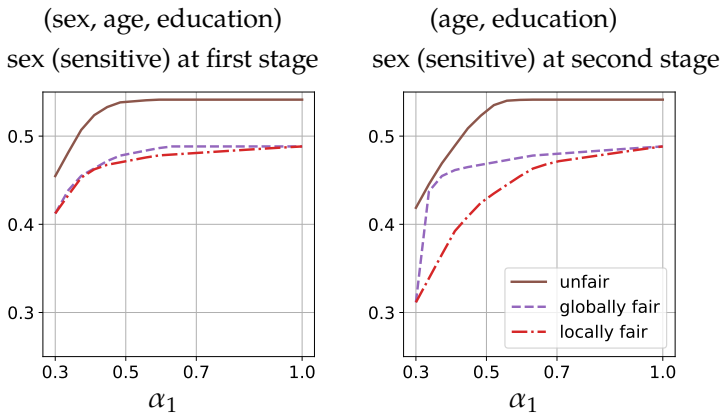


Figure 5.2 – Utility $\mathcal{U}(\alpha_1, \alpha_2 = 0.3)$ for `Adult` dataset (DP).

these three cases since it could happen that the use of the sensitive attribute G in decision making is forbidden at some stages or even at all (by law or other conventions). Our aim is to compare how the price of local fairness behaves in every case.

Let us start with a simple example. We leave 5 features from the `Adult` dataset: *sex*, *age*, *education*, *relationship* and *native country* and consider the attribute *sex* as sensitive. Fig. 5.2 then shows the values of $\mathcal{U}^{\{un,\, GF,\, LF\}}(\alpha_1, \alpha_2)$ as a function of α_1 for fixed $\alpha_2 = 0.3$ when using the features displayed on top of each subfigure at first stage and the rest at second stage. We make two important observations from this figure. *First*, the value of $PoLF$ can be significant. From Fig. 5.2-(right), we see that for $\alpha_1 \approx 0.33$, the value of $PoLF$ is about 1.3, meaning that the globally fair algorithm achieves 30% larger value of precision than the locally fair. *Second*, the gap between LF and GF algorithms is significantly larger when the sensitive attribute G is observed at the second stage.

To show that this behavior is significant we calculate the values of $\mathcal{U}(\alpha_1, \alpha_2)$ for every possible combination $X = \{X_1, \ldots, X_5\}$ of 5 features out of 6 as decision variables (X_1, X_2 at first stage and X_3, X_4 at second stage), with one sensitive attribute $G = X_5$ that can be observed at the first stage or at the second stage or not observed at all, and for every possible (discretized) value of $\alpha_1 \geq \alpha_2$. Due to space constraints we present our results only for the DP definition of fairness; we emphasize that the observations are robust among the three datasets and the two fairness notions (DP and EO) (see Section 5.9 for additional results). Fig. 5.3 shows the empirical cumulative distribution functions $\hat{F}_{PoLF}(x)$ of the values of $PoLF$ obtained. We observe that the *price of local fairness is significantly lower when the sensitive attribute G is revealed at the first stage* compared to the case where it is revealed later. This is consistent with the observation made in Fig. 5.2. A possible interpretation is that the LF algorithm has to make a conservative decision at the first stage and therefore cannot perform well compared to the GF algorithm that is able to compensate (when the sensitive feature G is observed) for the unfair decisions that have been made at the first stage. It is worth mentioning that we have the same observation for a three-stage algorithm: the later we reveal the sensitive attribute, the higher the values of $PoLF$ we obtain (see Section 5.9).

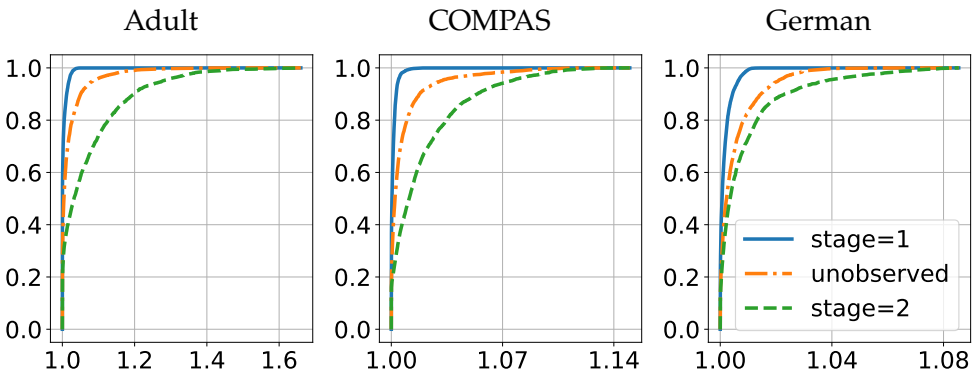


Figure 5.3 – Empirical CDFs of $PoLF$ for all datasets (DP, $\alpha_2 = 0.3$).

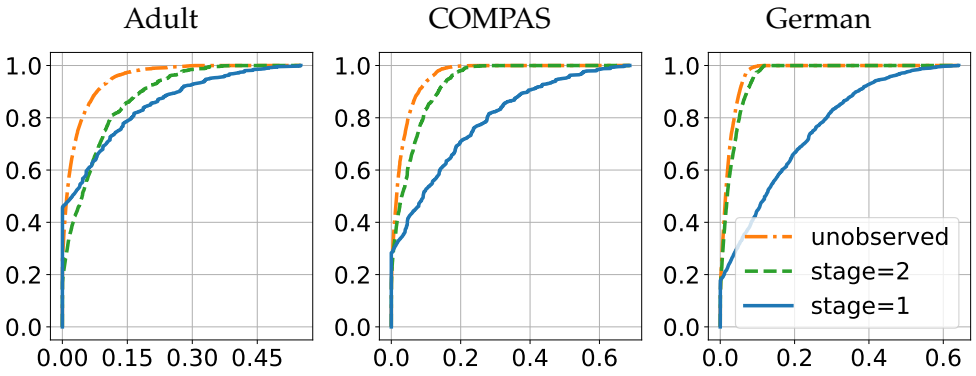


Figure 5.4 – Empirical CDFs of $VoLF$ for all datasets (DP, $\alpha_2 = 0.3$).

5.6.2 Violation of Local Fairness

By definition, a globally fair algorithm can violate fairness constraints at intermediate stages. For a given budget constraints α_1, α_2, we define the violation of local fairness ($VoLF$) as the absolute value of the fairness constraint violation at the first stage for the optimal globally fair algorithm. For instance, for DP, this quantity equals:

$$VoLF(\alpha_1, \alpha_2) = \left|\mathrm{P}(\hat{Y}_1 = 1|G = A) - \mathrm{P}(\hat{Y}_1 = 1|G = B)\right|.$$

In Fig. 5.4, we show the empirical cumulative distribution function of violation of fairness $\hat{F}_{VoLF}(x)$ for every value of $\alpha_1 \in [\alpha_2; 1]$ and for every feature combination. We observe that *the later the sensitive feature G is revealed (or even not revealed), the more fair at intermediate stages the globally fair algorithm is*. One possible explanation is that an algorithm that cannot observe the sensitive feature G at the first stage has to be more "cautious" at every stage to be able to satisfy global fairness since the exact value of sensitive attribute G is not available. This observation is again robust among different datasets and notions of fairness.

Finally, in Fig. 5.5 we represent the joint distribution of $PoLF$ and $VoLF$. As mentioned before, the globally fair algorithm is more unfair at the intermediate stages when the sensitive feature G is observed from the beginning (left panel), however the price of local fairness we pay in this case is the smallest one. When the sensitive feature

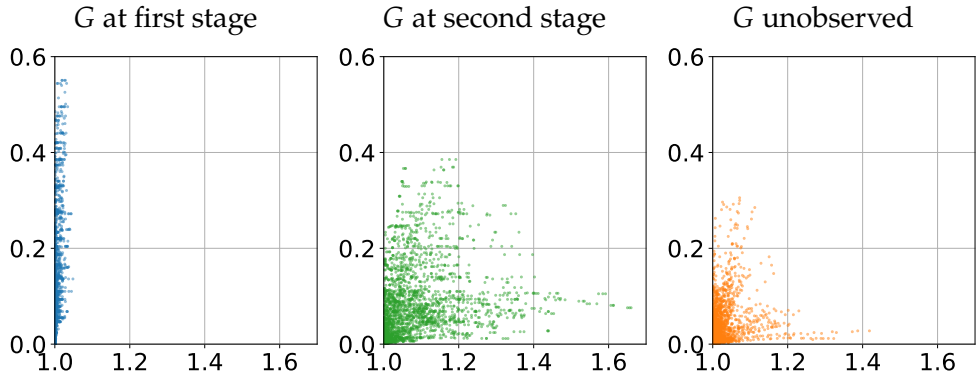


Figure 5.5 – $VoLF$ (y-axis) vs $PoLF$ (x-axis) for `Adult` dataset (DP, $\alpha_2 = 0.3$).

G is observed at the second stage (middle panel) the globally fair algorithm is more locally fair compared to the previous case, but the value of $PoLF$ is way larger. Finally, when G is never observed (right panel) the globally fair algorithm is the "most locally fair" among all three settings. We finally observe that, while most points have either $PoLF$ small (i.e., using a LF algorithm does not lose much) or $VoLF$ small (i.e., the GF algorithm is almost locally fair), there exist some points—when the sensitive feature is observed at the second stage—where both $PoLF$ and $VoLF$ are large; i.e., imposing local fairness even approximately comes at a significant cost.

5.7 Conclusion and Discussion

In this work we tackle the problem of multistage selection and the fairness issues it entails. We propose a stylized model based on a probabilistic formulation of the k-stage selection problem with constraints on the number of selected individuals at each stage that should hold in expectation. We introduce two different notions of fairness for the multistage setting: local (under two equivalent variants) and global fairness. Thanks to this framework, we show that maximizing precision under budget and fairness constraints can be done via linear programming, which enables for efficient computation as well as theoretical investigation. In particular, we analyze theoretically and empirically how the utility of locally and globally fair algorithms vary with selection budgets, and we find that globally fair algorithms can lead to non-negligible performance increases compared to locally fair ones.

One of the main findings of our work is that the stage at which the sensitive attribute is revealed greatly affects the difference between the performance of locally and globally fair algorithms: hiding the sensitive feature at early stages tends to make globally fair algorithm more fair at intermediate stages. While locally fair algorithms may be desirable, our results show that local fairness does not come for free. They also show that if a decision maker would like to encourage locally fair selection algorithms, there are essentially two choices: either hide the sensitive feature at the first stage or impose by rules the first stage to be fair.

Our model allows us to provide elegant insights into the fairness questions related

to multistage selection, yet it does a number of simplifying assumptions that naturally restrict its direct applicability. *First*, our model ignores the issue that the selection probability at a stage depends on which candidates got selected at the previous stages; i.e., it implicitly makes the approximation that at each stage the number of candidates selected for each feature combination is equal to its expectation. In Section 5.8.2, we show that this approximation becomes exact as n tends to infinity. *Second*, we assume perfect statistical knowledge of the joint distribution of features and label values, without bias. *Third*, we consider only discrete features and use a non-compact representation of the selection probabilities—this allows us to solve the exact selection problem by using an LP formulation. Relaxing these assumptions, in particular using a more compact representation of the selection algorithm (at the cost of a loss of precision) is an interesting direction of future work.

5.8 Omitted Proofs

In this section, we provide the proofs of all results stated in the chapter. We start by introducing notation that will be used throughout the proofs.

Notation

To ease the exposition, we introduce the following matrix notation for the problem (5.8).

- *Selection probabilities* $\mathbf{p}$. We concatenate all the selection probabilities in a single vector:

$$\mathbf{p} = (\tilde{p}^{(i|i-1)}_{x_1 \dots x_{d_i}})^T,$$

where $\tilde{p}^{(i|i-1)}_{x_1 \dots x_{d_i}}$ is the vector of selection probabilities at stage i for all possible values of $x_1 \dots x_{d_i}$ (whose size depends on i).

- *Constraints set* $C_{\boldsymbol{\alpha}_{-k}, \alpha_k}$. We have two types of constraints in problem (5.8).

1. The constraints that correspond to selection sizes α_i, $i = 1, \dots, k$. We separate them such that $\mathbf{A}\mathbf{p} \leq \boldsymbol{\alpha}_{-k}$ corresponds to selection at first $k-1$ stages, so

$$\boldsymbol{\alpha}_{-k} = (\alpha_1, \dots, \alpha_{k-1})^T.$$

   The constraint $\mathbf{b}^T\mathbf{p} = \boldsymbol{\alpha}_k$ corresponds to selection at the last stage, where we require a strict equality.

2. The constraints that do not depend on selection sizes are written in a form of $\mathbf{D}\mathbf{p} \leq \boldsymbol{\delta}$ for an appropriate $\mathbf{D}$, where $\boldsymbol{\delta} = (1, \dots, 1, 0, \dots, 0)^T$: 1's in $\boldsymbol{\delta}$ correspond to constraints $0 \leq \tilde{p}^{(1|0)}_{x_1 \dots x_{d_1}} \leq 1$ and 0's correspond to constraints $0 \leq \tilde{p}^{(i|i-1)}_{x_1 \dots x_{d_i}} \leq \tilde{p}^{(i-1|i-2)}_{x_1 \dots x_{d_{i-1}}}$.

Thus, we write every constraint in matrix form and introduce the following compactly formed constraint set:

$$C_{\boldsymbol{\alpha}_{-k}, \alpha_k} = \{\mathbf{p} \in [0,1]^d : \ \mathbf{A}\mathbf{p} \leq \boldsymbol{\alpha}_{-k}, \ \mathbf{b}^T\mathbf{p} = \alpha_k, \ \mathbf{D}\mathbf{p} \leq \boldsymbol{\delta}\}.$$

- *Utility function* $\mathcal{U}^{\boldsymbol{\alpha}_{-k},\alpha_k}(\mathbf{p})$, can be written as

$$\mathcal{U}^{\boldsymbol{\alpha}_{-k},\alpha_k}(\mathbf{p}) = \frac{1}{\alpha_k}\mathbf{c}^T\mathbf{p},$$

where $\mathbf{c} = (p^{y=1}_{x_1\dots x_d} \cdot p_{x_1\dots x_d})^T$.

Theorem 5.8.1 and Lemma 5.8.2 show that the problem of maximizing precision (or equivalently, other metrics, see Theorem 5.3.1) can be solved through a linear program when the selection sizes sizes $(\boldsymbol{\alpha}_{-k}, \alpha_k)$ are given constants. This shows that the utility maximization problem in general form can be written as:

$$\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k) = \max_{\mathbf{p}\in C_{\boldsymbol{\alpha}_{-k},\alpha_k}\cap C_f} \frac{1}{\alpha_k}\mathbf{c}^T\mathbf{p}, \tag{5.7}$$

where

$$\begin{aligned} C_{\boldsymbol{\alpha}_{-k},\alpha_k} &= \{\mathbf{p} \in [0,1]^d \ : \ \mathbf{A}\mathbf{p} \le \boldsymbol{\alpha}_{-k},\ \mathbf{b}^T\mathbf{p} = \alpha_k,\ \mathbf{D}\mathbf{p} \le \boldsymbol{\delta}\},\\ C_f &= \{\mathbf{p} \in [0,1]^d \ : \ \mathbf{F}\mathbf{p} = \mathbf{0}\}. \end{aligned}$$

5.8.1 Utility Maximization via Linear Programming

In this section, we formally justify that maximizing precision under budget constraints and fairness constraints can be done via linear programming. The following theorem shows how to do that with only budget constraints:

Theorem 5.8.1 (Utility maximization as a linear program)**.** *Let, for all* $i \in \{1, \cdots, k\}$ *and all* $x_1 \dots x_{d_i}$,

$$p^{(i|i-1)}_{x_1\dots x_{d_i}} = \begin{cases} \tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}} / \tilde{p}^{(i-1|i-2)}_{x_1\dots x_{d_{i-1}}}, & \text{if } \tilde{p}^{(i-1|i-2)}_{x_1\dots x_{d_{i-1}}} \neq 0 \\ 0, & \text{otherwise}, \end{cases}$$

where the variables $\tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}}$ *are solutions of the linear program*

$$\begin{aligned} \max_{\tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}}} \quad & \frac{1}{\alpha_k} \sum_{x_1\dots x_d} p^{y=1}_{x_1\dots x_d} \cdot p_{x_1\dots x_d} \cdot \tilde{p}^{(k|k-1)}_{x_1\dots x_{d_k}} \\ \text{s.t.} \quad & \sum_{x_1\dots x_d} p_{x_1\dots x_d} \cdot \tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}} \le \alpha_i,\ i < k, \\ & \sum_{x_1\dots x_d} p_{x_1\dots x_d} \cdot \tilde{p}^{(k|k-1)}_{x_1\dots x_{d_k}} = \alpha_k, \\ & 0 \le \tilde{p}^{(1|0)}_{x_1\dots x_{d_1}} \le 1, \\ & 0 \le \tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}} \le \tilde{p}^{(i-1|i-2)}_{x_1\dots x_{d_{i-1}}}, \qquad 1 < i \le k. \end{aligned} \tag{5.8}$$

Then the $p^{(i|i-1)}_{x_1\dots x_{d_i}}$ *are solutions of* (5.3) *without any fairness constraint.*

Proof. Using (5.4)–(5.5), we can rewrite problem (5.3) as

$$\begin{aligned} \max_{p^{(i|i-1)}_{x_1\dots x_{d_i}}} \quad & \frac{1}{\alpha_k} \sum_{x_1\dots x_d} p^{y=1}_{x_1\dots x_d} \cdot p_{x_1\dots x_d} \cdot \prod_{i=1}^{k} p^{(i|i-1)}_{x_1\dots x_{d_i}} \\ \text{s.t.} \quad & \sum_{x_1\dots x_d} p_{x_1\dots x_d} \cdot \prod_{j=1}^{i} p^{(j|j-1)}_{x_1\dots x_{d_j}} \le \alpha_i, \quad i < k, \\ & \sum_{x_1\dots x_d} p_{x_1\dots x_d} \cdot \prod_{j=1}^{k} p^{(j|j-1)}_{x_1\dots x_{d_j}} = \alpha_k, \\ & 0 \le p^{(i|i-1)}_{x_1\dots x_{d_i}} \le 1, \quad 1 \le i \le k. \end{aligned} \tag{5.9}$$

Let us define the new variables

$$\tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}} = \prod_{j=1}^{i} p^{(j|j-1)}_{x_1\dots x_{d_j}}. \tag{5.10}$$

By substitution, we get the linear program (5.8). Hence, assuming that $\tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}}$ are solutions of (5.8), any $p^{(i|i-1)}_{x_1\dots x_{d_i}}$ such that (5.10) is satisfied (as is the case for the $p^{(i|i-1)}_{x_1\dots x_{d_i}}$ defined in the theorem) is a solution of (5.3). □

In the following lemma, we then show that fairness constraints can also be written as linear homogeneous equations in terms of the transformed variables $\mathbf{p} = (\tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}})^T$.

Lemma 5.8.2 (Linearity of fairness constraints)**.** *For both local and global fairness, and for both EO and DP, there exists a matrix* $\mathbf{F}$ *such that the fairness constraint can be expressed as*

$$\mathbf{Fp} = \mathbf{0}.$$

Proof. We present the proof for demographic parity, the idea is the same for equal opportunity. Let us consider the fairness constraint corresponding to stage i, $1 \le i \le k$:

$$\mathrm{P}(\hat{Y}_i = 1 | G = A) = \mathrm{P}(\hat{Y}_i = 1 | G = B).$$

By expanding the left side, we obtain:

$$\mathrm{P}(\hat{Y}_i = 1 | G = g) = \frac{\sum_{x_i} \tilde{p}^{(i|i-1)}_{x_1\dots x_{d_i}} \cdot p_{x_1\dots g\dots x_d}}{\sum_{x_i} p_{x_1\dots g\dots x_d}}.$$

Recall that $p_{x_1\dots g\dots x_d}$ is a fixed parameter and not a decision variable. Thus, for both local and global fairness, the fairness constraint (equality of the probabilities for $G = A$ and $G = B$) can be represented in the form $\mathbf{Fp} = \mathbf{0}$ for an appropriate $\mathbf{F}$ simply by moving all terms on the left side of the equality. □

5.8.2 Utility Maximization in Limit of $n \to \infty$

In this section we provide an intuition on multistage selection process in limit of infinitely large n. In short, the optimal candidate selection problem with finite number of candidates n appears to be a k-stage stochastic optimization problem which is difficult to solve exactly. By letting the number of candidates n to be infinitely large allows us to reformulate the problem in a much simpler manner such that we are able to find an optimal selection probabilities easily.

To prove the statements in this section we will exploit the two following classical results from the probability theory, see (Rohatgi and Saleh, 2015).

Lemma 5.8.3 (Chebyshev-Bienaymé inequality)**.** *Let X be a random variable, then*

$$\mathrm{P}\left(|X - \mathrm{E}(X)| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Lemma 5.8.4 (Properties of convergence in probability)**.** *Let X_n and Y_n be sequences of random variables.*

1. *If $X_n \xrightarrow{P} X$ and a is a constant, then $aX_n \xrightarrow{P} aX$.*
2. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.*
3. *If $X_n \xrightarrow{P} X$, then $1/X_n \xrightarrow{P} 1/X$.*
4. *If $X_n \xrightarrow{L} X$ and $|X_n - Y_n| \xrightarrow{P} 0$, then $Y_n \xrightarrow{L} X$.*

The following lemma gives us the limit on the proportion of selected at stage i candidates as $n \to \infty$.

Lemma 5.8.5. *Let by $n^{(i)}_{x_1 \dots x_{d_i}}$ denote the number of candidates having features $x_1 \dots x_{d_i}$ that are selected at the stage i, then for $n \to \infty$:*

$$\frac{n^{(i)}_{x_1 \dots x_{d_i}}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_i}} \prod_{j=1}^{i} p^{(j|j-1)}_{x_1 \dots x_{d_j}}.$$

Before proving the above lemma let us define the budget $B_n(i)$ at the stage i as $B_n(i) = \frac{1}{n} \sum_{x_1 \dots x_{d_i}} n^{(i)}_{x_1 \dots x_{d_i}}$. Then using property 2 from Lemma 5.8.4 and Lemma 5.8.5 we obtain that $B_n(i) \xrightarrow{L} \sum_{x_1 \dots x_{d_i}} p_{x_1 \dots x_{d_i}} \prod_{j=1}^{i} p^{(j|j-1)}_{x_1 \dots x_{d_j}}$.

The precision is the proportion of good candidates among selected then using Lemma 5.8.4, the argument in Lemma 5.8.5 and fact the the final stage selection size is fixed to α_k, the precision converges in law to $\frac{1}{\alpha_k} \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} p^{y=1}_{x_1 \dots x_d} \prod_{j=1}^{k} p^{(j|j-1)}_{x_1 \dots x_{d_j}}$ as n goes to infinity. Hence, the equations (5.4)–(5.5) hold as the number of candidates $n \to \infty$. Let us prove Lemma 5.8.5 by induction on stage number i.

Proof. Base of induction. Before we do any selection:

$$n^{(0)}_{x_1\dots x_{d_1}} \sim \mathrm{Bin}\left(n, p_{x_1\dots x_{d_1}}\right),$$

then using Chebyshev–Bienaymé inequality:

$$P\left(\left|\frac{n^{(0)}_{x_1\dots x_{d_1}}}{n} - \frac{n\cdot p_{x_1\dots x_{d_1}}}{n}\right| \geq \varepsilon\right) \leq \frac{np_{x_1\dots x_{d_1}}(1-p_{x_1\dots x_{d_1}})}{\varepsilon^2 n^2} \leq \frac{1}{\varepsilon^2 n}$$

hence $\frac{n^{(0)}_{x_1\dots x_{d_1}}}{n} \xrightarrow{P} p_{x_1\dots x_{d_1}}$, $n\to\infty$.

After, when we perform the selection at the first stage:

$$n^{(1)}_{x_1\dots x_{d_1}} | n^{(0)}_{x_1\dots x_{d_1}} \sim \mathrm{Bin}\left(n^{(0)}_{x_1\dots x_{d_1}}, p^{(1|0)}_{x_1\dots x_{d_1}}\right).$$

$$P\left(\left|\frac{n^{(1)}_{x_1\dots x_{d_1}}}{n} - \frac{n^{(0)}_{x_1\dots x_{d_1}}}{n}p^{(1|0)}_{x_1\dots x_{d_1}}\right| \geq \varepsilon\right) \leq$$
$$\leq \frac{n^{(0)}_{x_1\dots x_{d_1}}p^{(1|0)}_{x_1\dots x_{d_1}}(1-p^{(1|0)}_{x_1\dots x_{d_1}})}{\varepsilon^2\cdot n^2} \leq \frac{1}{\varepsilon^2 n},$$

so using properties 1 and 4 from Lemma 5.8.4, we obtain that $\frac{n^{(1)}_{x_1\dots x_{d_1}}}{n} \xrightarrow{L} p_{x_1\dots x_{d_1}}p^{(1|0)}_{x_1\dots x_{d_1}}$.

Induction Step. Let us consider the stage $i+1$. By the assumption of induction:

$$\frac{n^{(i)}_{x_1\dots x_{d_i}}}{n} \xrightarrow{L} p_{x_1\dots x_{d_i}}\prod_{j=1}^{i} p^{(j|j-1)}_{x_1\dots x_{d_j}}.$$

After making the selection at the stage i we observe the new features $d_i+1,\dots,d_{i+1}$, so

$$n^{(i)}_{x_1\dots x_{d_{i+1}}} | n^{(i)}_{x_1\dots x_{d_i}} \sim \mathrm{Bin}\left(n^{(i)}_{x_1\dots x_{d_i}}, p_{x_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i}}\right),$$

where $p_{X_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i}} := \mathrm{P}(x_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i})$. Then

$$P\left(\left|\frac{n^{(i)}_{x_1\dots x_{d_{i+1}}}}{n} - \frac{n^{(i)}_{x_1\dots x_{d_i}}}{n}p_{x_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i}}\right| \geq \varepsilon\right) \leq$$
$$\leq \frac{n^{(i)}_{x_1\dots x_{d_i}}p_{x_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i}}(1-p_{x_{d_i+1}\dots x_{d_{i+1}}|x_1\dots x_{d_i}})}{\varepsilon^2\cdot n^2}$$
$$\leq \frac{1}{\varepsilon^2 n},$$

hence $\frac{n^{(i)}_{x_1\dots x_{d_{i+1}}}}{n} \xrightarrow{L} p_{x_1\dots x_{d_{i+1}}}\prod_{j=1}^{i} p^{(j|j-1)}_{x_1\dots x_{d_j}}$.

When we perform the selection at the stage $i+1$:

$$n^{(i+1)}_{x_1\dots x_{d_{i+1}}} | n^{(i)}_{x_1\dots x_{d_{i+1}}} \sim \text{Bin}\left(n^{(i)}_{x_1\dots x_{d_{i+1}}}, p^{(i+1|i)}_{x_1\dots x_{d_i+1}}\right)$$

then again using Chebyshev-Bienaymé inequality for $\frac{n^{(i+1)}_{x_1\dots x_{d_{i+1}}}}{n}$:

$$P\left(\left|\frac{n^{(i+1)}_{x_1\dots x_{d_{i+1}}}}{n} - \frac{n^{(i)}_{x_1\dots x_{d_{i+1}}} \cdot p^{(i+1|i)}_{x_1\dots x_{d_{i+1}}}}{n}\right| \geq \varepsilon\right) \leq$$
$$\leq \frac{n^{(i)}_{x_1\dots x_{d_{i+1}}} p^{(i+1|i)}_{x_1\dots x_{d_{i+1}}} (1 - p^{(i+1|i)}_{x_1\dots x_{d_{i+1}}})}{\varepsilon^2 \cdot n^2} \leq \frac{1}{\varepsilon^2 n},$$

so $\left|\frac{n^{(i+1)}_{x_1\dots x_{d_{i+1}}}}{n} - \frac{n^{(i)}_{x_1\dots x_{d_{i+1}}} \cdot p^{(i+1|i)}_{x_1\dots x_{d_{i+1}}}}{n}\right| \xrightarrow{P} 0,\ n \to \infty$ and finally:

$$\frac{n^{(i+1)}_{x_1\dots x_{d_{i+1}}}}{n} \xrightarrow{L} p_{x_1\dots x_{d_{i+1}}} \prod_{j=1}^{i+1} p^{(j|j-1)}_{x_1\dots x_{d_j}}.$$

□

5.8.3 Proof of Theorem 5.3.1

We prove the equivalence only for accuracy (denoted ACC); the proof for other metrics follows the same idea. By expanding ACC, we obtain:

$$\begin{aligned} ACC &= \mathrm{P}(\hat{Y}_k = y) = \mathrm{P}(\hat{Y}_k = 1, Y = 1) + \mathrm{P}(\hat{Y}_k = 0, Y = 0) \\ &= \mathrm{P}(\hat{Y}_k = 1, Y = 1) + \big(\mathrm{P}(Y = 0) - \mathrm{P}(\hat{Y}_k = 1, Y = 0)\big) \\ &= 2 \cdot \mathrm{P}(\hat{Y}_k = 1, Y = 1) + \mathrm{P}(Y = 0) - \mathrm{P}(\hat{Y}_k = 1) \\ &= 2 \cdot \mathrm{P}(\hat{Y}_k = 1) \cdot \mathrm{P}(Y = 1 | \hat{Y}_k = 1) \\ &\quad + \mathrm{P}(Y = 0) - \mathrm{P}(\hat{Y}_k = 1). \end{aligned}$$

Since the terms $\mathrm{P}(\hat{Y}_k = 1)$ and $\mathrm{P}(Y = 0)$ are constant, maximization of precision is equivalent to maximization of ACC.

5.8.4 Proof of Theorem 5.4.6

(1) We present the proof only for demographic parity, the proof for equal opportunity follows the same idea. We do the proof by induction. First consider a 2-stage selection algorithm $\hat{Y} = (\hat{Y}_1, \hat{Y}_2)$. By considering the following quantity:

$$\mathrm{P}(\hat{Y}_2 = 1 | \hat{Y}_1 = 1, G = g) = \frac{\mathrm{P}(\hat{Y}_2 = 1, \hat{Y}_1 = 1, G = g)}{\mathrm{P}(\hat{Y}_1 = 1, G = g)}$$
$$= \frac{\mathrm{P}(\hat{Y}_2 = 1, \hat{Y}_1 = 1, G = g) + \overbrace{\mathrm{P}(\hat{Y}_2 = 1, \hat{Y}_1 = 0, G = g)}^{=0}}{\mathrm{P}(\hat{Y}_1 = 1, G = g)}$$

$$= \frac{\mathrm{P}(\hat{Y}_2 = 1, G = g)}{\mathrm{P}(\hat{Y}_1 = 1|G = g)\,\mathrm{P}(G = g)} = \frac{\mathrm{P}(\hat{Y}_2 = 1|G = g)}{\mathrm{P}(\hat{Y}_1 = 1|G = g)},$$

the fairness constraint for LF1 at the second stage is:

$$\frac{\mathrm{P}(\hat{Y}_2 = 1|G = A)}{\mathrm{P}(\hat{Y}_1 = 1|G = B)} = \frac{\mathrm{P}(\hat{Y}_2 = 1|G = B)}{\mathrm{P}(\hat{Y}_1 = 1|G = B)}.$$

Since we impose fairness at the first stage, then $\mathrm{P}(\hat{Y}_1 = 1|G = A) = \mathrm{P}(\hat{Y}_1 = 1|G = B)$, so the condition above is equivalent to

$$\mathrm{P}(\hat{Y}_2 = 1|G = A) = \mathrm{P}(\hat{Y}_2 = 1|G = B),$$

that is exactly the second constraint for the LF2 notion. Thus, the statement is true for a 2 stage selection algorithm.

Second, assuming that the statement is true for $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_i)$, $i > 2$, let us consider the $(i+1)$-stage selection algorithm. By analogy, considering the quantity

$$\mathrm{P}(\hat{Y}_{i+1} = 1|\hat{Y}_i = 1, G = g) = \frac{\mathrm{P}(\hat{Y}_{i+1} = 1, \hat{Y}_i = 1, G = g)}{\mathrm{P}(\hat{Y}_i = 1, G = g)}$$
$$= \frac{\mathrm{P}(\hat{Y}_{i+1} = 1, G = g)}{\mathrm{P}(\hat{Y}_i = 1|G = g)\,\mathrm{P}(G = g)} = \frac{\mathrm{P}(\hat{Y}_{i+1} = 1|G = g)}{\mathrm{P}(\hat{Y}_i = 1|G = g)}$$

we obtain that the fairness constraint for LF1 at the stage $i+1$ is

$$\frac{\mathrm{P}(\hat{Y}_{i+1} = 1|G = A)}{\mathrm{P}(\hat{Y}_i = 1|G = A)} = \frac{\mathrm{P}(\hat{Y}_{i+1} = 1|G = B)}{\mathrm{P}(\hat{Y}_i = 1|G = B)}.$$

As $\mathrm{P}(\hat{Y}_i = 1|G = A) = \mathrm{P}(\hat{Y}_i = 1|G = B)$ by the assumption of induction, we have $\mathrm{P}(\hat{Y}_{i+1} = 1|G = A) = \mathrm{P}(\hat{Y}_{i+1} = 1|G = B)$.

The point (2) follows from the definitions of LF2 and GF, since the problem GF is less constrained than LF2.

5.8.5 Proof of Theorem 5.5.1

(1) For a given $\boldsymbol{\alpha}_{-k}$, assume that the program attains its maximum $\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ at point $\mathbf{p}$ and let $\boldsymbol{\alpha}'_{-k} \geq \boldsymbol{\alpha}_{-k}$, where $\geq$ is meant component-wise.

By setting $\mathbf{p}' = \mathbf{p}$, we obtain that $\mathbf{p}' \in C_{\boldsymbol{\alpha}'_{-k}, \alpha_k}$ and thus:

$$\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k) = \mathcal{U}^{\boldsymbol{\alpha}_{-k}, \alpha_k}(\mathbf{p}) = \mathcal{U}^{\boldsymbol{\alpha}'_{-k}, \alpha_k}(\mathbf{p}') \leq \mathcal{U}(\boldsymbol{\alpha}'_{-k}, \alpha_k).$$

Let us consider a problem, when $\boldsymbol{\alpha}_{-k} = \boldsymbol{\alpha}'_{-k}$, it attains its maximum $\mathcal{U}(\boldsymbol{\alpha}'_{-k}, \alpha_k)$ at the point $\mathbf{p}'$. Analogously, let for the second problem $\boldsymbol{\alpha}_{-k} = \boldsymbol{\alpha}''_{-k}$, and it attains its maximum $\mathcal{U}(\boldsymbol{\alpha}''_{-k}, \alpha_k)$ at the point $\mathbf{p}''$. Then for any $\lambda \in [0, 1]$ the point $\lambda\mathbf{p}' + (1-\lambda)\mathbf{p}'' \in C_{\lambda\boldsymbol{\alpha}'_{-k} + (1-\lambda)\boldsymbol{\alpha}''_{-k}, \alpha_k}$ and:

$$\lambda\mathcal{U}(\boldsymbol{\alpha}', \alpha_k) + (1-\lambda)\mathcal{U}(\boldsymbol{\alpha}''_{-k}, \alpha_k) = \lambda\frac{1}{\alpha_k}\mathbf{c}^T\mathbf{p}'$$
$$+ (1-\lambda)\frac{1}{\alpha_k}\mathbf{c}^T\mathbf{p}''$$

$$
\begin{aligned}
&= \frac{1}{\alpha_k}\mathbf{c}^T(\lambda\mathbf{p}' + (1-\lambda)\mathbf{p}'') \\
&= \mathcal{U}^{\lambda\boldsymbol{\alpha}'_{-k}+(1-\lambda)\boldsymbol{\alpha}''_{-k},\alpha_k}(\lambda\mathbf{p}' + (1-\lambda)\mathbf{p}'') \\
&\leq \mathcal{U}(\lambda\boldsymbol{\alpha}'_{-k} + (1-\lambda)\boldsymbol{\alpha}''_{-k}, \alpha_k).
\end{aligned}
$$

(2) For a given α_k, assume that the program attains its maximum $\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ at point $\mathbf{p}$. Let $\alpha'_k = \alpha_k/\gamma$, where $\gamma \in [1, +\infty)$. Then consider $\mathbf{p}' = \mathbf{p}/\gamma$. We have $\mathbf{p}' \in C_{\boldsymbol{\alpha}_{-k},\alpha'_k}$ and $\mathbf{p}' \in C_f$ and:

$$
\begin{aligned}
\mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha_k) &= \mathcal{U}^{\boldsymbol{\alpha}_{-k},\alpha_k}(\mathbf{p}) = \frac{1}{\alpha_k}\mathbf{c}^T\mathbf{p} \\
&= \frac{1}{\alpha_k/\gamma}\mathbf{c}^T\mathbf{p}/\gamma = \mathcal{U}^{\boldsymbol{\alpha}_{-k},\alpha'_k}(\mathbf{p}') \leq \mathcal{U}(\boldsymbol{\alpha}_{-k}, \alpha'_k).
\end{aligned}
$$

5.8.6 Proof of Theorem 5.5.3

Let us consider the trivial locally fair algorithm. It selects candidates randomly with probability α_1 at the first stage and with probability α_i/α_{i-1}, $\forall 1 < i \leq k$. The utility of such random algorithm is equal to $\mathcal{U}^{random}(\boldsymbol{\alpha}_{-k}, \alpha_k) = \mathrm{P}(Y = 1)$. It is obvious by definition that

$$
\mathcal{U}^{random}(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \mathcal{U}^{\mathrm{LF}}(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k).
$$

To obtain an upper bound of $\mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k)$ we suppose that all features are available for the selection, meaning that $\alpha_i = 1$, $\forall i < k$. Then $\mathcal{U}^{\mathrm{GF}}(\boldsymbol{\alpha}_{-k}, \alpha_k) \leq \mathcal{U}^{\mathrm{un}}(\alpha_1 = 1, \ldots, \alpha_{k-1} = 1, \alpha_k) \leq \min(\mathrm{P}(Y = 1)/\alpha_k, 1)$. Thus,

$$
\begin{aligned}
PoLF(\boldsymbol{\alpha}_{-k}, \alpha_k) &\leq \frac{\min(\mathrm{P}(Y = 1)/\alpha_k, 1)}{\mathrm{P}(Y = 1)} \\
&= \min\left(\frac{1}{\mathrm{P}(Y = 1)}, \frac{1}{\alpha_k}\right).
\end{aligned}
$$

5.9 Additional Experimental Results

In this section, we provide additional experimental results that support the claims in the chapter—in particular by reproducing the curves in the chapter for different datasets and different fairness metrics. Below, we describe the data preparation for our experiments in more detail.

`Adult dataset`

In the `Adult` dataset from the UCI repository (Dua and Graff, 2017) there are 48842 candidates, each described by 14 features. The label *income* denotes if candidate gains more than 50.000 dollars annually. For all our experiments we binarize and leave only the 6 following features: *sex* (is male), *age* (is above 35), *native-country* (from the EU or US), *education* (has Bachelor or Master degree), *hours-per-week* (works more than 35 hours per week) and *relationship* (is married).

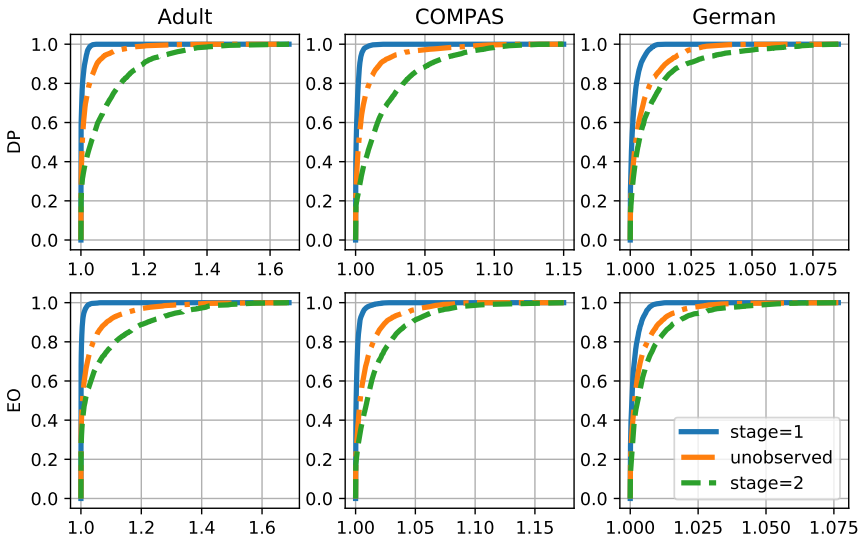


Figure 5.6 – The empirical CDF of $PoLF$ for $\alpha_2 = 0.3$.

COMPAS Dataset

The COMPAS dataset (Larson et al., 2016) is a dataset that is used to train the COMPAS algorithm. It contains information about prisoners, such as their name, gender, age, race, start of the sentence, end of the sentence, charge description etc. and a label *y=recidivism*, that is $Y = 1$ if person is likely to reoffend and 0, otherwise.

We prepare original COMPAS dataset for our means by selecting statistics only for Caucasian and African-American defendants, leaving only 6 features and binarizing them. The features that we use are following: *sex* (is male), *young* (younger than 25), *old* (older than 45), *long sentence* (sentence was longer than 30 days), *drugs* (the arrest was due to selling or possessing drugs), *race* (is Caucasian).

German Dataset

The German Credit data from (Dua and Graff, 2017) contains information about applicants for credit. As with other datasets, we binarize feature values. The label feature *y=returns* shows if applicant payed for his loan, and we binarize and use 6 features: *job* (is employed), *housing* (owns house), *sex* (is male) *savings* (greater than 500 DM), *credit history* (all credits payed back duly), *age* (older than 50).

5.9.1 The Price of Local Fairness ($PoLF$)

Fig. 5.6 displays the CDF of $PoLF$ as in Fig. 5.3 but including the results for EO as well.

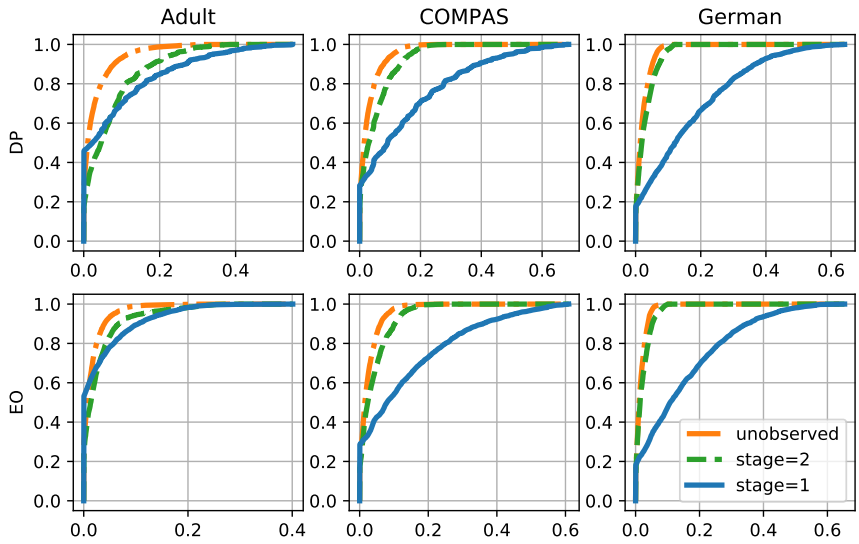


Figure 5.7 – The empirical CDF of $VoLF$ for $\alpha_2 = 0.3$.

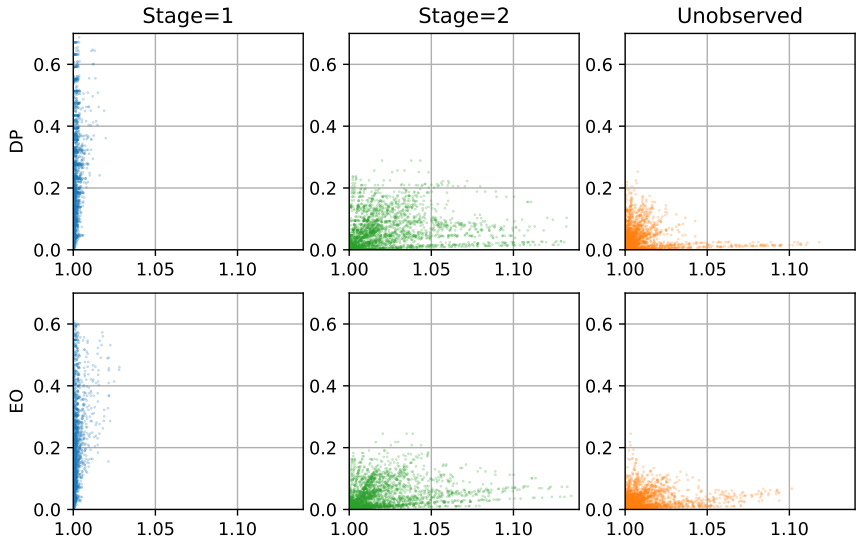


Figure 5.8 – Joint distribution of $VoLF$ and $PoLF$ for `COMPAS` dataset and $\alpha_2 = 0.3$.

5.9.2 The Violation of Local Fairness ($VoLF$)

Fig. 5.7 displays the CDF of $VoLF$ as in Figure 5.7 but including the results for EO as well.

5.9.3 $VoLF$ vs $PoLF$ for Various Datasets

Fig. 5.8 and Fig. 5.9 display the joint distribution of $VoLF$ (y-axis) and $PoLF$ (x-axis) as in Fig. 5.5 but for the other two datasets: `COMPAS` and `German` respectively.

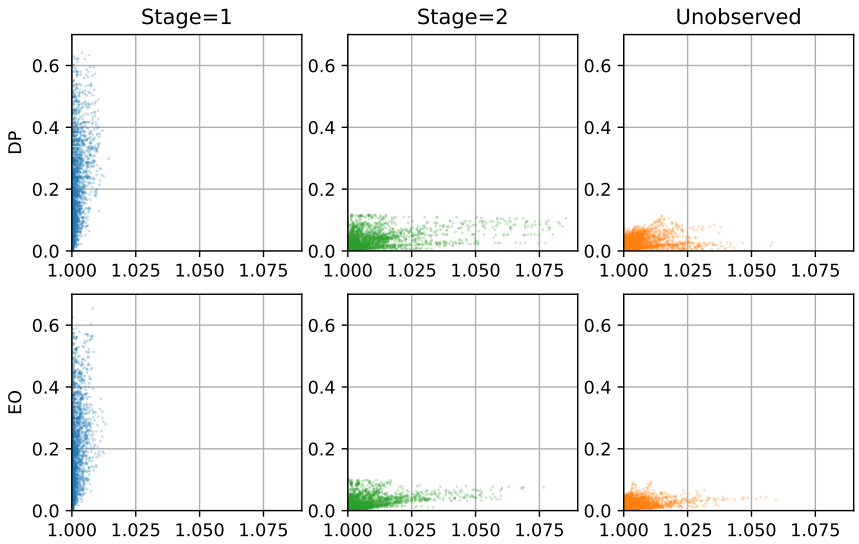


Figure 5.9 – Joint distribution of $VoLF$ and $PoLF$ for `German` dataset and $\alpha_2 = 0.3$.

5.9.4 $PoLF$ of 3-stage algorithm

In this subsection we present the results for $PoLF$ of 3-stage selection algorithm. The procedure to calculate the $PoLF$ values is similar to the one we used for two-stage algorithm. We suppose that we observe only one feature at every stage, we suppose that the one of the rest features is a sensitive G and consider the cases when it is observed at first, second or third stage of selection process. We calculate the value of $PoLF$ for every possible 4 feature combinations (three decision variables and one being sensitive) out of 6 and for every discretized value of α_1 and α_2, such that $\alpha_3 = 0.3 \leq \alpha_2 \leq \alpha_1$. Fig. 5.10 displays the empirical CDFs of $PoLF$ for 3-stage algorithm. The observations are the same as in two-stage case: later the sensitive attribute G is revealed, larger is the price of imposing local constraints.

In Fig. 5.11 we show the joint distribution of $PoLF$ when the G (shown on top of each subfigure) is observed at the first stage (we call it $PoLF_1$) and $PoLF$ when the G is observed at the second stage ($PoLF_2$) for `Adult` dataset. We observe that the value of $PoLF_1$ is sufficiently smaller than the corresponding value of $PoLF_2$.

In Fig. 5.12 we show the joint distribution of $PoLF$ when the G is observed at the second stage (we call it $PoLF_2$) and $PoLF$ when the G is observed at the third stage ($PoLF_3$) for `Adult` dataset. We again observe that the value of $PoLF_2$ is smaller than the corresponding value of $PoLF_3$, since the most of the points lie above the diagonal line which is marked as dashed red line.

In Fig. 5.13-5.16 we display the joint distributions in the same manner as in Fig. 5.11 and Fig. 5.12 but for `COMPAS` and `German Credit` datasets.

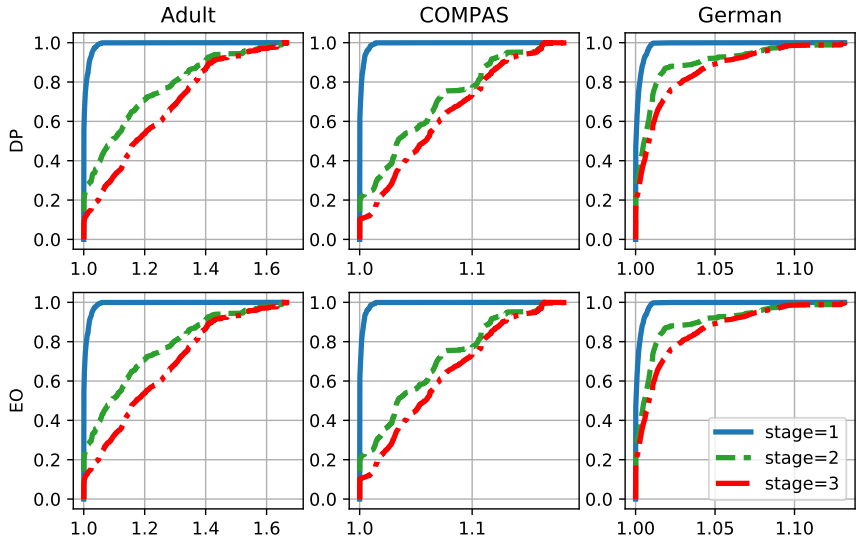


Figure 5.10 – The empirical CDF of $PoLF$ of 3-stage algorithm for $\alpha_3 = 0.3$.

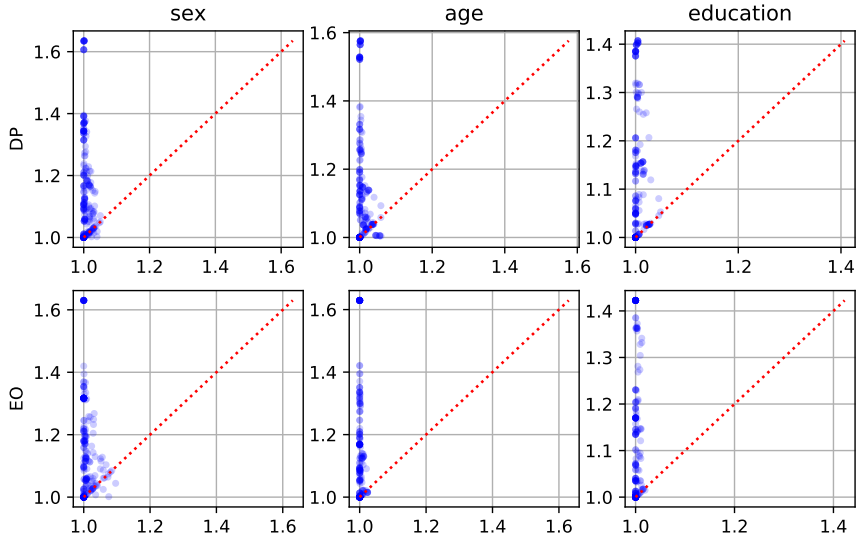


Figure 5.11 – Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for `Adult` dataset and $\alpha_3 = 0.3$.

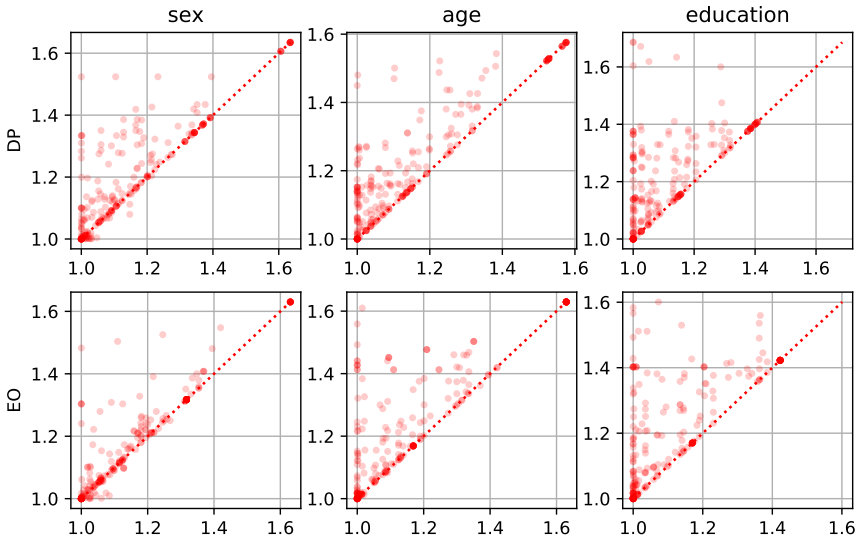


Figure 5.12 – Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for `Adult` dataset and $\alpha_3 = 0.3$.

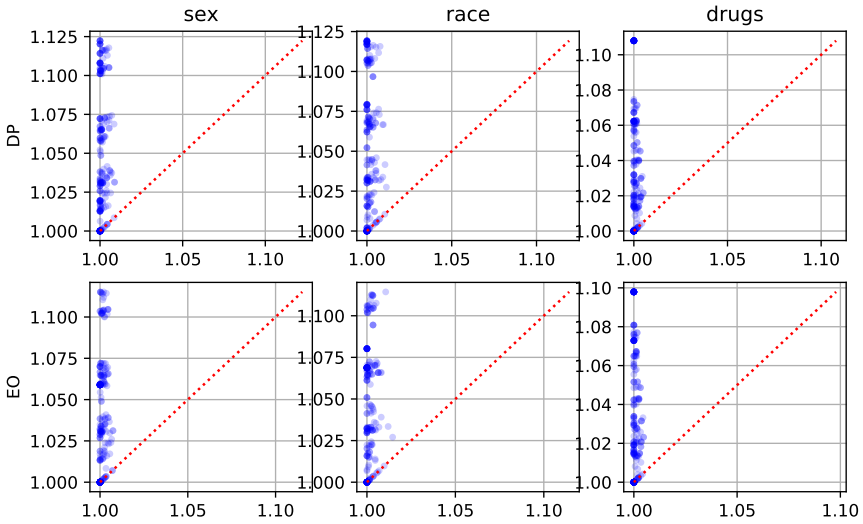


Figure 5.13 – Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for `COMPAS` dataset and $\alpha_3 = 0.3$.

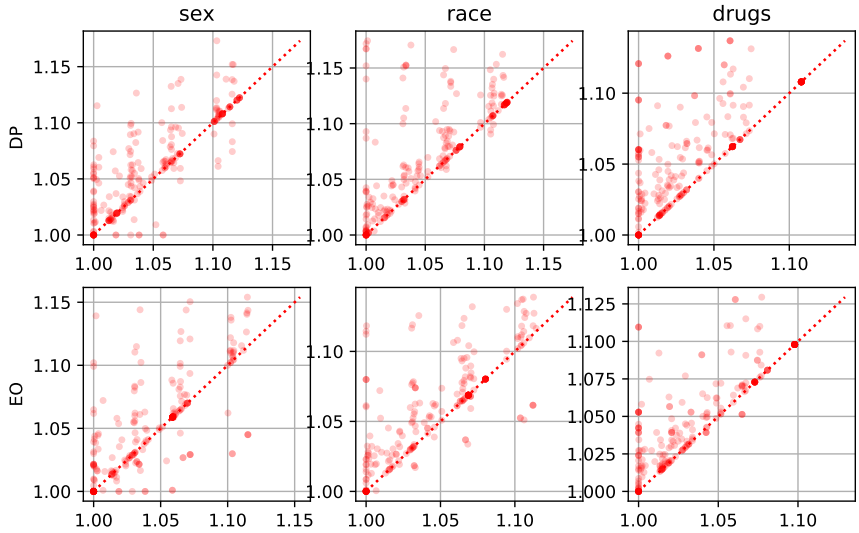


Figure 5.14 – Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for `COMPAS` dataset and $\alpha_3 = 0.3$.

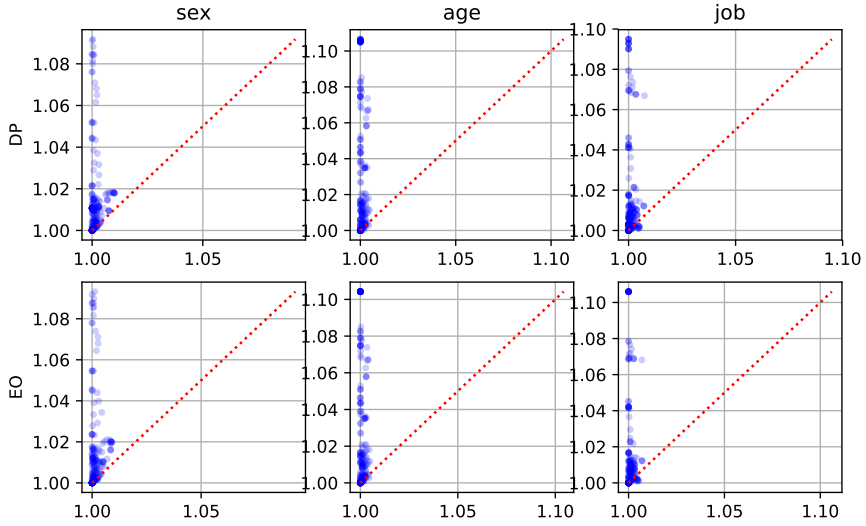


Figure 5.15 – Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for `German Credit` dataset and $\alpha_3 = 0.3$.

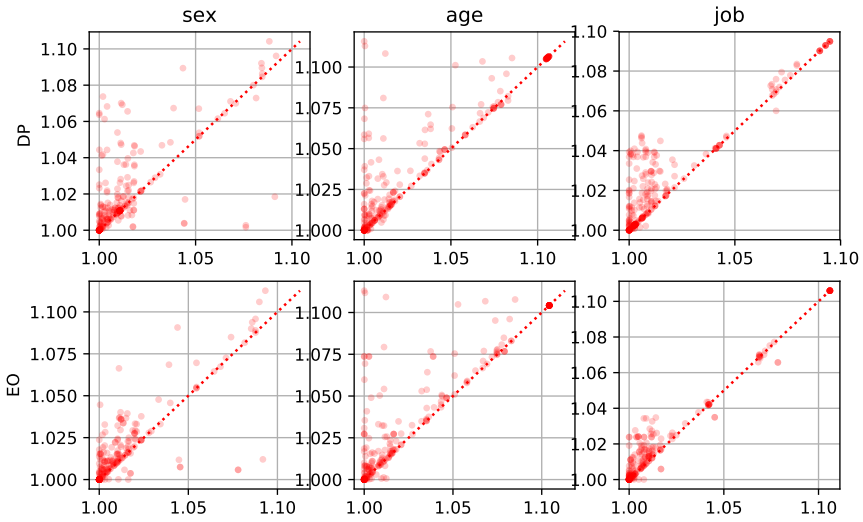


Figure 5.16 – Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for `German Credit` dataset and $\alpha_3 = 0.3$.

Chapter 6

General Conclusion and Discussion

In this thesis, we studied the problem of fairness in selection procedures. We proposed three models that capture different aspects of the selection problem. First, in Chapter 3, we proposed a model of selection with *differential variance*. We showed that the differential variance results in discrimination; hence, we studied how different group fairness mechanisms affect the selection utility in this model. Second, in Chapter 4, based on our model of selection with differential variance from Chapter 3, we proposed a model of selection with *strategic individuals* that obtain qualities that maximize the individual utility of taking part in selection procedures. This model of selection results in a population game that attains a unique equilibrium which we characterized. Finally, in Chapter 5, we proposed a model of multistage selection, and introduced two fairness notions for multistage setting—*local* and *global* fairness. We studied both theoretically and empirically the *price of local fairness* which is the ratio of selection utilities for the globally fair to that of the locally fair algorithms. We showed bounds on the price of local fairness, and we identified how the access to sensitive information (e.g., race or gender) affects it.

6.1 Implications of the Thesis

Our work can be used for designing and evaluating selection policies. The high-level implications of our results are as follows:

(i) *The effect of noise in estimating candidates' quality on the resulting discrimination is non-negligible* and must be taken into account in addition to the implicit bias. In Chapter 3 and Chapter 4, we show that even in the absence of implicit bias but in the presence of differential variance, one of the groups of candidates (high-noise or low-noise) is always underrepresented.

(ii) *Policy-makers must evaluate what information is available to decision-makers.* In our model in Chapter 3 we show that the resulting discrimination is opposite for the Bayesian decision-maker (i.e., for which the distribution of qualities is available) compared to that of the group-oblivious decision-maker (i.e., for which

only the estimates of qualities are available). If the decision-maker is group-oblivious, then the low-noise candidates will be underrepresented. In contrast, if the decision-maker is Bayesian, the high-noise candidates will be underrepresented. In Section 5, we also study how the access to sensitive attributes affects the price of local fairness. We empirically show that in most cases the earlier the sensitive attribute is observed in the selection pipeline, the lower the price of local fairness.

(iii) *It is important to consider the strategic nature of individuals.* We show that, in the non-strategic setting (Chapter 3), the high-noise candidates are always underrepresented if the decision-maker is Bayesian. This contrasts with the results in the strategic setting (Chapter 4): we prove that if the reward for being selected is large, then the low-noise candidates will be underrepresented by the Bayesian decision-maker. In addition, we show that the Bayesian decision-maker is not optimal anymore in the selection with strategic individuals. Hence, policy-makers must evaluate if individuals can behave strategically.

(iv) *Discrimination is nuanced and depends on the interactions among different parameters of the selection problem.* For example, in Chapter 4, we show that when the cost-of-effort coefficients are group-dependent, the effects of differential variance on selection are negligible if the reward is large enough; this is not true for small rewards. Our work, thus, can be used by policy-makers to estimate the effects of different parameters on the selection outcome.

6.2 Perspectives of the Thesis

Our work can be extended in multiple ways. We provided each chapter of our contributions with a corresponding discussion. This section presents some of the main directions for future work.

Matching problems with noisy preferences The selection problem can be seen as a special case of a so-called *college admission problem* (Gale and Shapley, 1962). The college admission problem is a matching problem: there are multiple colleges (of a limited capacity) and multiple applicants. Both colleges and applicants have listed preferences. The aim of a college admission problem is to provide a stable matching of students to colleges. A matching is called *stable* if there exists no unmatched pair of student s and college c, where student s prefers college c and college c still has some unfilled place, or college c admitted a student that it prefers less than student s.

Some authors argue that college preferences can be noisy as colleges have only a little information about students (Chade et al., 2013). The noise can be of different magnitude for different groups of students, since decision-makers might be less familiar with students from some demographic groups. Preferences can also be biased since decision-makers might prefer some students to others due to stereotyping beliefs or other factors. The general study of a college admission problems with noisy and

biased preferences, hence, is an important direction for future work. Note that the selection problem we studied in Chapter 3 can be seen as a matching problem with a single college.

Long-term effects of affirmative actions In our work, we assume that the selection decision is made only once, and we were interested in studying only the intermediate effects of fairness mechanisms on the utility of the selection and/or group representation. In reality, decisions are usually made at multiple consecutive times and the earlier decisions might affects the performance of future generations. For example, due to different opportunities, candidates of a disadvantaged groups might have poor performance, however, providing opportunities to current generations by performing affirmative actions (even at the price of utility of selection), might lead to an increase of the performance of previously disadvantaged groups at future generations. This is due to the fact that opportunities can be inherited, e.g., children of educated parents would have a lower cost of getting an education since their parent could teach them or have enough income to provide preparatory courses. This brings the questions on the long-term effects of affirmative actions, and whether they must be sustained through generations or they can be lifted at some point in the future (Celis et al., 2021; Coate and Loury, 1993; Heidari and Kleinberg, 2021; Jehiel and Leduc, 2021). In our model, both in Chapter 3 and Chapter 4, we assume that decision-maker receives noisy estimates of candidates' qualities and the noise variance is constant. We can assume a bit different setting compared to one studied above: a *finite* number of candidates are available for selection at time $t = 1$, the decision-maker performs the selection and updates the variance of noise for groups depending on the respective selection size. This introduces a certain trade-off between exploitation (selecting candidates for which the expected quality is high) and exploration (learning better the candidates of previously high-noise group).

Fair online selection problem Most selection problems are performed in an online manner: individuals arrive sequentially (or in batches), and the decision-maker should momentarily decide to accept or reject an individual based on his attributes (which are an estimate of her quality). For example, in lending, the decision-maker studies the data of an applicant and decides whether they will repay their loans or will default. In this case, the decision-maker receives feedback for its decisions only from one side: whether the prediction is correct becomes known only for accepted applicants, e.g., the information about whether an applicant repays or defaults is known only for those who actually were accepted to get loans. Also, at the beginning of the learning process, only a little or no information is available, which may naturally lead to degradation of performance for minority groups due to less statistical data. The decision-making algorithm should find a proper balance between exploitation and exploration to guarantee a large cumulative reward; at the same time the algorithm must guarantee a sublinear growth of fairness and budget violations over time.

The online classification problem with partial feedback appears in the literature

under the name of an *apple tasting* problem (Helmbold et al., 2000). The standard assumption is to convert this problem to a contextual bandit problem by providing a reward matrix for each of the actions (classify as 1, or 0). In our approach, we could use the same transformation and convert the problem to a contextual bandit problem. We then could start with an assumption of infinite data where the reward maximization problem with fairness and budget constraints can be solved using linear programming as we show in Chapter 5. In order to apply linear programming to an online setting, we could estimate the parameters of the linear program (distribution parameters) using an upper confidence bound approach as in (Chen et al., 2018) or a Bayesian approach similar to one in (Saxena et al., 2020).

List of Publications

Emelianov, Vitalii, George Arvanitakis, Nicolas Gast, Krishna Gummadi, and Patrick Loiseau (2019). "The Price of Local Fairness in Multistage Selection". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (cit. on pp. 7, 14, 85).

Emelianov, Vitalii, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau (2022a). "On fair selection in the presence of implicit and differential variance". In: *Artificial Intelligence* 302, p. 103609 (cit. on pp. 6, 14, 16, 18).

— (2020). "On Fair Selection in the Presence of Implicit Variance". In: *Proceedings of the 21st ACM Conference on Economics and Computation* (cit. on pp. 6, 14, 16, 18, 25, 36).

Emelianov, Vitalii, Nicolas Gast, and Patrick Loiseau (2022b). "Fairness in Selection Problems with Strategic Candidates". In: *Proceedings of the 23rd ACM Conference on Economics and Computation* (cit. on pp. 7, 12, 14, 16, 50, 55, 68, 71).

Bibliography

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach (2018). "A reductions approach to fair classification". In: *International Conference on Machine Learning*. PMLR, pp. 60–69 (cit. on p. 11).

Ahmed, Ali, Mark Granberg, and Shantanu Khanna (2021). "Gender discrimination in hiring: An experimental reexamination of the Swedish case". In: *PLOS ONE* 16.1, pp. 1–15 (cit. on p. 4).

Aigner, Dennis J. and Glen G. Cain (1977). "Statistical Theories of Discrimination in Labor Markets". In: *Industrial and Labor Relations Review* 30.2, pp. 175–187 (cit. on pp. 16, 55).

Akerlof, George A. and Rachel E. Kranton (2000). "Economics and Identity*". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753 (cit. on p. 16).

Arrow, Kenneth J. (1973). "The theory of discrimination". In: *Discrimination in Labor Markets*. Ed. by Orley Ashenfelter and Albert Rees. Princeton University Press, pp. 3–33 (cit. on pp. 6, 15, 16, 54, 55).

Arunachaleswaran, Eshwar Ram, Sampath Kannan, Aaron Roth, and Juba Ziani (2021). "Pipeline Interventions". In: *ITCS* (cit. on p. 13).

Balcan, Maria-Florina, Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia (2019). "Envy-Free Classification". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (cit. on p. 16).

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. `http://www.fairmlbook.org`. fairmlbook.org (cit. on pp. 4, 65).

Baye, Ariane and Christian Monseur (2016). "Gender differences in variability and extreme scores in an international context". In: *Large-scale Assessments in Education* 4 (cit. on p. 20).

Becker, Gary (1971). *The Economics of Discrimination*. 2nd ed. University of Chicago Press (cit. on p. 15).

Benabou, Roland and Jean Tirole (2011). "Identity, Morals, and Taboos: Beliefs as Assets". In: *The quarterly journal of economics* 126, pp. 805–55 (cit. on p. 16).

Bertrand, Marianne and Sendhil Mullainathan (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". In: *American Economic Review* 94.4, pp. 991–1013 (cit. on pp. 4, 19).

Blum, Avrim and Kevin Stangl (2019). "Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?" In: *1st Symposium on Foundations of Responsible Computing (FORC)* (cit. on p. 24).

Bower, Amanda, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alex Vargo, and Suresh Venkatasubramanian (2017). "Fair Pipelines". In: *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML)* (cit. on pp. 13, 87).

Brandt, Felix, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (2016). *Handbook of Computational Social Choice*. 1st. USA: Cambridge University Press (cit. on p. 17).

Braverman, Mark and Sumegha Garg (2020). "The Role of Randomness and Noise in Strategic Classification". In: *Proceedings of The Symposium on Foundations of Responsible Computing (FORC)* (cit. on pp. 12, 51).

Carvalho, Jean-Paul and Bary S. R. Pradelski (2019). "Identity and Underrepresentation: Interactions between Race and Gender". In: *SSRN* (cit. on p. 16).

Cavicchia, Marilyn (2015). "How to fight implicit bias? With conscious thought, diversity expert tells NABE". In: *American Bar Association: Bar Leader* 40.1 (cit. on p. 19).

Celis, L. Elisa, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi (2021). "The Effect of the Rooney Rule on Implicit Bias in the Long Term". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (cit. on pp. 13, 51, 55, 114).

Celis, L. Elisa, Anay Mehrotra, and Nisheeth K. Vishnoi (2020). "Interventions for Ranking in the Presence of Implicit Bias". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, pp. 369–380 (cit. on pp. 6, 13, 20, 23, 24, 29, 36, 43, 44, 51, 55).

Chade, Hector, Gregory Lewis, and Lones Smith (2013). "Student Portfolios and the College Admissions Problem". In: *The Review of Economic Studies* 81 (cit. on p. 113).

Chakraborty, Abhijnan, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau (2019). "Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 129–138 (cit. on p. 17).

Chen, Kun, Kechao Cai, Longbo Huang, and John C.S. Lui (2018). "Beyond the Click-Through Rate: Web Link Selection with Multi-level Feedback". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (cit. on p. 115).

Chouldechova, Alexandra (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2, pp. 153–163 (cit. on pp. 9, 87, 89).

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman (2009). "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4, pp. 661–703 (cit. on p. 42).

Coate, Stephen and Glenn Loury (1993). "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" In: *American Economic Review* 83, pp. 1220–40 (cit. on pp. 16, 55, 114).

Cohen, Patricia (2019). "New Evidence of Age Bias in Hiring, and a Push to Fight It". In: *The New York Times* (cit. on p. 4).

Collins, Brian (2007). "Tackling Unconscious Bias in Hiring Practices: The Plight of the Rooney Rule". In: *NYU Law Review* 82 (cit. on pp. 13, 19, 23).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017). "Algorithmic Decision Making and the Cost of Fairness". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 797–806 (cit. on pp. 9, 87).

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu (2018). "Strategic Classification from Revealed Preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 55–70 (cit. on pp. 12, 51).

Dooley, Samuel et al. (2021). "Comparing Human and Machine Bias in Face Recognition". In: arXiv: 2110.08396 [cs.CV] (cit. on p. 4).

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository* (cit. on pp. 93, 104, 105).

Dutta, Sanghamitra, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney (2020). "Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing". In: *Proceedings of the 37th International Conference on Machine Learning and Systems (PMLR)*, pp. 5067–5077 (cit. on p. 24).

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). "Fairness through awareness". In: *Proceedings of the 3rd conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226 (cit. on pp. 9, 11, 87).

Dwork, Cynthia and Christina Ilvento (2019). "Fairness Under Composition". In: *Proceedings of the 10th conference on Innovations in Theoretical Computer Science (ITCS)*, 33:1–33:20 (cit. on pp. 13, 87).

Estornell, Andrew, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik (2021). *Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents*. arXiv: 2112.02746 [cs.MA] (cit. on p. 12).

Fang, Hanming and Andrea Moro (2011). "Chapter 5 - Theories of Statistical Discrimination and Affirmative Action: A Survey". In: ed. by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson. Vol. 1. Handbook of Social Economics. North-Holland, pp. 133–200 (cit. on pp. 6, 16, 55).

Finocchiaro, Jessie, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis (2021). "Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 489–503 (cit. on p. 4).

Fu, Qiang and Zenan Wu (2019). "Contests: Theory and Topics". In: *Oxford Research Encyclopedia of Economics* (cit. on p. 55).

Gale, D. and L. S. Shapley (1962). "College Admissions and the Stability of Marriage". In: *The American Mathematical Monthly* 69.1, pp. 9–15 (cit. on p. 113).

Garg, Nikhil, Hannah Li, and Faidra Monachou (2021). "Standardized Tests and Affirmative Action: The Role of Bias and Variance". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, p. 261 (cit. on pp. 25, 55, 71).

Geirhos, Robert, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann (2018). "Generalisation in Humans and Deep Neural Networks". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 7549–7561 (cit. on p. 4).

Gersen, Jeannie Suk (2019). "The many sins of college admissions". In: *The New Yorker* (cit. on p. 4).

Gillen, Stephen, Christopher Jung, Michael Kearns, and Aaron Roth (2018). "Online Learning with an Unknown Fairness Metric". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 2605–2614 (cit. on p. 11).

Gordaliza, Paula, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes (2019). "Obtaining Fairness using Optimal Transport Theory". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2357–2365 (cit. on p. 11).

Greenwald, Anthony and Linda Krieger (2006). "Implicit Bias: Scientific Foundations". In: *California Law Review* 94, p. 945 (cit. on pp. 12, 19, 23).

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters (2016a). "Strategic Classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122 (cit. on pp. 12, 51).

Hardt, Moritz, Eric Price, and Nathan Srebro (2016b). "Equality of Opportunity in Supervised Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pp. 3323–3331 (cit. on pp. 9, 10, 11, 87, 89, 90).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. (cit. on p. 9).

Heidari, Hoda and Jon Kleinberg (2021). "Allocating Opportunities in a Dynamic Model of Intergenerational Mobility". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 15–25 (cit. on pp. 14, 114).

Heidari, Hoda and Andreas Krause (2018). "Preventing Disparate Treatment in Sequential Decision Making". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2248–2254 (cit. on p. 87).

Helmbold, David, Nick Littlestone, and Philip Long (2000). "Apple Tasting." In: *Inf. Comput.* 161, pp. 85–139 (cit. on p. 115).

Holzer, Harry and David Neumark (2000). "Assessing Affirmative Action". In: *Journal of Economic Literature* 38.3, pp. 483–568 (cit. on pp. 19, 24, 28).

Hossain, Safwan, Andjela Mladenovic, and Nisarg Shah (2020). "Designing Fairly Fair Classifiers Via Economic Fairness Notions". In: *Proceedings of The Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, pp. 1559–1569 (cit. on p. 16).

Hu, Lily and Yiling Chen (2018). "A Short-Term Intervention for Long-Term Fairness in the Labor Market". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 1389–1398 (cit. on p. 14).

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan (2019). "The Disparate Effects of Strategic Manipulation". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (cit. on p. 12).

Human Rights Watch (2020). "China: gender discrimination in hiring persists". In: *Human Rights Watch* (cit. on p. 4).

IIT-JEE dataset (2019). https://github.com/AnayMehrotra/Ranking-with-Implicit-Bias. [Online; accessed Jan 29, 2020] (cit. on pp. 43, 44).

Isabelle Kocher, seule femme dirigeante du CAC 40 (n.d.). L'Express, Feb. 5, 2020. https://lentreprise.lexpress.fr/actualites/1/actualites/isabelle-kocher-seule-femme-dirigeante-du-cac-40_2117393.html (cit. on p. 20).

Jabbari, Shahin, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth (2017). "Fairness in Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1617–1626 (cit. on p. 87).

Jehiel, Philippe and Matthew V Leduc (2021). "On the Permanent Nature of Affirmative Action Policies". In: working paper or preprint, 33 p. (Cit. on pp. 14, 114).

Joseph, Matthew, Michael Kearns, Jamie H Morgenstern, and Aaron Roth (2016). "Fairness in Learning: Classic and Contextual Bandits". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 325–333 (cit. on pp. 11, 87).

Kannan, Sampath, Aaron Roth, and Juba Ziani (2019). "Downstream Effects of Affirmative Action". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (cit. on p. 13).

Kearns, Michael, Aaron Roth, and Zhiwei Steven Wu (2017). "Meritocratic Fairness for Cross-Population Selection". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 1828–1836 (cit. on p. 11).

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf (2017). "Avoiding Discrimination through Causal Reasoning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pp. 656–666 (cit. on pp. 11, 87).

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *Proceedings of the 8th conference on Innovations in Theoretical Computer Science (ITCS)*, 43:1–43:23 (cit. on pp. 9, 19, 87).

Kleinberg, Jon and Manish Raghavan (2019). "How Do Classifiers Induce Agents to Invest Effort Strategically?" In: *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pp. 825–844 (cit. on pp. 12, 51).

— (2018). "Selection Problems in the Presence of Implicit Bias". In: *Proceedings of the 9th conference on Innovations in Theoretical Computer Science (ITCS)*, 33:1–33:17 (cit. on pp. 5, 6, 12, 13, 15, 20, 21, 23, 26, 29, 36, 42, 51, 55).

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva (2017). "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4066–4076 (cit. on p. 11).

Lambrecht, Anja and Catherine E. Tucker (2018). *Algorithmic Bias? A study of data-based discrimination in the serving of ads in Social Media*. SSRN Electronic Journal (cit. on p. 4).

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (cit. on pp. 4, 93, 105).

Lazear, Edward and Sherwin Rosen (1981). "Rank-Order Tournaments as Optimum Labor Contracts". In: *Journal of Political Economy* 89.5, pp. 841–64 (cit. on p. 55).

Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova (2018). "Does mitigating ML's impact disparity require treatment disparity?" In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 8125–8135 (cit. on pp. 9, 87).

Liu, Lydia T., Nikhil Garg, and Christian Borgs (2021). *Strategic Ranking*. arXiv: 2109.08240 [cs.GT] (cit. on p. 56).

Liu, Lydia T., Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes (2020). "The disparate equilibria of algorithmic decision making when individuals invest rationally". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (cit. on p. 14).

Locatello, Francesco, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem (2019). "On the Fairness of Disentangled Representations". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (cit. on pp. 11, 24).

Lundberg, Shelly J. and Richard Startz (1983). "Private Discrimination and Social Intervention in Competitive Labor Market". In: *The American Economic Review* 73.3, pp. 340–347 (cit. on pp. 15, 16, 54).

Mehrotra, Anay, Bary S. R. Pradelski, and Nisheeth K. Vishnoi (2022). *Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints*. arXiv: 2202.01661 [cs.CY] (cit. on pp. 51, 55).

Miller, John, Smitha Milli, and Moritz Hardt (2020). "Strategic Classification is Causal Modeling in Disguise". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)* (cit. on p. 51).

Milli, Smitha, John Miller, Anca D. Dragan, and Moritz Hardt (2019). "The Social Cost of Strategic Classification". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 230–239 (cit. on p. 51).

O'Dea, R. E., M. Lagisz, M. D. Jennions, and S. Nakagawa (2018). "Gender differences in individual variation in academic grades fail to fit expected patterns for STEM". In: *Nature Communications* 9.1, p. 3777 (cit. on p. 20).

Passariello, Christina (2016). "Tech Firms Borrow Football Play to Increase Hiring of Women". In: *Wall Street Journal* (cit. on p. 19).

Patil, Vishakha, Ganesh Ghalme, Vineet Nair, and Y. Narahari (2020). "Achieving Fairness in the Stochastic Multi-Armed Bandit Problem". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04, pp. 5379–5386 (cit. on p. 11).

Patro, Gourab K, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg (2022). *Fair ranking: a critical review, challenges, and future directions*. arXiv: 2201.12662 [cs.IR] (cit. on pp. 6, 51).

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini (2008). "Discrimination-aware Data Mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 560–568 (cit. on pp. 9, 87).

Petersen, Felix, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin (2021). "Post-processing for Individual Fairness". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (cit. on p. 11).

Phelps, Edmund (1972). "The Statistical Theory of Racism and Sexism". In: *American Economic Review* 62.4, pp. 659–61 (cit. on pp. 6, 15, 16, 24, 33, 52, 54).

Pitoura, Evaggelia, Kostas Stefanidis, and Georgia Koutrika (2021). "Fairness in rankings and recommendations: an overview". In: *The VLDB Journal* (cit. on p. 14).

Quillian, Lincoln, Devah Pager, Arnfinn H. Midtbøen, and Ole Hexel (2017). "Hiring Discrimination Against Black Americans Hasn't Declined in 25 Years". In: *Harvard Business Review* (cit. on p. 4).

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy (2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 469–481 (cit. on p. 24).

Rohatgi, Vijay K. and A.K.M.E. Saleh (2015). *An Introduction to Probability and Statistics*. Wiley Series in Probability and Statistics. Wiley (cit. on p. 100).

Romano, Yaniv, Stephen Bates, and Emmanuel Candes (2020). "Achieving equalized odds by resampling sensitive attributes". In: *Advances in Neural Information Processing Systems* 33, pp. 361–371 (cit. on p. 11).

Sandholm, William H (2010). *Population games and evolutionary dynamics*. MIT press (cit. on pp. 58, 59).

Saxena, Vidit, Joakim Jalden, and Joseph Gonzalez (2020). "Thompson Sampling for Linearly Constrained Bandits". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1999–2009 (cit. on p. 115).

Schotter, Andrew and Keith Weigelt (1992). "Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results". In: *The Quarterly Journal of Economics* 107.2, pp. 511–539. eprint: https://academic.oup.com/qje/article-pdf/107/2/511/5203088/107-2-511.pdf (cit. on pp. 55, 56).

Schumann, Candice, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson (2019). "The Diverse Cohort Selection Problem". In: *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 601–609 (cit. on p. 87).

Senator, Ted E. (2005). "Multi-Stage Classification". In: *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)*, pp. 386–393 (cit. on p. 87).

Slivkins, Aleksandrs (2019). *Introduction to Multi-Armed Bandits*. Vol. 12. 1–2. Now Publishers, pp. 1–286 (cit. on p. 11).

Suksompong, Warut (2021). "Constraints in Fair Division". In: *SIGecom Exch.* 19.2, pp. 46–61 (cit. on p. 17).

Temnyalov, Emil (2019). "An information theory of efficient differential treatment". In: *SSRN* (cit. on p. 25).

Trapeznikov, Kirill, Venkatesh Saligrama, and David Castañón (2012). "Multi-Stage Classifier Design". In: *Proceedings of the Asian Conference on Machine Learning*, pp. 459–474 (cit. on p. 87).

Tsirtsis, Stratis and Manuel Gomez Rodriguez (2020). "Decisions, Counterfactual Explanations and Strategic Behavior". In: *Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)* (cit. on pp. 12, 51).

"Uniform Guidelines on Employee Selection Procedures" (1978). In: *Code of Federal Regulations*. US (cit. on pp. 10, 19).

Valera, Isabel, Adish Singla, and Manuel Gomez Rodriguez (2018). "Enhancing the Accuracy and Fairness of Human Decision Making". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 1774–1783 (cit. on p. 87).

Wick, Michael, Swetasudha Panda, and Jean-Baptiste Tristan (2019). "Unlocking Fairness: a Trade-off Revisited". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8783–8792 (cit. on p. 24).

Woolley, Suzanne (2017). "How More Americans Are Getting a Perfect Credit Score". In: *Bloomberg* (cit. on p. 51).

Zafar, Muhammad B., Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi (2017a). "Fairness Constraints: Mechanisms for Fair Classification". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 962–970 (cit. on p. 11).

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi (2017b). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 1171–1180 (cit. on pp. 9, 29, 87, 89).

Zemel, Richard, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork (2013). "Learning Fair Representations". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, III–325–III–333 (cit. on pp. 11, 24).

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, pp. 335–340 (cit. on p. 11).

Zhang, Hanrui, Yu Cheng, and Vincent Conitzer (2019). "When Samples Are Strategically Selected". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 7345–7353 (cit. on p. 51).

Zhang, Xueru and Mingyan Liu (2021). "Fairness in Learning-Based Sequential Decision Algorithms: A Survey". In: Springer International Publishing, pp. 525–555 (cit. on p. 11).

Zhang, Xueru, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang (2020). "How do fair decisions fare in long-term qualification?" In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 18457–18469 (cit. on p. 14).