Summer 9-13-2023

Consequences and Incentives of Fair Learning

Andrew Estornell
Washington University – McKelvey School of Engineering

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Dissertation Examination Committee:
Chien-Ju Ho, Chair
Sanmay Das
Brendan Juba
Paulen Kim
Yevgeniy Vorobeychik

Consequences and Incentives of Fair Learning
by
Andrew Estornell

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisors Yevgeniy Vorobeychik and Sanmay Das for their guidance and support throughout the whole of my Ph.D. I am also grateful to my committee members, Chien-Ju Ho, Brendan Juba, and Paulen Kim who have been essential in steering the direction of this thesis. Lastly I would like to thank Yang Liu for his continued insight and advice.

This work was supported by both the National Science Foundation and the Department of Defense.

Andrew Estornell

Washington University in St. Louis
August 2023

ABSTRACT OF THE DISSERTATION

Consequences and Incentives of Fair Learning

by

Andrew Estornell

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2023

Professor Chien-Ju Ho, Chair

As algorithmic decision-making systems become increasingly entrenched in human-centric domains such as hiring and lending, it is crucial that these systems do not perpetuate historical bias or unfairly discriminate against sensitive demographics. However, in these domains, considering group fairness as the sole factor often leads to two significant consequences: 1) incentivizing individuals to strategically alter their behavior to obtain desired outcomes (e.g., hiding debt to qualify for a loan), and 2) achieving fairness at the expense of individual welfare (e.g., equalizing lending rates between groups by offering fewer total loans). We explore these consequences from both theoretical and empirical perspectives with the goal of characterizing when and why such phenomena occur, as well as developing solutions to mitigate their negative side effects. Our findings suggest that traditional group-fair learning, i.e., optimizing solely for group fairness and predictive performance, can frequently result in both of the aforementioned consequences, implying that an isolated focus on group fairness can lead to increased manipulative behavior and widespread decreases in individual welfare. Notably, the former has the potential to decrease model fairness, suggesting that optimizing for group fairness can be counterproductive (i.e., resulting in less fair models) when the model creates incentives for strategic behavior. In light of the pitfalls of group-fair learning, we propose several approaches to mitigate their adverse effects. From the perspective of strategic behavior, we propose an auditing mechanism that discourages manipulative behavior and promotes true feature changes (i.e., promotes recourse). From the perspective of individual welfare, we develop two learning schemes that preserve individual welfare while achieving high levels of performance and group fairness. In addition to providing theoretical guarantees for both these welfare-aware learning schemes and the auditing mechanism, we also demonstrate their practical efficacy through experiments on a multitude of datasets from several different domains. Our results indicate that by adopting a more nuanced approach to group-fair learning, it is possible to build models that avoid these negative side effects without compromising performance or group fairness.

Part I

Background and Overview

Chapter 1

Introduction

The ever-increasing role of algorithmic decision-making systems in human-centered domains has sparked deep investigation into the efficacy of these systems. A common lens through which such investigations are conducted is that of *group-fairness*, which aims to ensure that when machine learning models are deployed in high-stakes domains, such as social service assistance or employment, the decisions of these models adhere to notions of fairness between demographics defined by protected features, e.g., equalizing false positive rates between males and females in recidivism prediction. There is a large body of work focusing on both defining what it means for a model to be fair, as well as developing techniques for building fair models [2, 62, 95, 33, 34, 68, 42, 38]. While this line of work has seen great success in developing models that achieve group fairness, these models can often have subtle but consequential side-effects, such as decreasing individual welfare or incentivizing individuals to engage in manipulations. The work presented in this thesis aims to characterize these side-effects, provide tools for model designers to identify when they can occur, and to develop algorithms that mitigate the severity of these side-effects. Our investigations into these topics will fall under three primary lines of inquiry. First, we will examine the ways in which strategic incentives arise in group-fair learning, as well as the impacts that the resulting strategic behavior has on model fairness and performance. Second, we explore the individual-level harms that arise as a result of imposing group-fairness constraints on predictive models, e.g., decreases to individual welfare or loss of desired opportunities among already ”disadvantaged” individuals. We will see that there is a fundamental connection between the individual harms caused by the imposition of group-fairness and the impacts of strategic behavior on model fairness and performance. Third, we develop several algorithmic techniques for mitigating both the impacts of strategic behavior on group-fair models and the individual-level harms caused by group-fairness constraints.

We begin by examining the impacts of strategic behavior on model fairness. In particular, we compare the fairness of a group-fair model to a fairness-agnostic model (also called a *conventional* model). The deployment of fair models involves replacing a preexisting conventional model that is currently in use. The fair learning literature primarily focuses on fair models in isolation and pays little attention to conventional predecessors, except to the extent that the predecessor is unfair, while the fair model is not. However, a deep understanding of the relationship between fair and conventional learning schemes is vital to the efficacy of these fair schemes. For example, if a model designer deploys a group-fair model in lieu of a conventional model, but strategic agent behavior causes the group-fair model to become *less* fair than the conventional model, the change in model was ultimately counterproductive to the model designer's intention. As most group-fair learning schemes presume faithful behavior from agents, there is a need to better understand the relationship between strategic behavior and group-fair learning. To this end, we define the notion of *fairness reversals*, which capture the event where strategic agent behavior causes the group-fair model to become *less* fair than its conventional counterpart. Fairness reversals can be interpreted as a counterfactual analysis of model choice, in which the system would have been more fair if the model designer had elected *not* to add fairness constraints. Our key finding is that fair models that are more selective (i.e., those that negatively classify more individuals) than their conventional counterpart are precisely the models that result in fairness reversals. We also characterize conditions under which the fair model will indeed be more selective. Further details are discussed in Part II.

Next, we investigate the complementary problem of how group-fairness impacts individual agents. We again take the perspective of counterfactual comparisons between a fair model and "reasonable" choices of conventional models. From an individual-centered view, we investigate the ways in which each individual's treatment differs as a result of the model designer's choice to impose group fairness. In particular, we aim to answer two primary questions about the relationship between the use of group-fair and fairness-agnostic models: 1) at an individual level, are there agents who are reliably made worse off by the use of fair learning schemes over conventional learning schemes, and if so, are there any meaningful similarities between such agents; and 2) are these negative impacts distributed unequally between groups, i.e., do individuals in one group, compared to other groups, bear more of the impact associated with imposing fairness.

To answer the first question, we look at both classification and scarce resource allocation. In the context of classification, we investigate whether an individual's probability of being positively classified is decreased under a fair learning scheme. We find that common fair learning schemes frequently result in models where roughly 25%-50% of individuals have their chance of positive classification decreased. In the scarce resource allocation setting, a score function is used to rank agents in a population, and the top-k scoring agents receive a desired resource. We investigate whether individuals *reliably* lose their resource when using a fair score function in place of a conventional score function. Unlike classification, individual decisions can no longer be made in isolation since allocations are dependent on the distribution of scores across the population. As such, we propose a more nuanced manner of examining whether an individual loses their resource due to the inclusion of group-fairness or due to the innate randomness in both the choice of score functions and the population. More specifically, we define *perceived-impact*, which captures an individual's perception of whether or not they lost allocation simply when changing score functions without accounting for innate randomness, and *realized-impact*, which captures whether the individual would have lost their resource when switching to a fair score function while accounting for both randomness in the choice of score function and population. We find that across a multitude of datasets and score function choices, both perceived- and realized-impact are frequently high. Further, we observe that perceived-impact is often a significant overestimate of realized-impact, implying that individuals may believe themselves to be more harmed than they actually are.

To answer the second question, we look at both individual welfare in classification (chance of being positively classified), as well as perceived- and realized-impact across groups. We find that in cases of classification, it is typically the advantaged group whose individual welfare is decreased. Similarly, in the case of scarce resource allocation, both perceived- and realized-impact fall predominantly on individuals in the advantaged group. However, when resources are abundant (more than one resource for every two individuals), the disadvantaged group frequently suffers a larger impact than the advantaged group. Further details are provided in Part II.

Lastly, we investigate techniques to mitigate the aforementioned side-effects: both the negative impacts on individuals as well as the negative impacts of strategic behavior on model fairness. In particular, to mitigate harm at the individual level, we propose a general post-processing approach for learning models that are both fair and maintain a specified level of individual welfare. This postprocessing framework can be used in both classification and scarce resource allocation settings and works for a broad class of *additive* fairness metrics (e.g., FPR, PR, TPR, ERR, etc.). We find that this framework is capable of producing models that have high levels of individual welfare while also maintaining high group fairness and predictive efficacy. Furthermore, we demonstrate that these postprocessing techniques can also be used to prevent the selectivity of the fair classifier, which is ultimately the root cause of fairness reversals.

While this postprocessing technique can reduce fairness reversals, it does not directly address strategic manipulations by agents. Thus, we also propose an auditing framework that can be used to directly reduce strategic agent behavior. An audit constitutes a verification that features submitted by an agent were not the result of manipulation (e.g., the IRS verifying information on tax returns); as such verifications may be costly, the auditor has a limited budget with which to audit. Agents found to be manipulating may be subject to rejection and an additional fine. Within our auditing framework, we characterize optimal auditing policies for three primary objectives: 1) minimizing the incentive to manipulate, 2) maximizing the number of agents making true *feature improvements* rather than manipulations, and 3) maximizing system utility. The first objective corresponds to finding an audit policy that makes truthful reporting an ε-equilibrium for the minimum value of ε, meaning that no agent can gain more than ε utility by choosing to manipulate when all other agents are truthful. When $\varepsilon = 0$, truthful reporting is an equilibrium strategy, and model designers can recover both the fairness and performance of their model using the truthful data. The second objective can be seen as a form of social good in that the auditor aims to maximize the number of agents obtaining a desired outcome via feature improvement, which, unlike manipulations, can improve the agent's true qualification (e.g., paying off debt improves an individual's creditworthiness while hiding debt does not). The third objective captures a self-interested auditor who aims only to improve their own utility (e.g., a bank aiming to maximize profits made on loans). We characterize necessary and sufficient conditions for the latter two objectives to align, meaning that an auditor aiming to maximize their own utility will also maximize the fraction of agents electing to perform feature improvements.

Furthermore, we outline several advantages of auditing compared to traditional methods for achieving robustness. In particular, auditing will never result in agents being required to manipulate in order to maintain positive classification, as is the case with almost all other traditional approaches (e.g., adversarial retraining). We also investigate the role of subsidies in auditing, where subsidies correspond to the auditor electing to perform fewer audits and instead use a portion of their audit budget to decrease the cost (e.g., allotting a fraction of the audit budget to produce and distribute educational material on financial literacy, thus decreasing the cost for agents to improve their creditworthiness). We provide necessary and sufficient conditions under which the auditor will allot a nonzero fraction of their audit budget to subsidies, thus decreasing the cost of feature improvements.

Chapter 2

Related Work

Here, we provide an overview of works that are relevant to the scope of this thesis. In some chapters, we also offer a more in-depth discussion of works that are of particular relevance to the topic of that chapter. The work presented in this thesis primarily falls under three fields: fair learning, strategic classification, and mechanism design.

2.1 Fair Learning

The field of fair learning is focused around both defining what it means for machine learning model to be fair (e.g., equal error rate between different demographics), as well as opperationalizing those definitions in order to develop learning schemes which yield fair models. While fairness can be can be defined in many different ways, the definitions found in the fair-learning literature can be broken into two categories: *individual fairness* which captures the idea that similar individuals should be treated similarly [33, 118, 89, 93, 103, 69, 74], and *group fairness* which captures the idea that groups or subpopulations should be treated similarly [43, 95, 62, 38, 53, 22, 52, 50, 48, 68, 51]. We will focus primarily on the latter, but the former will be relevant for several components of later chapters. Works in fair learning can be also be broken into three categories based on the way in fair fairness is opperationalized: *preprocessing* which modifies data such that models trained on that modified data are fair [38, 121, 53, 30, 83, 81, 75, 117, 19], *inprocessing* which modifies the learning procedure in order to obtain a fair model [23, 62, 2, 68, 95, 52, 91, 51, 112, 25], and *post-processing* which modifies the decisions of a pretrained model such that the model becomes fair [43, 50, 93, 77, 98, 7]. In the context of group-fair learning, each of these works aims to ensure that predictions (almost always binary predictions) made between groups, subgroups, or subpopulations, defined by sensitive features (e.g., race, age, gender, etc.) have roughly equal value under some type of metric (typically an error metric). For example

[43, 62, 95, 2, 25] can produce classifiers which have roughly equal false positive rate, or true positive rate, between different demographics. Some works have also investigated fair learning in the context of regression [12, 3, 66]. We will focus on group-fairness in the context of binary classification, which trivially extended to multi-class classification in nearly all settings. Most of the aforementioned group-fair learning schemes allow the model designer to specify a set of desired fairness metrics as well as threshold for "unfairness", e.g., the model designer can specify that the false positive rate between any two group cannot exceed 0.2. Of particular relevance to the scope of this thesis is the assumption, made by each of these works, that data is reported truthfully to the model (both at test- and decision-time) and that any alterations are the result of "natural" noise.

2.2 Strategic Classification

The field of strategic classification aims to capture settings in which strategic agents manipulate a decision making system in order to gain more favorable outcomes, e.g., an applicant under-reporting debt in order to be approved for a loan. Agents are modeled as being selfish, and manipulate as a means of increasing their own utility [41], which is in contrast to the malevolent objectives of agents (attackers) found in adversarial machine learning [47]. As such, the objectives of the model designer and the strategic agents are not necessarily at odds, and can sometimes even be aligned [71]. When deciding the optimal best response to a given model, agents weigh both their valuation of a desired outcome along with the associated cost of obtaining that outcome, and select the action (manipulation) which yields the highest value (possibly in expectation). Manipulation costs capture Works in strategic classification have development of classifiers robust to strategic manipulation [41, 87, 70, 31], to measuring the ways in which robustness and strategic behavior leads to negative outcomes for the population [87, 46, 4].

While there are some works which aim to merge the ideas of group fairness and strategic agents [87, 46, 23, 91, 116, 118], these works either study the effects of strategic behavior of relatively specialized definitions of fairness (such as the cost to remain positively classified) [46, 87, 104], frame strategic agents from the perspective of adversarial machine learning (i.e., agents wish too degrade model performance, not increase their own likelihood of positive outcomes) [116, 120, 91, 23], or fair classification which is robust to non-strategic noise

[51, 112]. That is to say that the intersection of group fairness and strategic classification is still widely unexplored.

A generalization of strategic behavior, refereed to as *performative prediction* and first proposed in [92], captures the ways in which predictions made by machine learning models can influence individuals and populations, one such type of influence being the incentive to strategically manipulate. Followup works in the area have further generalized this concept to notions of populations adapting to model decisions through strategic, or non-strategic, means [97, 86, 16, 40]. These works typically take a model-centered approach by examining algorithms which can cope with population shifts caused by model influence, such as repeated risk minimization [16, 92] (similar to the technique of adversarial training).

A particularly relevant type of agent adaption is *actionable recourse* [109], in which agents seek to obtain a desired outcome through earnest means; rather than manipulating or misreporting their features, agent make actual *feature improvements* such as paying off debt instead of imply hiding debt. Within the actionable recourse literature the objective of model designers is to develop models which enable large fractions of a population to achieve positive classification [109, 100, 100, 58, 56, 24]. From a mathematical perspective, objectives of this form can be viewed an inverse form of the model designer's objective in strategic classification; the former aims to maximize the number of agents capable of achieving positive classification, while latter aims to minimize the number of such agents.

2.3 Mechanism Design

Lastly the field of mechanism design aims to develop systems which elicit a particular type of behavior from agents such as truthful reporting [15]. In the scope of our work, we will be interested in using mechanism as a means of shaping agent behavior such that either they do not act strategically, i.e. the classifier incentive compatible [90], or such that such that impacts of strategic behavior on model performance is minimized, i.e. increase model robustness. One type mechanism of particular relevance to our work is auditing. Auditing constitutes a verification that the information submitted by an agent (e.g., a loan application) is truthful. First formalized in game theoretic context by [14], auditing has seen use in a wide array of domains, particular in the context of scare recourse allocation [79, 10, 36, 13]. While these works make use of auditing in a somewhat similar context to ours, they presume the agents' ability to directly modify their score or requested resource, rather than accounting for the necessary feature changes or feature manipulations required to obtain a particular score or resource. As such the auditing schemes proposed in these works do not apply to the case of strategic classification where agents must modify their features in order to obtain a desired outcome.

Chapter 3

Preliminaries

In this section, we outline general notation a definitions used throughout the thesis. When necessary, a set of more comprehensive preliminaries is provided in each chapter.

f	model, typically a classifier
$\mathcal{X}, \mathcal{Y}, G$	features, labels, and groups respectively
$\mathcal{D}$	distribution over $(\mathcal{X}, \mathcal{Y}, G)$
$\mathcal{L}(f, \mathcal{X}, \mathcal{Y})$	loss of model f w.r.t. features $\mathcal{X}$ and labels $\mathcal{Y}$
$c(\mathbf{x}, \mathbf{x}')$	cost of changing feature $\mathbf{x}$ to $\mathbf{x}'$
B	manipulation budget, i.e., $c(\mathbf{x}, \mathbf{x}') \leq B$
$\mathcal{M}(f\|g)$	efficacy metric of model f on group g
$U(f; \mathcal{M})$	unfairness of model f w.r.t. metric $\mathcal{M}$
$\beta, \ \alpha$	hard and soft fairness constraints respectively i.e. $U(f; \mathcal{M}) \leq \beta$ or $\mathcal{L}(f, \mathcal{X}, \mathcal{Y}) \mathrel{+}= \alpha U(f, \mathcal{M})$
$f^{(c,B)}$	classifier resulting from agents best responding to f with budget B and manipulation cost function c

Table 3.1: Notation

Throughout we presume a population of agents given as $(\mathcal{X}, \mathcal{Y}, G)$, where agent $(\mathbf{x}, y, g)$ is characterized by a feature vector $\mathbf{x} \in \mathcal{X}$, a group $g \in G$ to which they belong, and a binary (true) label $y \in \mathcal{Y} \equiv \{0, 1\}$. For ease of analysis we presume binary groups, i.e., $G \equiv \{0, 1\}$, however all results (unless otherwise stated) extend to the case of three or more groups. Let $\mathcal{D}$ be the joint distribution over $G \times \mathcal{X} \times \mathcal{Y}$.

Classification and Fairness We denote fairness-agnostic classifiers, also called *conventional classifiers*, as f_C, and group-fair classifiers as f_F. Both classifiers map from the domain of features $\mathcal{X}$ to the set of binary labels $\mathcal{Y}$, i.e. $f_C, f_F : \mathcal{X} \rightarrow \mathcal{Y}$. Note that either classifier can be group-aware or group-agnostic, i.e. the classifier makes use of, or does not make use of, the sensitive features G at decision time ($\mathcal{X}$ could simply contain a copy of G).

When relevant we will make whether the models are group-aware or group-agnostic. Let $\mathcal{M}(f;g)$ be a measure of efficacy (e.g., positive rate) of f restricted to a group g, and define $U(f;\mathcal{M}) = \big|\mathcal{M}(f|g=1) - \mathcal{M}(f|g=0)\big|$. We shorten this notation to $U(f)$ where $\mathcal{M}$ is clear from context. We assume that the conventional classifier aims to minimize loss,

$$f_C = \arg\min_f \mathcal{L}(f, \mathcal{X}, \mathcal{Y}) \tag{3.1}$$

while f_F aims minimize loss subject to a fairness constraint,

$$\begin{gathered} f_F = \arg\min_f \mathcal{L}(f, \mathcal{X}, \mathcal{Y}) \\ U(f;\mathcal{M}) \leq \beta \end{gathered} \tag{3.2}$$

where $\beta \in [0,1]$ specifies allowed "unfairness". Again for ease of analysis we presume a single fairness constraint, but all results extend to multiple fairness constraints unless otherwise stated.

For the majority of this paper we will focus on four common group-fairness metrics defined next. In subsequent chapters, we will explicitly state when results extend to other fairness.

Definition 3.0.1. *Common fairness metrics:*

$$\begin{aligned}
\textbf{\textit{Positive rate (PR):}}\ U(f, PR) &= \big|\mathbb{P}\big(f(\mathbf{x})=1|g=1\big) - \mathbb{P}\big(f(\mathbf{x})=1|g=0\big)\big| \\
\textbf{\textit{True positive rate (TPR):}}\ U(f, TPR) &= \big|\mathbb{P}\big(f(\mathbf{x})=1|g=1, y=1\big) \\
&\quad - \mathbb{P}\big(f(\mathbf{x})=1|g=0, y=1\big)\big| \\
\textbf{\textit{False positive rate (FPR):}}\ U(f, FPR) &= \big|\mathbb{P}\big(f(\mathbf{x})=1|g=1, y=0\big) \\
&\quad - \mathbb{P}\big(f(\mathbf{x})=1|g=0, y=0\big)\big| \\
\textbf{\textit{Error rate (ERR):}}\ U(f, ERR) &= \big|\mathbb{P}\big(f(\mathbf{x})\neq y|g=1,\big) - \mathbb{P}\big(f(\mathbf{x})\neq 1|g=0\big)\big|
\end{aligned}$$

In all cases, we refer to the "advantaged" group (e.g. the group with higher PR for PR based fairness) as group 1, or G_1, while the disadvantaged group is referred to as 0 or G_0.

Strategic Behavior We consider the agents strategically responding to classifier f (conventional or group-fair). Specifically, we suppose that each agent with features $\mathbf{x}$ can manipulate these features to produce features $\mathbf{x}'$ that are then reported to the classifier. Manipulations do *not alter* the agents true label y. When manipulating, the agent incurs a cost, captured by a manipulation cost function $c(\mathbf{x}, \mathbf{x}') \geq 0$ [41, 87, 46]. For example, if $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$, then "larger" manipulations are more expensive.

The utility of an agent is then,

$$u(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}') - f(\mathbf{x}) - {}^{1}\!/_{B} \cdot c(\mathbf{x}, \mathbf{x}'),$$

where B is a parameter trading off costs and benefits of manipulation. Following the standard setting in strategic classification or adversarial machine learning, we assume any misreporting behavior would not change the true label y associated with $\mathbf{x}$. We assume that all agents are rational utility maximizers. Thus, since $f(\mathbf{x}') - f(\mathbf{x}) \leq 1$, the agent will misreport its features only when $c(\mathbf{x}, \mathbf{x}') \leq B$. Additionally, the agent will not misreport if $f(\mathbf{x}) = 1$ (they are selected even when truthfully reporting $\mathbf{x}$). Consequently, we can equivalently view B as an upper bound on the costs that agents are willing to incur from misreporting their features, that is, the *manipulation budget*.

Chapter 4

Datasets, Models, and Algorithms for Experiments

Here we outline datasets and learning schemes. For consistency, the same datasets and learning schemes are used throughout each chapter. Deviations, when necessary, are stated explicitly.

Datasets For our empirical study, we use five datasets commonly used as benchmarks for group-fair classification: **Adult:** Dataset of working professionals where the goal is to predict high or low income (protected feature: gender) [67, 32]. **Crime:** Dataset of communities where the objective is to predict if the community has high crime (protected feature: race) [99, 32]. **Law:** Dataset of law students where the objective is to predict bar-exam passage (protected feature: race) [114]. **Student:** Dataset of students where the objective is to predict a student receiving high math grades (protected feature: race) [28, 32]. **Credit:** Dataset of people applying for credit where the objective is to predict creditworthiness (protected feature: age) [32]. All five datasets have binary outcomes, and we label the more *desirable* outcome for the individuals by $y = 1$ (e.g., having a high income in the Adult dataset), with the less desirable outcome labeled by $y = 0$. Sensitive features are also considered as binary, for example, the *age* feature is an indicator that the individual is Young or Old.

Learning Schemes In our experimentation we use the following models: logistic regression (LGR), support vector machines with an RBF kernel (SVM), neural networks (NN), and gradient boosting trees (GB). For fair models we use four common learning paradigms: *Reductions* [2] an inprocessing technique which works via cost-sensitive learning and is suited to PR, TPR, FPR, ERR fairness, *GerryFair* [62] an inprocessing technique which works via

name	sensitive features	binary label	size
Adult [67]	age, race, gender, nationality	earns $\geq$ 50k per-year	5,200
Crime [99]	race	has “high” crime rate	2,000
Law [114]	age, race gender)	pass bar on first attempt	1,700
Student: [28]	age, gender	pass AP exams	395
Credit: [32]	gender, age	is creditworthy	1,000

Table 4.1: Dataset Description

minmax fictitious play and is suited for TPR and FPR fairness, *EqOdds* [95] a postprocessing technique which works via randomized group-wise thresholds and is suited for TPR and FPR fairness, and *KDE* [25], an inprocessing technique which works via kernel density estimation and is suited for TPR, FPR, PR, and ERR fairness.

Strategic Manipulations Agents best responding to a classifier f will be a function of the classifier itself, feature types (discrete or continuous), and the cost function c. In the majority of experiments we study cost functions which are l_p-norms, i.e., $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p$. When c is an l_p-norm and the classifier is differentiable (LGR, SVM, and NN) we use a mix of projected gradient descent (PGD) for continuous features [82] use a mix of local search for categorical features [106, 72]. When the classifier is non-differentiable (GB), we make use local search methods for ensemble-based classifiers [122]. Several chapters make use of cost functions different from l_p-norms, in those chapters the methods for computing optimal manipulations is discussed in detail. In cases where the feature space is small, more effective attacks can be found by approximating the decision boundary of the classifier as a set of discrete points and binary searching for the lowest cost point.

Part II

Impacts of Group-Fair Learning

Chapter 5

Group-Fairness and Strategic Behavior

We being our investigation by examining the relationship between strategic manipulation and group fair classification. When model decisions are consequential to individuals' utility, the deployment of machine learning models inevitably creates a environments in which individuals are incentivized to behave strategically. Such strategic behavior can potentially undermine the model efficacy. This phenomenon has been studied extensively in the context of fairness agnostic learning [41], remains relatively under-studied in the context of fairness-aware learning. Most analysis of fair algorithms proceeds from the assumption that the people affected by algorithmic decisions will not modify their features (manipulate) in-order to improve their outcome. However, this assumption is impractical in many real world settings as the manipulation of automated decision making systems in fairness-critical domains such as welfare assistance and lending [20, 70] has been well documented. We investigate the role of strategic manipulation in the context of fair learning and find a plethora of settings in which strategic behavior decreases model fairness. In particular we find that strategic behavior frequently leads to *fairness reversal*, with a conventional classifier, in which the fair model exhibited higher unfairness than a classifier trained with no consideration of fairness. We further contextualize the fairness reversal phenomenon by providing conditions under which fairness reversals occur; in particular, we show that fairness reversal occurs as a result of a group-fair classifier becoming more *selective*, achieving fairness largely by excluding individuals from the advantaged group. In contrast, if group fairness is achieved by the classifier becoming more *inclusive*, fairness reversal does not occur.

We investigate the effects of such strategic manipulation of a binary *group-fair* classifier. In the social services example, the classifier may decide whether an applicant receives assistance, and the fairness criterion could be approximate equality of selection rate between male and female applicants. First, we observe that the ability of individuals to manipulate the features a classifier uses can lead to *fairness reversal*, with the conventional (accuracy-maximizing) classifier exhibiting greater fairness than a group-fair classifier. We demonstrate this phenomenon on several standard benchmark datasets commonly used in evaluating group-fair classifiers. Next, we theoretically investigate conditions under which fairness reversal occurs. We prove that the key characteristic that leads to fairness reversal is that the group fair classifier becomes more selective, excluding some of the individuals in the advantaged group from being selected. Moreover, we show that this condition is sufficient for fairness reversal for several classes of functions measuring feature misreporting costs. In contrast, we experimentally demonstrate that when a group-fair classifier exhibits inclusiveness instead by selecting additional individuals from the disadvantaged group, fairness reversal does not occur.

5.1 Summary of Results

We begin by observing empirically the phenomenon of fairness reversal, exhibited on a number of datasets commonly used in bench-marking group-fair classification efficacy. We demonstrate that the key factor resulting in fairness reversals is the extent to which group fairness is achieved through increased selectivity (the fair classifier f_F positively classifies fewer inputs than the conventional classifier f_C) as opposed to increased inclusiveness (f_F positively classifies more inputs than f_C). In particular, classifiers skewed towards the former frequently exhibit fairness reversals as where those skewed towards the latter do not.

Next, we examine this issue theoretically, and prove that selectivity is indeed a sufficient condition for fairness reversal. Further, we show that under some additional conditions, selectivity is also a necessary condition. These theoretical results hold for two common classes of manipulation cost functions (features- and outcome-monotonic costs). Lastly we investigate *why* some fair classifiers are selective, while others are not, and provide a set of sufficient conditions on both the distribution of labels and features as well as the fair learning scheme. These conditions helps explain our empirical observations as they hold frequently for each of the dataset and classifier combinations used in our experiments.

5.2 Preliminaries

We begin by extending the model found in Chapter 3 with additional details relevant the effects of strategic behavior on group-fair classifiers. For results in Part II we focus on the three most common types of fairness metrics found in the literature: false positive rate (FPR), true positive rate (TPR) and positive rate (PR), [43, 38, 62, 121, 2]. Additionally we focus on conventional and fair models which are optimized for expected accuracy, i.e. the conventional and fair objectives given for a general loss function $\mathcal{L}$ in Equations 3.1 and 3.2 now become

$$f_C = \arg\min_f \mathbb{P}(f(\mathbf{x}) \neq y) \tag{5.1}$$

and

$$\begin{gathered} f_F = \arg\min_f \mathbb{P}(f(\mathbf{x}) \neq y) \\ \text{s.t.} U(f; \mathcal{M}) \leq \beta \end{gathered} \tag{5.2}$$

For ease of analysis we will consider fairness in terms of so called *soft constrained fairness* rather than the *hard constrained fairness* found in 3.2. That is, for a given *fairness importance* weight $\alpha \in [0, 1]$,

$$f_F = \arg\min_f (1 - \alpha)\mathbb{P}(f(\mathbf{x}) \neq y) + \alpha U(f; \mathcal{M}) \tag{5.3}$$

For the fairness metrics of interest, FPR, TPR, and PR, the objectives in Equations 5.2 and 5.3 are equivalent. We formally state this next as a theorem, but first remark that this equivalence implies that our theoretical and experimental results hold for both soft constrained and hard constrained fairness.

Theorem 5.2.1. *Let $\mathcal{M}$ be defined by PR, FPR, or TPR. For any $\alpha \in [0, 1]$ there exists a $\beta \in [0, 1]$ such that the optimal classifier for Equation 5.2 is also an optimal solution to Equation 5.3. Conversely for any $\beta \in [0, 1]$ there exists an $\alpha \in [0, 1]$ such that the optimal classifier for Equation 5.3 is also an optimal solution to Equation 5.2.*

Proof. The full proof is provided in Section A of the Appendix. The intuition for this result follows similarly to that of strong-duality results found in optimization theory. □

Further, we focus our analysis on the two most common families of cost functions in strategic classification literature, namely feature-monotonic costs [41] and outcome-monotonic cost function [87].

Definition 5.2.2. ***(Feature-Monotonic Costs)****: A cost function c is said to be* feature-monotonic *if $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $||\mathbf{x} - \mathbf{x}'||$, i.e., larger manipulations are more costly.*

Definition 5.2.3. ***(Outcome-monotonic costs)****: A cost function c is said to be* outcome-monotonic *if $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $\mathbb{P}(y = 1|\mathbf{x}') - \mathbb{P}(y = 1|\mathbf{x})$ where $c(\mathbf{x}, \mathbf{x}') = 0$ for any $\mathbf{x}'$ such that $\mathbb{P}(y = 1|\mathbf{x}) > \mathbb{P}(y = 1|\mathbf{x}')$, i.e., manipulations leading to better outcomes are more costly.*

5.3 Fairness Reversals

To capture the effects of strategic behavior on group-fair classifiers, we contrast these classifiers against fairness-agnostic (or conventional) classifiers, i.e. those corresponding to Equation 5.1. In particular we will be interested in the ways in which strategic behavior impacts the *relative* fairness of each model. In particular, we propose the notion of a *fairness reversal* in the presence of strategic agents, i.e., strategic agent behavior to leads to the fair model f_F becoming less fair than conventional model f_C.

Definition 5.3.1. ***(Fairness Reversal)*** *Let $\mathcal{M}$ be a measure of efficacy, f_F be a classifier which is* group-fair *with respect to $U(f; \mathcal{M})$ and f_C be a conventional accuracy-maximizing classifier. Suppose that $U(f_F; \mathcal{M}) < U(f_C; \mathcal{M})$. Let $f_C^{(c,B)}, f_F^{(c,B)}$ be the induced classifiers when agents best respond to f_C and f_F respectively with manipulation cost $c(\mathbf{x}, \mathbf{x}')$ and budget B. We say that a* budget B leads to fairness reversal *between f_C and f_F if $U(f_F^{(c,B)}; \mathcal{M}) \geq U(f_C^{(c,B)}; \mathcal{M})$.*

We will then say that fairness reversal between f_F and f_C occurs if there is some strategic manipulation budget B which leads to fairness reversal, that is, for this budget, f_C becomes more fair than f_F after manipulation. Note that if the budget B is 0, f_F will be more fair than f_C by construction, whereas if the budget is infinite, as long as any input is classified as the positive class, all individuals can misreport their features to be this class, and consequently both classifiers are fair in the sense that every input is predicted as 1. As a result, our analysis is focused solely on the intermediate cases between these extremes.

Selectivity We find that it is classifier *selectivity* which leads to fairness reversals. More specifically, examining FPR, TPR, and PR fairness in binary classification, there are two ways in which the fair classifier f_F can "correct" the unfairness of the conventional classifier f_C. The fair classifier can either increase TPR, FPR, or PR on the disadvantaged group (the group with lower TPR, FPR, or PR), or decrease TPR, FPR, or PR on the advantaged group (the group with the higher TPR, FPR, or PR). The relative frequency at which these two actions occur will ultimately determine whether a fairness reversal occurs.

Definition 5.3.2. *Let* $\mathcal{X}_{f_C} = \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$ *and* $\mathcal{X}_{f_F} = \{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\}$*. We say that* f_F is more selective than f_C *if* $\mathcal{X}_{f_F} \subset \mathcal{X}_{f_C}$*.*

That is, f_F is more selective than f_C if the set of positively classified examples under f_F is a subset of those positively classified under f_C. While this definition of selectivity is slightly more restrictive than the type of selectivity found in our empirical results, the subset propriety is a driving force behind the fairness reversals observed in practice. Selectivity can be interpreted as the fair model f_F, achieving fairness by "excluding" additional agents from positive classification, compared to f_C. As an example, under PR-based fairness let G_0 be the group with lower PR and G_1 be the group with higher PR under f_C (TPR and FPR hold similarly). A model designer could improve the fairness of f_C by positively classifying more agents in G_0 or negatively classifying more agents in G_1 (or a combination of both). In the latter case, members of G_0 are "excluded" from positive classification, and the resulting model is considered to be *more selective*. Note that this type of exclusion is precisely the means through which fairness is achieved in Figure 5.2 (center).

In the context of our empirical results, we look at a more general notion of selectivity which we refer to as *soft-selectivity* defined as,

$$S(f_C, f_F) = \underbrace{\mathbb{P}\big(f_C(\mathbf{x}) = 1 \neq f_F(\mathbf{x})\big)}_{\textbf{x losses} \text{ positive classification when switching to } f_F} - \underbrace{\mathbb{P}\big(f_C(\mathbf{x}) = 0 \neq f_F(\mathbf{x})\big)}_{\textbf{x gains} \text{ positive classification when switching to } f_F}$$

Note that if Definition 5.3.2 holds, then $S(f_C, f_F) \leq 0$. In the context of our theoretical results we find that *selectivity* of f_F leads to fairness reversals; empirically we observe that high values of $S(f_C, f_F)$ leads to fairness-reversals, while low values of $S(f_C, f_F)$ do not.

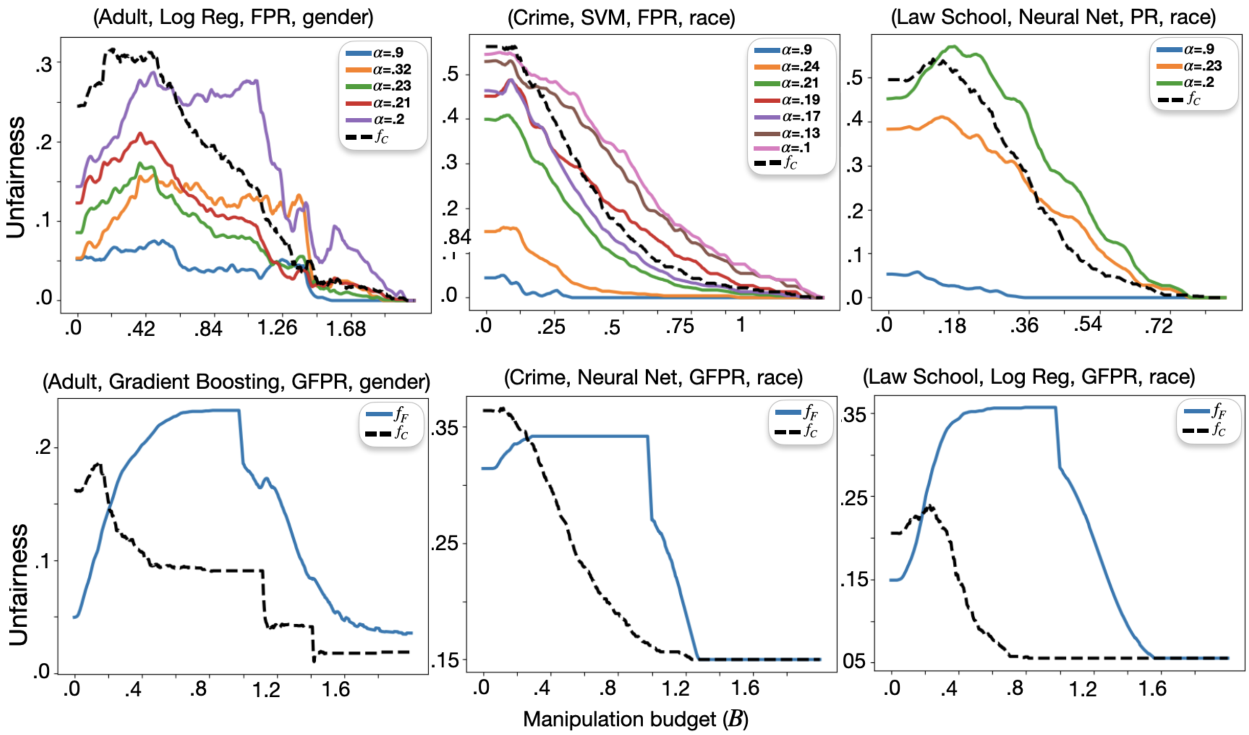

Figure 5.1: Difference in unfairness between groups on several datasets as a function of the manipulation budget B when manipulation cost is $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$. Dashed black lines correspond to f_C and colored lines correspond to f_F. A fairness reversal occurs when one of the colored lines is above the black line. The top row displays results when f_F is learned via the Reductions algorithm, with fairness defined in terms of PR, TPR, or FPR, for several different values of α. The bottom row displays results when f_F is learned via the EqOdds algorithm, with fairness defined in terms of generalized false positive rate GFPR (i.e. expected FPR). Reductions is group-agnostic, and EqOdds is group-aware.

5.4 Experimntal Results

In this section, we study phenomenon of fairness-reversals empirically, demonstrating that they are commonly observed for several benchmark datasets described in Chapter 4. Our central goal is to understand the conditions under which fairness-reversals occur as the result of strategic behavior, that is, when a fair classifier f_F becomes *less fair* than its conventional counterpart f_C when agents act strategically.

Fairness Reversals Under Strategic Behavior

In Figure 5.1 we investigate fairness reversals on three datasets with both Reductions and EqOdds fairness methods. Consider first Figure 5.1 (top), which examines settings where predictions do not take the sensitive features as an input (we call these *group-agnostic* classifiers). In these three plots, the dashed line corresponds to f_C, and the rest are group-fair classifiers f_F for different values of α (recall that higher α entails greater importance of group fairness). What we observe is that in many cases, particularly when α is not very high, there is a range of budget values B for which f_F becomes less fair than f_C. Moreover, in many cases, this range is considerable. In Figure 5.1 (bottom plots), where group-fair classifiers are *group-aware*, including the sensitive feature as an input, the fairness reversal phenomenon is even more dramatic (note that EqOdds attempts to achieve 0 unfairness between groups, and thus we do not show multiple values of α in these plots). In this experiment, when best responding agents are capable of misreporting their group as if it where a feature in $\mathbf{x}$ (fairness is still computed with true group membership). Due to the particular nature of EqOdds, specifically its handling of agents from different groups, we observe a sharp change in fairness at $B = 1$, the precise budget for which misreporting group membership is feasible.

Figure 5.1 exhibits several additional phenomena. Note, in particular, that in many cases the unfairness (i.e., FPR difference between the groups) initially *increases* as the budget increases, but in all cases as budgets B keep increasing, eventually unfairness vanishes *as a result of strategic behavior by agents*. Furthermore, much as we observe this initial unfairness increase for both f_C and f_F, it appears *amplified* for some of the group fair classifiers f_F.

What Causes Fairness Reversals?

As we formally prove below, the essential condition is *selectivity* of fair classifier f_F compared to f_C. Specifically, in binary classification, there are, roughly, two ways one can improve fairness on a given dataset (that is, without any consideration of strategic behavior); either through *inclusiveness* (positively classifying additional agents from the disadvantaged group by changing their predicted class to 1), or through *selectivity* (negative classifying some of the members of the advantaged group by changing their predicted class 1 to 0).

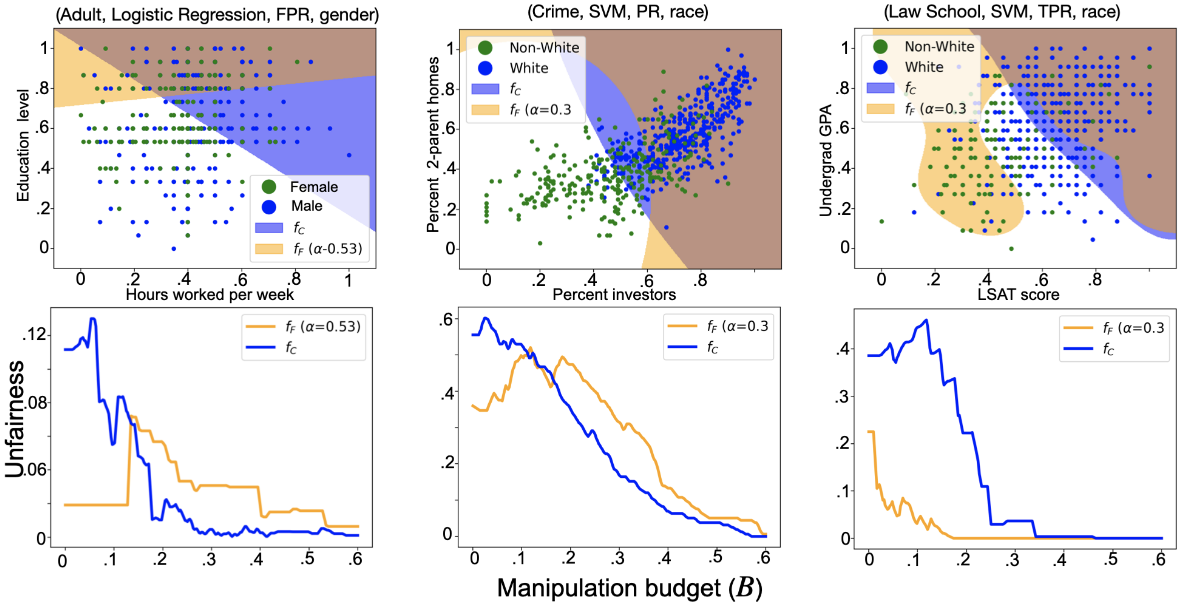

Figure 5.2: Fairness reversals and selectivity of classifiers on two ordinal features. The top row shows regions with positive predictions (blue for f_C and orange for f_F) using two features (corresponding to the axes), and dot colors correspond to the sensitive demographics. The bottom row shows the relative unfairness between demographic groups (for the classifiers shown in the top row) as a function of strategic manipulation budget B (lower means more fair).

Our key observation is that *selectivity* leads to fairness reversals, while *inclusiveness* does not. Specifically, we observe that as the number of agents positively classified under f_C, but negatively under f_F, is larger than the number of agents negatively classified by f_C, but positively under f_F, fairness reversals are more commons.

We illustrate this in Figure 5.2, which shows the decision boundaries of f_F and f_C (top row), as well as associated fairness as a function of budget (bottom row) for several combinations of dataset, classifier, and fairness definition. On the Adult and Crime datasets (first two columns), fairness is achieved predominantly through selectivity, as the orange region (f_C) includes few additional green points (disadvantaged group) compared to the blue region (f_C), but excludes many blue points (advantaged group). This is given more precisely in terms of the respective group-wise positive rates for f_C and f_F; in the first two examples the positive rates on both groups drops when switching from f_C to f_F, while in the third case the positive rate for both groups increases. This, in turn, leads to instances of fairness reversal (bottom row first column). Qualitatively similar behavior is also observed on the Crime dataset (second column). In the Law School dataset (third column), in contrast, fairness is achieved primarily through inclusiveness, and f_F remains more fair than f_C over a broad range of strategic manipulation budgets B.

The reason that selectivity leads to fairness reversal is that those from the advantaged group who are excluded tend as a result to be closer to the decision boundary than those from the disadvantaged group. In Section A of the Appendix we provide further results linking selectivity of the fair classifier to fairness reversals. In this section we also observe that when strategic agent behavior (for some manipulation budget) results in a fairness reversal between f_F and f_C, the relative accuracy of the classifiers is also reversed (for some potentially different manipulation budget), implying a fundamental relationship between fairness and accuracy when agents are strategic.

Unfairness of f_F in Isolation

Lastly we remark on the relationship of the between the manipulation budge B and the unfairness of the fair classifier f_F. As seen in Figures 5.2 and 5.1, the unfairness of f_F is frequently increasing in B (for small values of B). To provide insight into this phenomenon we look to the case of single variable prediction as showing in Figure 5.3. This figure shows the error and unfairness of a single variable classifier (i.e., a threshold classifier with threshold θ) when using a student's LSAT score to predict whether they will pass the bar exam. Since manipulations change model decisions only in a single direction (negative predictions become positive), predicting on strategically altered data amounts to predicting on unaltered data with a lower threshold . As the manipulation budget B grows, the corresponding threshold becomes increasingly smaller. Thus, when f_F is more selective than f_C, i.e. $\theta_F > \theta_C = 0.57$, the unfairness of f_F will initially increase as B increases. In the case of multivariate prediction, the increased unfairness of f_F stems from a similar

Next, we study fairness and accuracy reversals in strategic classification settings theoretically, demonstrating that selectivity is indeed a sufficient (and, under some additional qualifications, necessary) condition for fairness reversal.

5.5 Theoretical Analysis

Here we provide theoretical characterizations of the results observed in Section 5.4. In particular, we provide provide three primary types of results: 1.) selectivity of the fair classifier is a sufficient condition — and under some mild assumptions also a necessary — for fairness-reversals to occur, 2.) instances in which fairness-reversals occur are also instances in which accuracy-reversals occur, and 3.) outlining conditions on both the fairness-importance weight α and data distribution which cause the fair classifier to be more selective.

We begin with single-variable classifiers and then proceed to generalize our observations to multi-feature classifiers. Analysis of single-variable classifiers is valuable not only in building intuition, but also in that many of the results in the single-variable setting have direct extensions to the general classifiers with outcome-monotonic costs. Throughout, our key finding is that *selectivity* is in fact a sufficient condition for fairness reversal, providing a theoretical underpinning for the empirical observations above. Additionally, we investigate the underlying *causes* of fair classifiers becoming more selective, and provide conditions on the underlying distribution for this to be the case. In the cases of single variable classifiers with feature-monotonic costs and multivariable classifiers with outcome-monotonic costs, we further demonstrate that selectivity also leads to *accuracy reversals* (strategic behavior causes the fair classifier to become *more* accurate than the conventional model), and outline conditions on the underlying distribution such that selectivity is not just sufficient, but also necessary for both of these phenomena. When strategic agent behavior results in both a fairness and accuracy reversal, the functionality of both classifiers has fundamentally swapped; the accuracy driven (conventional) model f_C is no longer the most accurate model and the fairness driven (fair) model f_F is no longer the most model.

5.5.1 Single Variable Classifiers

We begin our theoretical exploration of fairness reversals with an exemplar case: a single variable threshold classifier. In this setting agents possess a single ordinal feature x. For simplicity we demonstrate our results for a continuous feature $x \in [0, 1]$, but the results hold for any ordinal feature (discrete or continuous) . Both the conventional classifier (selected for maximal accuracy) and fair classifier (selected for a weighted combination of accuracy and fairness with respect to a fairness metric M) can be expressed as a single parameter $\theta_C, \theta_F \in [0, 1]$ respectively where $f(x) = \mathbb{I}[x \geq \theta]$.

Prior to our main theoretical results, we first provide a helping lemma which demonstrates that threshold classifiers making predictions on manipulated distributions, can be expressed as threshold classifiers on the original unmanipulated distribution.

Lemma 5.5.1. *Suppose f is a threshold classifier with threshold θ, (i.e.,$f(x) = \mathbb{I}[x \geq \theta]$). Suppose further that the cost of manipulation, $c(x, x')$ is feature monotonic (i.e., monotonic in $|x - x'|$) with budget B, and an agent with true feature x can misreport any x' s.t. $c(x, x') \leq B$. Then there exists a classifier $f^{(c,B)}(x) = \mathbb{I}[x \geq \theta^{(c,B)}]$ which makes identical predictions on the true distribution $\mathcal{D}$ as f makes on the manipulated distribution $\mathcal{D}_f^{(c,B)}$, i.e. when agents behave strategically $f(x') = f^{(c,B)}(x)$ for all $x \in \mathcal{X}$. Moreover, the manipulated threshold $\theta^{(c,B)}$ acting on the true distribution $\mathcal{D}$ is given by,*

$$\begin{aligned} \theta^{(c,B)} = \arg\min_{x} & x \\ s.t. \quad & c(x, \theta) \leq B. \end{aligned}$$

Lemma 5.5.1 implies that strategic agent behavior can be examined through both the perspective of the original classifier f making predictions on the modified distribution $\mathcal{D}_f^{(c,B)}$ or a modified classifier $f^{(c,B)}$ on the original distribution $\mathcal{D}$. Since our investigation involves comparing the error and unfairness of two classifiers, f_C and f_F, the latter perspective is particularly useful given that the distribution $\mathcal{D}$ remains invariant between the two manipulated classifiers.

Proof: (Lemma 5.5.1). When all agents prefer positive predictions to negative predictions, their manipulations will change the classifier in only a single direction, namely manipulations cause negatively predicted examples to become positively predicted. Thus, only agents with feature x, where $f(x) = 0$ need be considered.

Suppose f is a threshold classifier with threshold θ, then the best response of an agent with true feature x is,

$$\begin{aligned} x^* = \arg\max_x \ & \mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta] \\ \text{s.t.} \ \ & c(x, x') \leq B \end{aligned}$$

Since the cost function $c(x, x')$ is monotone w.r.t. $|x' - x|$ the above best response has solution

$$x^* = \begin{cases} \theta & \text{if } c(x, \theta) \leq B \ \text{ and } x < \theta \\ x & \text{otherwise} \end{cases}$$

Moreover, the monotonicity of $c(x, x')$ also implies that if an agent with feature x has best response $x^* = \theta$, then any other agent with true feature x_1 where $x \leq x_1 < \theta$, also has best response $x_1^* = \theta$.

Thus, the distribution shift of $\mathcal{D}$ caused by strategic behavior, can be quantified in terms of the agent with the smallest feature which is able to misreport a value of θ as their feature, i.e.,

$$\begin{aligned} x_{\min} = \arg\min_x \ & x \\ \text{s.t.} \ & c(x, \theta) \leq B \end{aligned}$$

Thus when agents are strategic, any agent with feature $x \geq x_{\min}$ will be positively classified by f. Therefore, the threshold $\theta^{(c,B)} = x_{\min}$ makes the same classifications on the unmanipulated distribution $\mathcal{D}$ as θ makes on the manipulated distribution $\mathcal{D}_\theta^{(c,B)}$. □

Our first result is that in single-feature classification, higher selectivity of the group-fair classifier (i.e. $\theta_C < \theta_F$) is sufficient for fairness reversal.

Theorem 5.5.2. *Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x' - x|$, θ_C is the most accurate, and θ_F the optimal α-fair, threshold. If $\theta_C < \theta_F$, then there exists a budget B that leads to fairness reversal between f_F and f_C.*

Proof. The unfairness of threshold θ w.r.t. to the distribution $\mathcal{D}$ and fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$ is expressed as,

$$U_{\mathcal{D}}(\theta) = \big|\mathcal{M}_{\mathcal{D}}(\theta|g = 1) - \mathcal{M}_{\mathcal{D}}(\theta|g = 0)\big|,$$

For a given threshold θ and manipulation budget B the best response of an agent with true feature x is

$$x_\theta^{(B)} = \text{argmax}_{x'}\big(\mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]\big) \text{ s.t. } c(x, x') \leq B,$$

When agents from $\mathcal{D}$ play this optimal response, let the resulting distribution be $\mathcal{D}_\theta^{(c,B)}$. The difference in unfairness between classifiers when agents are strategic is $U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F)$. Since both f_C and f_F are thresholds, and c is feature-monotonic, the decisions of θ_C, θ_F on the modified distribution $\mathcal{D}_\theta^{(c,B)}$ can be expressed as decisions of modified thresholds $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ on the original distribution $\mathcal{D}$, i.e.,

$$U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F) = U_{\mathcal{D}}(\theta_C^{(c,B)}) - U_{\mathcal{D}}(\theta_F^{(c,B)})$$

where

$$\theta_C^{(c,B)} = \arg\min_x \text{ s.t. } c(x, \theta_C) \leq B \qquad \text{and} \qquad \theta_F^{(c,B)} = \arg\min_x \text{ s.t. } c(x, \theta_F) \leq B$$

Given these modified threshold, we see that strategic agent behavior results in a *lowering* of each threshold as more agents are now able to achieve positive classification; this is due to the fact that only negatively classified agents will behavior strategically, their goal being to achieve positive classification. Moreover, when considering $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ as functions of B, both are monotonically decreasing in B (due to the the monotonicity of c), and $\theta_C^{(c,B)} \leq \theta_F^{(c,B)}$ for all B (due to $\theta_C < \theta_F$).

Since fairness is defined in terms of PR, FPR, or TPR the constant function $f(x) = 1$ has unfairness 0 for any distribution. Thus, $\theta_C^{(c,B)} = 0$ implies $U_{\mathcal{D}}(\theta_C^{(c,B)}) = 0$. Let

$$B' = \sup\{B \in \mathbb{R}_+ : U_{\mathcal{D}}(\theta_C^{(c,B)}) > 0\},$$

then since $U_{\mathcal{D}} \geq 0$ and c is continuous, there must exist some $\varepsilon > 0$ over the interval $B \in [B' - \varepsilon, B']$ the unfairness $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is strictly decreasing in B. If

$$U_{\mathcal{D}}(\theta_F^{(c,B'-\varepsilon)}) \geq U_{\mathcal{D}}(\theta_C^{(c,B'-\varepsilon)}) > 0,$$

then a fairness reversal has already occurred for budget $B' - \varepsilon$, so assume otherwise. Combining the difference in relative fairness for budget $B' - \varepsilon$ with the fact that $\theta_C^{(c,B)} \leq \theta_F^{(c,B)}$ for all B, we get $\theta_C^{(c,B'-\varepsilon)} < \theta_F^{(c,B'-\varepsilon)}$. Since c is monotonic and continuous there must exist some budget $B_F > B' - \varepsilon$ such that $\theta_C^{(c,B'-\varepsilon)} = \theta_F^{(c,B_F)}$. Since $B_F \geq B' - \varepsilon$, and $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is decreasing for $B \geq B' - \varepsilon$, it must be the case that

$$U_{\mathcal{D}}(\theta_C^{(c,B_F)}) = U_{\mathcal{D}}(\theta_F^{(c,B'-\varepsilon)}) \leq U_{\mathcal{D}}(\theta_F^{(c,B_F)}),$$

and a fairness reversal occurs for budget B_F. □

We now turn our attention to a complementary observation: fairness reversal is accompanied by *accuracy reversal*, that is, strategic behavior leads to f_F having higher accuracy than f_C. This is primarily due to the fact that f_F becomes more selective and therefore more resilient to manipulation. Note that the fairness reversal and accuracy reversal need not occur for the same budget B.

Theorem 5.5.3. *Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x' - x|$, θ_C is the most accurate threshold, and θ_F the optimal α-fair threshold. If $\theta_C < \theta_F$, then there exists a budget B s.t. f_F is more accurate than f_C.*

Proof Sketch. The error of threshold θ on distribution $\mathcal{D}$ is given by

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{P}\big(\mathbb{I}[x \geq \theta] = y\big)$$

By definition the definition of θ_C, we have $\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta)$ for all $\theta \in [0, 1]$, i.e. $\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta_F)$. Similar to the proof of Theorem 5.5.2, agents strategically responding to threshold classifiers θ_C, θ_F can be viewed as modified thresholds $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ operating on the true distribution $\mathcal{D}$. Both $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ are monotonically decreasing in B. Moreover, $\theta_C^{(c,B)} = 0$ implies $\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{P}(y = 0)$, since the threshold classifies all agents positively.

Let

$$B' = \sup\{B : \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B)}) < \mathbb{P}(y=0)\},$$

i.e. B' is the "largest" manipulation budget such that the conventional threshold is not a trivial classifier (i.e., not making constant predictions) in the presence of strategic agent behavior. In a similar line of reasoning to the case of fairness reversals, there must exist some interval $[B'-\varepsilon, B']$ over which $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)})$ is strictly increasing. Again, by the fact that $\theta_C < \theta_F$, there must exist some $B_F > B' - \varepsilon$ such that $\theta_C^{(B'-\varepsilon)} = \theta_F^{(c,B_F)}$. Thus,

$$\mathcal{L}_{\mathcal{D}}(\theta_F^{(c,B_F)}) = \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B'-\varepsilon)}) \geq \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B_F)}),$$

implying that an accuracy reversal occurs for budget B_F. □

Unimodality and Necessary Conditions for Fairness Reversals

We have showed thus far that selectivity is *sufficient* for fairness and accuracy reversals, but under what conditions is it also *necessary*? Loosely speaking, when a feature x is a good predictor of both y and g, the error and unfairness of f_C and f_F are *unimodal* (defined next) with respect to the manipulation budget B.

Definition 5.5.4. ***(Unimodal):*** *A function* $g : [a,b] \to \mathbb{R}$ *is* negatively unimodal *(*positively unimodal*) on the interval* $[a,b]$ *if there exists an* inflection *point* $r \in [a,b]$ *such that* f *is monotone decreasing (increasing) on* $[a,r]$ *and monotone increasing (decreasing) on* $[r,b]$*. (All convex functions are* negatively unimodal *and all concave functions are* positively unimodal*).*

Unimodality is relevant to fairness and accuracy reversals as we will see that when error is negatively unimodal and unfairness is positively unimodal, both fairness and accuracy reversals occur. We empirically demonstrate that unimodality of both functions holds frequently on real data. The condition of unimodality of error and unfairness can be interpreted as both functions possessing a "sweet spot" which yields best case accuracy (or worst case unfairness). In the former, x is good predictor of true label y and in the latter x is a good predictor of g.

As an example, in Figure 5.3 we see this phenomenon occur on the Law School dataset when using a student's LSAT score as the predictive feature x. Both error and unfairness (in terms

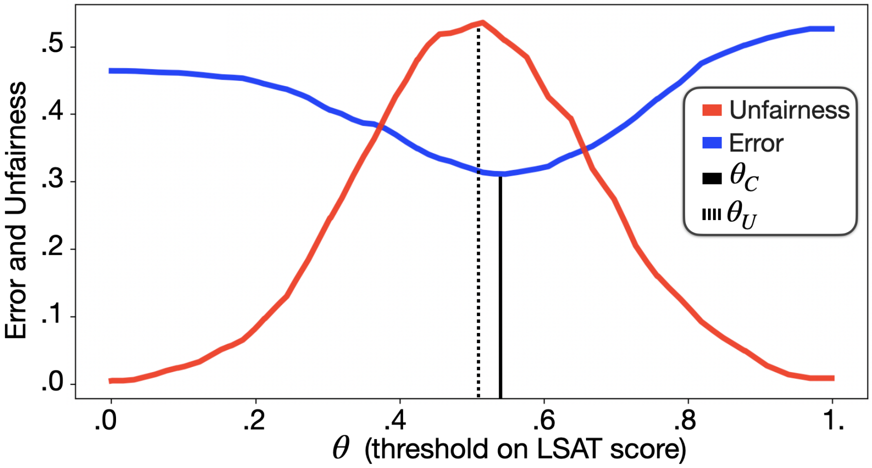

Figure 5.3: Error (blue) and PR-based unfairness between White and Non-White individuals (red) of a single variable classifier on the Law School dataset when using the student's LSAT score as a single predictive feature. All individuals with an LSAT score above the threshold θ are predicted positively. The thresholds θ_C and θ_U are the most accurate and least fair thresholds respectively.

of positive rate difference between groups) are unimodal in the threshold θ. That is, LSAT score is a "good" predictor of both the target variable (bar passage) and the sensitive feature (race); this is a well documented source of bias within the Law School dataset.

We further document this relationship in Section A of the Appendix, most ordinal features produce threshold classifiers which have (approximately) unimodal error and unfairness. In Section A, we also theoretically outline the precise conditions under which error and unfairness would be unimodal; these conditions essentially boil down to the feature x being a good predictor of both group and label (which we observe to be the case for most ordinal features across the datasets we study). When this occurs, the selectivity of f_F is not only sufficient for fairness and accuracy reversals, but also necessary. We next formalize this in the following theorem; further details on the necessary and sufficient conditions required for fairness and accuracy reversals are provided in Section A.

Theorem 5.5.5. *Let θ_C and θ_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic, and that error (and unfairness) are negatively (positively) unimodal in θ. Then there exists a budget B such that strategic agent behavior leads to a fairness reversal if an only if f_F is more selective than F_C.*

Proof. When error $\mathcal{L}_{\mathcal{D}}(\theta)$ and unfairness $U_{\mathcal{D}}(\theta)$ are both unimodal in θ, the optimal conventional threshold θ_C and optimal α-fair threshold θ_F can be expressed in terms of the inflection points $x_{\mathcal{L}}$ and x_U of error and unfairness respectively. The most accurate threshold is then $\theta_C = x_{\mathcal{L}}$, and the most unfair threshold is then $\theta_u = x_U$. The forward direction, i.e. when $\theta_C < \theta_F$, follows a similar of reasoning to the proof of Theorem 5.5.2, let $\theta_C^{(c,B)}$ and $\theta_F^{(c,B)}$ be the modified thresholds induced by agents best responding to either threshold with cost function c and budget B. Then, since $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ are monotonically decreasing in B and $\theta_C < \theta_F$, there must exist a B' such that $\theta_C^{(c,B')} \leq \theta_F^{(c,B')} = \theta_C$. Thus

$$U_{\mathcal{D}}(\theta_C^{(c,B')}) \leq U_{\mathcal{D}}(\theta_C) = U_{\mathcal{D}}(\theta_F^{(c,B')}) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\theta_F^{(c,B')}) = \mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B')}),$$

implying that a fairness and accuracy reversal occurs for budget B'.

The reverse direction, follows from the relationship between θ_F and the two inflection points θ_C, θ_U. Given the assumption that $\theta_F < \theta_C$, there are only three possible cases for the relationship between these points

$$(1)\ \theta_F < \theta_C \leq \theta_U, \quad (2)\ \theta_F < \theta_U \leq \theta_C, \quad (3)\ \theta_U < \theta_F < \theta_C$$

the strict inequalities being due to the fact that $\theta_F \neq \theta_C$ and $\theta_F \neq \theta_U$ by definition. In cases (1) and (2), no fairness or accuracy reversal will occur. Only in case (3) can a fairness or accuracy reversal occur, however we will show that such a case is impossible.

Beginning with case (1), both error and unfairness are unimodal in $\theta_F^{(c,B)}, \theta_C^{(cB)}$, each of which is monotonically increasing in B. Thus if $\theta_F < \theta_C$, then no accuracy reversal can occur. Similarly if $\theta_F < \theta_C \leq \theta_U$, no fairness reversal can occur, i.e. in case (1), neither reversal can occur.

In case (2) since $U_{\mathcal{D}}(\theta_F) < U_{\mathcal{D}}(\theta_C)$, and $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is monotonically increasing until $\theta_C^{(c,B)} = \theta_U$, no fairness reversal will occur. Similar to case (1), $\theta_F < \theta_C$, implies that no accuracy reversal occurs either.

Thus it remains only to show that case (3) can never occur. To see this, not for any $0 < \varepsilon < \theta_C - \theta_F$, it must be the case that both

$$U_{\mathcal{D}}(\theta_F + \varepsilon) \leq U_{\mathcal{D}}(\theta_F) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\theta_F + \varepsilon) \leq \mathcal{L}_{\mathcal{D}}(\theta_F)$$

implying that θ_F is not the optimal fair threshold. Thus case (3) is not possible, and in cases (1) and (2), i.e., the only possible cases, fairness and accuracy reversals can never occur. □

Now that we have established the critical role of selectivity in fairness reversal, we next analyze *why* that is. As mentioned previously, there are roughly two ways to achieve fairness: *inclusiveness* (classifying more examples as positive) or *selectivity* (classifying fewer examples as positive). Which of these will be the predominant outcome of training f_F depends intimately on the data distribution. We outline these conditions, as well as conditions for error and unfairness to be unimodal, via Lemmas A.1.2, A.1.3, A.1.1. These lemmas can be succinctly, stated as follows.

Remark 5.5.6. *(Summary of Lemmas A.1.2, A.1.3, A.1.1) Suppose that for the true conditional distributions $\mathbb{P}(y = 1|x)$ and $\mathbb{P}(g = 1|x)$ there exists some x_y, x_g such that any threshold $\theta \geq x_y$ has accuracy at least .5 on $\mathbb{P}(y = 1|x)$ and any $\theta \leq x_y$ has accuracy no more than .5 on $\mathbb{P}(y = 1|x)$ (similarly for x_g). Then error and unfairness are unimodal in the threshold θ. That is, x is a reasonably good predictor of both true label y and group g. This can be equivalently stated that $\mathbb{P}(y = 1|x)$ and $\mathbb{P}(g = 1|x)$ both have a* single crossing *with .5 as functions of x. This condition frequently holds on real data.*

Next, Theorem 5.5.7 provides conditions on the underlying distribution such that the optimal fair classifier will achieve fairness via selectivity.

Theorem 5.5.7. *Suppose fairness is defined by PR, TPR, or FPR. Suppose further that $\mathbb{P}\big(y = 1|x\big)$ has a single crossing with $\mathbb{P}(y = 1)$, and that $\mathbb{P}\big(g = 1|x\big)$ has a single crossing with the respective value given in Lemmas A.1.2 and A.1.3, call this value p_g. Let x_y and x_g be defined by*

$$\mathbb{P}\big(y = 1|x_y\big) = \mathbb{P}(y = 1) \quad \textit{and} \quad \mathbb{P}\big(g = 1|x_g\big) = p_g$$

If $x_g < x_y$, then there exists a nonempty interval $[\alpha_0, \alpha_1]$ s.t. for any $\alpha \in [\alpha_0, \alpha_1]$ the optimal α-fair classifier f_F, has the propriety that $\theta_C < \theta_F$ (implying that strategic agent behavior leads to f_F becoming less fair than f_C as outlined by Theorem 5.5.2).

Proof. Given $\alpha \in (0, 1)$, fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$, and data distribution $\mathcal{D}$, the objective of the fair learning scheme is to find θ_F such that

$$\theta_F = \text{argmin}_\theta (1-\alpha)\mathbb{P}\big(\mathbb{I}[x \geq \theta] \neq y\big) + \alpha U_\mathcal{D}(\theta) \tag{5.4}$$

where

$$U_\mathcal{D}(\theta) = \big|\mathcal{M}(\theta|g=1) - \mathcal{M}(\theta|g=0)\big|$$

By Lemma A.1.1 the error term $\mathbb{P}\big(\mathbb{I}[x \leq \theta] = y\big)$ is negatively unimodal in θ and achieves a minimum at θ_C where $\mathbb{P}\big(y = 1|x = \theta_C\big) = \mathbb{P}(y = 1)$. Similarly, by Lemmas A.1.2, A.1.3 the unfairness term $U_\mathcal{D}(\theta)$ is positively unimodal in θ and achieves a maximum at θ_U where $\mathbb{P}\big(g = 1|x = \theta_U\big) = \mathbb{P}\big(g = 1\big)$. Thus for any α the fair learning objective (Equation 5.4) is monotonically increasing, implying $\theta_F \notin [\theta_U, \theta_C]$. So either $\theta_F \in [0, \theta_U)$ or $\theta_F \in (\theta_C, 1]$. By the continuity of Equation 5.4 w.r.t. to θ. For some $\varepsilon > 0$, any $\varepsilon' < \varepsilon$ yields $\mathbb{P}\big(\mathbb{I}[x \geq \theta_C + \varepsilon'] \neq y\big) \leq \mathbb{P}\big(\mathbb{I}[x \geq \theta_U] \neq y\big)$ and $U_\mathcal{D}(\theta_C) \leq U_\mathcal{D}(\theta_U - \varepsilon')$. Thus implying that for small enough since both $|\theta_C - \theta_F|$ and $|\theta_U - \theta_F|$ are monotonic w.r.t. to α, it must be the case that that for α small enough $\theta_F = \theta_C + \varepsilon'$ is the optimal α-fair classifier, thus implying the existence of of fairness coefficients $[\alpha_1, \alpha_2]$ such that $\alpha \in [\alpha_1, \alpha_2]$ implies $\theta_F > \theta_C$. □

The condition in this theorem can be intuitively interpreted as follows. Suppose that S is the set of individuals selected (i.e., classified as 1) by f_C, who are also near the decision boundary of f_C. If the advantaged group (i.e., group with better average outcomes) is overrepresented in S, there is a range of parameters α such that the optimal α-fair classifier is more selective than f_C (recall that higher α places greater importance on group-fairness in learning).

5.5.2 General Classifiers

Next we discuss general multi-variate classifiers, generalizing several of the results from Section 5.5.1. First we show that when f_F is more selective than f_C, fairness reversal occurs for both feature-monotonic and outcome-monotonic cost functions. Second, we give conditions which lead to f_F being more selective than f_C. For outcome-monotonic costs, we provide two additional results: 1) greater selectivity of f_F also leads to accuracy reversal, and

2) unimodality of each classifier's error and unfairness causes selectivity to be both necessary and sufficient for fairness and accuracy reversal.

Outcome-Monotonic Costs

We begin with the case of outcome-monotonic costs. As shown by [88], outcome-monotonic manipulation costs result in the following best response for classifier f. Let

$$\mathbf{x}^* = \mathrm{argmin}_{\mathbf{x}}\ \mathbb{P}(y=1|\mathbf{x}) \quad \text{s.t. } f(\mathbf{x}) = 1.$$

If $c(\mathbf{x}, \mathbf{x}^*) \leq B$ then the best response is $\mathbf{x}' = \mathbf{x}^*$ otherwise $\mathbf{x}' = \mathbf{x}$. With this best response in hand we show that f_F having greater selectivity than f_C leads to fairness reversal.

Theorem 5.5.8. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic. Then if f_F is more selective than f_C, there exists a budget B such that strategic agent behavior leads to a fairness reversal.*

Proof. Intuitively, the case of outcome-monotonic costs with general classifiers follows via a similar line of reasoning to that of feature-monotonic costs with single variable classifier.

For a given classifier f, let

$$p_{\min} = \min_{\mathbf{x}: f(\mathbf{x})=1} \mathbb{P}(y=1|\mathbf{x})$$

and let $\mathbf{x}_{\min}$ be the feature associated with $p_{\min}$; $\mathbf{x}_{\min,C}, p_{\min,C}$ and $\mathbf{x}_{\min,F}, p_{\min,F}$ correspond to f_C and f_F respectively. When agents best respond to f the resulting manipulated classifier can be expressed as a threshold on the underlying probabilities $\mathbb{P}(y=1|\mathbf{x})$. More specifically, let

$$\mathbf{x}^* = \mathrm{argmin}_{\mathbf{x}} \mathbb{P}(y=1|\mathbf{x}) \quad \text{s.t.} \quad c(\mathbf{x}, \mathbf{x}_{\min}) \leq B.$$

Then when agents best resound to f (inducing classifier $f^{(c,B)}$) any agent $\mathbf{x}$ with $\mathbb{P}(y=1|\mathbf{x}) \geq \mathbb{P}(y=1|\mathbf{x}^*)$ will be positively classified under $f^{(c,B)}$, i.e.

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|\mathbf{x}) \geq \mathbb{P}(y=1|\mathbf{x}^*) \\ 0 & \text{otherwise} \end{cases}$$

Thus $f^{(c,B)}$ can expressed as the threshold $\mathbb{P}(y = 1|\mathbf{x}^*)$ operating on the conditional distribution $\mathbb{P}(y = 1|\mathbf{x})$.

Since f_F is more selective than f_C, (i.e. $f_F(\mathbf{x}) = 1 \implies f_C(\mathbf{x}) = 1$), it must be the case that

$$f_F(\mathbf{x}_{\min,F}) = 1 = f_C(\mathbf{x}_{\min,F}) \quad \text{implying that} \quad p_{\min,C} \leq p_{\min,F}$$

Therefore the induced conventional and fair thresholds $\mathbb{P}(y = 1|\mathbf{x}^*_C), \mathbb{P}(y = 1|\mathbf{x}^*_F)$ acting on $\mathbb{P}(y = 1|\mathbf{x})$ have the relationship that $\mathbb{P}(y = 1|\mathbf{x}^*_C) \leq \mathbb{P}(y = 1|\mathbf{x}^*_C)$. Thus, we see that selectivity of the fair classifier in the case outcome-monotonic costs yields a fair threshold (on a modified distribution) which is larger than the induced conventional threshold (operating on the same distribution as the fair threshold). While this setting is not entirely equivalent to the single variable case, the remainder of the proof follows in similar fashion to that of Theorem 5.5.2. In particular, the monotonicity of $\mathbb{P}(y = 1|\mathbf{x}^*)$, as a function of B, implies $\mathbb{P}(y = 1|\mathbf{x}^*_C) \leq \mathbb{P}(y = 1|\mathbf{x}^*_F)$ for any B, which in turn implies the existence of a budget interval over which the unfairness of $f_C^{(c,B)}$ decreases below $f_F^{(c,B)}$, thus resulting in a fairness reversal. □

Similar to the single-variable case, selectivity also result in accuracy reversal.

Theorem 5.5.9. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome-monotonic. Then if f_F is more selective than f_C, then there exists a budget B under which f_F becomes more accurate than f_C.*

Proof. This proof follows via a combination of Theorem 5.5.8 and Theorem 5.5.2. In particular using the approach presented in 5.5.8, both f_C and f_F can again be expressed as threshold classifiers on the underlying conditional distribution $\mathbb{P}\big(y = 1|\mathbf{x}\big)$. From here, accuracy reversals for single variable classifiers in Theorem 5.5.8 imply accuracy reversals of these induced threshold classifiers acting on the conditional distribution. □

Before outlining settings in which selectivity is not only sufficient but also necessary for fairness and accuracy reversals to occur, we first remark on the connection between selectivity, accuracy, and fairness. As previously noted, errors caused by strategic agent behavior are single-directional in the sense that manipulation can only induce false positive errors. As such, classifiers which are more selective are thus more robust to manipulation than their less selective counterparts. Generally speaking, this implies that for some range of manipulation budgets, a model that is more selective than the accuracy-maximizing model f_C will increase in its performative ability compared to f_C. As the performative ability of most classifiers on biased datasets is naturally tied with unfairness, the unfairness of the more robust model (more selective model) will likewise increase. Thus, we see a fundamental, albeit not necessarily universal, connection between selectivity (which in turn increases robustness) and model unfairness (which is increasing in model performance).

We next discuss unimodality in the context of outcome-monotonic costs. Empirically we observe that when costs are outcome-monotonic, the majority of classifiers tend to have error and unfairness which is (approximately) unimodal with respect to the manipulation budget B. When this occurs, selectivity of f_F becomes both necessary and sufficient, as outlined in the next theorem.

Theorem 5.5.10. *Let f_C and f_F the optimal conventional and fair classifiers respectively. Suppose fairness is defined in terms of PR, TPR, or FPR fairness, and $c(x, x')$ is outcome-monotonic. When error (and unfairness) are negatively (positively) unimodal with respect to the manipulation budget B, a fairness and accuracy reversal will occur between f_F and f_C if and only if f_F is more selective than f_C (each reversal may occur at different budgets B) .*

Proof. Similar to previous results involving outcome-monotonic costs, we can use results from single-variable classifiers to do much of the heavy-lifting. The intuition for this proof follows from the single variable case of Theorem 5.5.5. As shown in the proof of Theorem 5.5.8 when agents best respond to classifier f, the decisions of f can be expressed as threshold classifier acting on the conditional probability $\mathbb{P}(y = 1|\mathbf{x})$ of the original distribution $\mathcal{D}$, namely

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \ \mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1|\mathbf{x}^*) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{x}^*$ is determined by the cost function c and budget B. Since $\mathbb{P}(y = 1|\mathbf{x}^*)$ is monotonically decreasing in B, we recover a setting similar to 5.5.5, in which the forward direction of the claim holds from the fact that $\mathbb{P}(y = 1|\mathbf{x}^*_C) \leq \mathbb{P}(y = 1|\mathbf{x}^*_F)$. The reverse direction holds due to the fact that when $\mathbf{x}^*_F \leq \mathbf{x}^*_C$, unfairness is monotonically decreasing for both classifiers. □

Remark 5.5.11. *To better contextualize unimodality of error and unfairness with respect to the manipulation budget B, we can view this condition in terms of the calibration of the score function h of the classifier f. As is typical, classifiers are defined via thresholds on their underlying score functions, i.e. $f(\mathbf{x}) = \mathbb{I}[h(\mathbf{x}) \geq \theta]$. Suppose that h is reasonably well calibrated, then for every $p \in [0, 1]$, $\mathbb{P}(y = 1|h(\mathbf{x}) = p) \approx p$, i.e. $h(\mathbf{x})$ is a good approximation of the conditional distribution given by $\mathbb{P}(y = 1|\mathbf{x})$. When h is reasonably well calibrated, the condition that error and unfairness are unimodal w.r.t. to the manipulation budget B is equivalent to the error and unfairness of f being unimodal w.r.t. to the choice of threshold θ. Through this lens, one can see that the assumption of unimodality is likely to hold (at least approximately so) in practice as it is typically the case there is one "good" choice of threshold θ and any deviation (increasing or decreasing θ) results in strictly worse performance of f.*

Feature-Monotonic Costs

Laslty, we demonstrate that selectivity remains sufficient for fairness reversal in general when costs are feature-monotonic.

Theorem 5.5.12. *Let f_C and f_F be the most accurate and the optimal α-fair classifier, respectively. Suppose fairness is defined by PR, FPR, or TPR and $c(\mathbf{x}, \mathbf{x}')$ is feature-monotonic. If f_F is more selective than f_C, then there exists a budget B that leads to fairness reversal between f_F and f_C.*

Proof Sketch. The full proof is deferred to Section A of the Appendix and follows similar to the cases of outcome-monotonic costs. The intuition behind this results is that trivial classifiers (i.e., those that predict $f(\mathbf{x}) = 1$ for all $\mathbf{x}$) have 0 unfairness for PR, FPR, and TPR based fairness. As B increases, both $f_C^{(c,B)}$ and $f_F^{(c,B)}$ (the classifiers resulting from agents best responding to either classifier with budget B and cost function c) will approach 0 unfairness, not necessarily monotonically, as they become more like trivial classifiers. At some point prior to reaching trivial classification, the conventional classifier f_C will be at least as fair as f_F. This can be seen through a combination of the fact that f_F is more selective than f_C and the way in which manipulations alter the positively predicted region of a classifier when costs are feature-monotonic. In particular, f_F being more selective than f_C implies that,

$$\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}.$$

Feature-monotonic cost functions preserve this subset propriety under manipulation, i.e., for any B,

$$\{\mathbf{x} \in \mathcal{X} : f_F^{(c,B)}(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C^{(c,B)}(\mathbf{x}) = 1\}.$$

Thus f_F is *always* more selective than f_C, regardless of the manipulation budget B. As such, the positive rate of f_F will never exceed the positive rate of f_C, implying that $f_F^{(c,B)}$ approaches a trivial classifier more "slowly" than $f_C^{(c,B)}$, with respect to B. Moreover, prior to approaching triviality $f_F^{(c,B)}$ will effectively approach f_C, thus partially absorbing some of the original unfairness of f_C, resulting in a fairness reversal. □

Next, we provide a condition which leads f_F to be more selective than f_C. Here, we provide this condition for the PR fairness metric; analogous results for TPR and FPR are given in Section A of the Appendix. For this result, we define the following notation

$$P_{G_z} = \mathbb{P}(g = z), \quad g(\mathbf{x}) = P(g = 1|\mathbf{x})$$

and

$$\mathcal{X}_0 = \{\mathbf{x} \in \mathcal{X} : g(x) < P_{G_1} \text{ and } \mathbb{P}(y = 1|\mathbf{x}) < 1/2\}.$$

The set $\mathcal{X}_0$ represents the set of features which are less likely than chance to correspond to $g = 0$ and $y = 0$.

Theorem 5.5.13. *Let f_C and f_F be the most accurate and optimal α-fair classifiers respectively, and fairness defined by PR. Then f_F is more selective than f_C if and only if $0 < \alpha \leq \alpha^*$, where*

$$\alpha^* = \min_{\mathbf{x} \in \mathcal{X}_0} \frac{P_{G_0} P_{G_1}(2\mathbb{P}(y = 1|\mathbf{x}) - 1)}{g(\mathbf{x}) + P_{G_1}\big(P_{G_1} - 2g(\mathbf{x}) - 2P_{G_1}\mathbb{P}(y = 1|\mathbf{x})\big)}.$$

Proof. Both the conventional and fair objectives can be written as follows:

$$\begin{aligned} f_C =& \operatorname{argmin}_f \mathbb{P}(f(\mathbf{x}) \neq y) \\ f_F =& \operatorname{argmin}_f (1-\alpha)\mathbb{P}(f(\mathbf{x}) \neq y) + \alpha\big|\mathbb{P}(f(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f(x) = 1|g = 0)\big| \end{aligned}$$

Assuming the optimal f_F has higher positive rate for group 1 (the group 0 holds symmetrically), the fair objective function can be simplified to,

$$(1-\alpha)\sum_{\mathbf{x}\in\mathcal{X}}\big((1-f(\mathbf{x}))\mathbb{P}(y=1|\mathbf{x})+f(\mathbf{x})\mathbb{P}(y=0|\mathbf{x})\big)\mathbb{P}(\mathbf{x})$$
$$+\alpha\sum_{\mathbf{x}\in\mathcal{X}}f(\mathbf{x})\left(\frac{\mathbb{P}(g=1|\mathbf{x})}{\mathbb{P}(g=1)}-\frac{\mathbb{P}(g=0|\mathbf{x})}{\mathbb{P}(g=0)}\right)\mathbb{P}(\mathbf{x})$$

Thus $f_F(\mathbf{x})=1$ is optimal if

$$\alpha\frac{(\mathbb{P}(g=1|\mathbf{x})+(\mathbb{P}(g=1)-2)\mathbb{P}(g=1))}{(1-\mathbb{P}(g=1))\mathbb{P}(g=1)}-(1-\alpha)2\mathbb{P}(y=1|\mathbf{x})+1\geq 0 \tag{5.5}$$

and $f_C(\mathbf{x})=1$ is optimal if $\mathbb{P}(y=1|\mathbf{x})\geq\mathbb{P}(y=1)$. Thus, the only case in which f_F positively classifies an example $\mathbf{x}$, which is negatively classified by f_C (i.e., $f_f(\mathbf{x})=1\neq f_C(\mathcal{X})=0$), is when the left-hand side of Inequality 5.5 is nonnegative and $\mathbb{P}(y=1|\mathbf{x})\geq\mathbb{P}(y=1)$. Simplifying the condition in Equation 5.5 yields α^*. □

The key observation from Theorem 5.5.13 is that fairness reversal is a *small-α* phenomenon. This may seem surprising, since f_F is likely to be most similar to f_C for smaller values of α (in particular, the two are identical when $\alpha=0$). However, when α is high, the fairness term is sufficiently dominant that reversals are unlikely. Consequently, it is precisely the intermediate values of α, where we aspire to preserve high accuracy while improving group-fairness that are most susceptible to fairness reversal. This is indeed consistent with our empirical observations in Section 5.4, which indicate that for intermediate values of α fairness reversals are not only more common, but occur with greater magnitude. Lastly, note that for some distributions, $\alpha^*\leq 0$, which means that fairness reversals are luckily not guaranteed.

5.6 Discussion

We demonstrate a fairness-reversal phenomenon, where a trained-to-be fair classifier exhibits more unfairness than a conventional accuracy-maximizing classifier if human agents can strategically respond to a classifier. We show that a sufficient condition for observing fairness reversal is “selectivity”, that is, a group-fair classifier making fewer positive predictions than its conventional counterpart. Additionally, we demonstrated that this condition of “selectivity” also results in an accuracy reversal. The aggregate of these results indicates that when fairness is achieved through an overall decrease in positive rate (compared to the conventional classifier), strategic agent behavior can lead to a reversal of the core functionality of both models (i.e., the performance based model becomes less accurate than the fair model, and the fair model becomes less fair than the fairness-agnostic model).

Further, fairness reversals are accompanied by accuracy reversals in most cases, as demonstrated by both our empirical and theoretical results. In cases where the fair classifier is more selective than its conventional counterpart, accuracy reversals can be viewed as an indication that the fair classifier is more robust to manipulation. As such, these findings also indicate a fundamental trade-off between fairness and robustness to strategic manipulation. Classifiers which are more selective are in turn more robustness under manipulation, but are likewise less fair under that same manipulation.

These results are not as a critique of fair-learning, but rather as a caution towards the expectation of fairness guarantees when a fair classifier sees real-world deployment. The successful deployment of fair-learning models requires the consideration of many nuanced factors, strategic agent response to model choice being on such consideration. While we have outlined several necessary and sufficient conditions regarding both classifier selectivity as well as fairness reversals, a more nuanced understanding of when fair classifiers may suffer from such problems in cases where the classifier s designed to anticipate strategic behavior. Mitigating fairness and accuracy reversals is an important direction for future work.

Chapter 6

Group Fairness and Individual Welfare

As discussed in Chapter 5 we observed that strategic agent behavior can negatively impact both the fairness and performance of group-fair models. In this chapter we will investigate the complementary problem of the ways in which group-fair learning can negatively impact individuals. In particular we will examine the ways in which the imposition of group-fairness can result in decreases to individual welfare in both the case of classification and scarce resource allocation. Individual welfare is measured with respect to agents gaining a desired outcome, i.e., positive classification or allocation of a resource. Similar to the case of fairness-reversals discussed in Part II, we focus on the way in which individual welfare changes which switching from a conventional (fairness agnostic) model, to a fairness-aware model.

To capture these changes in individual welfare we propose two definitions: *popularity* which measures the fraction of a population which does *not* suffer a decrease to likelihood of receiving a desired outcome when switching from a conventional to a fair model, and *impact* which measures the likelihood that a given individuals losses a desired outcome when switching from a conventional to a fair model while accounting for the possibility that any change in model choice may result in different decisions among the population (i.e., impacts captures whether the loss of a desired outcome was due to the addition of fairness rather than simply due to classifier change). These two definitions can be though of as complementary to one another in the sense that popularity captures the degree to which a population is made worse off by the addition of fairness constraints, while impact captures the level of certainty that fairness is indeed the driving force behind individuals being made worse off.

6.1 Summary of Results

We begin by analyzing the popularity of current state-of-the-art fair learning schemes and demonstrate that fair models are frequently *unpopular* across several different domains and model types; i.e., roughly half of the population experiences a decrease in their individual welfare when switching from a conventional model to a fair model. Further, we demonstrate that popularity is not uniformly distributed among groups; members of the disadvantaged group prefer fair models at rates much higher than those of the advantaged group. Disadvantaged groups often comprise a significantly smaller fraction of the total population compared to advantaged groups. As such, even a model that is unanimously popular among the disadvantaged group can have low popularity across the entire population. This observation can be thought of as an example of the *tyranny of the majority*.

We then move on to examine the impact in the case of scarce resource allocation and demonstrate that nontrivial fractions of the population experience impact, even when accounting for loss in individual welfare that could result from *any* change in classifier. This suggests that it is indeed the imposition of group fairness that leads to decreases in individual welfare. As was the case with popularity, impact is also distributed disparately among groups. While we observe that impact most commonly falls on the advantaged group, we outline several cases in which the disadvantaged group suffers the majority of impact. Lastly, we examine the relationship between resource availability and impact. Surprisingly, we find that disadvantaged groups suffer higher rates of impact when resources are more abundant.

6.2 Additional Related Work

Several recent papers look at the potential negative consequences of applying group fairness [76, 123, 27, 61, 9]. In particular [76, 9] demonstrate that specific types of group equity can be decreased by the use of fair algorithms. Others have merged notions of welfare and fairness [45, 29, 115]. Both the notion of popularity, as well as our proposed techniques for satisfying popularity and fairness (introduced in Chapter 8), differ from these lines of work in that popularity casts welfare in terms of the fraction of a population which prefers a fair model compared with a fairness-agnostic model. While the idea of agent preference over models has received some recent attention [108] (which aims to classify a population using multiple models such that each agent prefers their assigned model over all others), popularity in the context of group fair learning has remained unexplored thus far.

Additional we note the difference between individual-fairness [35, 33, 93, 89] which captures the idea that similar individuals should receive similar treatment from a given classifier (e.g., classifiers which are Lipschitz continuous are often considered individually fair). Our notions of individual welfare differ in that they capture the idea that a *given* individual (or population) should be treated at least as well under a fair model as compared to a conventional model. In this, individual-fairness compares similar individuals under a fixed classifier, while individual-welfare compares similar classifiers acting on a fixed individual.

6.3 Preliminaries

Similar to the case of fairness reversals discussed previously, we are interested in how a change from a conventional learning scheme, to a fair learning scheme, affects individuals. In particular, we consider the situation in which a conventional learning scheme C is initially in place, and a *principal* considers a switch from C to a group-fair scheme F, and wishes to ensure that F is γ-*popular* in the sense that it is preferred to C by at least a fraction γ of the target population. We formalize preference over learning schemes by assuming that an individual prefers schemes which yield higher expected outcomes for them, that is, they prefer being selected to not being selected, as in [42]. Thus, an individual i with features $\mathbf{x}_i$ prefers F over C if

$$f_C(\mathbf{x}_i) \leq f_F(\mathbf{x}_i) \text{ or } I_{C,i}(\mathcal{X}, h, k) \leq I_{F,i}(\mathcal{X}, h, k) \tag{6.1}$$

when decisions are deterministic and

$$\mathbb{E}\big[f_C(\mathbf{x}_i)\big] \leq \mathbb{E}\big[f_F(\mathbf{x}_i)\big] \quad \text{or} \quad \mathbb{E}\big[I_{C,i}(\mathcal{X}, h, k)\big] \leq \mathbb{E}\big[I_{F,i}(\mathcal{X}, h, k)\big] \tag{6.2}$$

when decisions are stochastic.

Note that our analysis is in the space of outcomes, rather than scores. Consequently, if decisions are deterministic, either in classification or allocation settings, agents only have a definitive preference over scores produced by h if this is consequential to outcomes (e.g., pushing them above or below θ). In the stochastic case, on the other hand, agents prefer the classifier or allocation scheme which yields the higher expected outcome (that is, higher probability of being selected). Armed with this model of individual preference, we now define what it means for F to be popular.

Definition 6.3.1. ***(γ-popularity)****: A learning scheme F is said to be γ-popular with respect to a population $(\mathcal{X}, Y, G)$ and conventional scheme C, if Condition (6.1) (for deterministic models), or Condition (6.2) (for randomized models), holds for at least $\gamma|\mathcal{X}|$ individuals.*

Popularity thus captures the fraction γ of a population which is weakly better off (or, equivalently, *not* made worse) from the use of F over C. Similar to the concept of β-fairness, in which a model designer can specify the desired level of fairness β, the definition of popularity, as well as our postprocessing techniques described later, allow the model designer to *directly* specify, and control, the desired level of popularity. Note that we do not capture the *degree* to which individuals are made better or worse off as a result of switching from C to F, but only *whether* they are.

Example 6.3.2. *To illustrate the relationship between fairness, accuracy, and popularity, consider the following example. Let G_1 and G_0 have four and two members respectively, with true labels $\langle 1, 1, 1, 1\rangle$ and $\langle 0, 0\rangle$. A randomized conventional classifier f_C, predicts each member of G_1 to be positive with probability 0.75 and each member of G_0 to be positive with probability 0.25. Under demographic parity fairness, G_1 is advantaged as this group has a positive rate 0.5 greater than that of G_0. Consider two choices for a fair model. f_{F_1} predicts members of G_1 to be positive with probability 0.75 and members of G_0 to be positive with probability 0.55. f_{F_2} predicts one member of G_1 to be positive with probability 1, and the others with probability $\frac{2}{3}$; it predicts one member of G_0 to be positive with probability 1 and the other with probability 0.1. Note that both models have identical accuracy and unfairness, namely $.65$ and $.2$ respectively. However, f_{F_1} has* not *decreased the score of any agent in the population; all six prefer f_{F_1} at least as much as the original f_C. In contrast, f_{F_2} has decreased the scores of three agents from G_1 and one agent from G_0; only two agents prefer f_{F_2} at least as much as f_C. This example illustrates that popularity should be viewed as a different axis than either accuracy or fairness, and there may be space to innovate by enabling popularity comparisons among fair(er) models.*

As mentioned earlier, our setting is one of a concrete choice by a principal between a particular conventional approach C and a particular group-fair approach F. This reflects a decision by the principal to switch from C—which is currently deployed—to F in order to reduce impact to a disadvantaged group (or groups). Of course, different pairs of C and F (e.g., using different loss functions, different learning algorithms, etc) would yield different judgments about popularity of F, which is, by construction, relative to C. Consequently, these will also yield different decisions about improving popularity of F based on algorithms we discuss below. Nevertheless, our framework generalizes immediately to a setting in which neither C nor F are fixed, and there is uncertain about either, or both. In such a case, we treat uncertainty about either C or F as a distribution over approaches and, consequently, over outcomes induced. This can then be immediately captured within our framework dealing with randomized schemes, and all definitions above, and technical results below, go through unchanged.

In addition to classifier popularity, we also examine individual impact Both types of impact are presented concisely in the following definition.

Definition 6.3.3. (Individual Impact): *Let f_C and f_F be a conventional and fair classifier respectively, let $I_i(f)$ be a binary indicator of agent i receiving a resource under f, let $\theta \in [0,1]$, and let $p_i = \mathbb{P}_{f'_C}\big(I_i(f'_C) = 1\big)$ be the probability that agent i receives a resource under f_F, then perceived and realized impact are defined as,*

- *Perceived Impact:* $\mathbb{I}\big[I_i(f_C) \text{ and not } I_i(f_F)\big]$,
- *Realized Impact:* $\mathbb{I}\big[I_i(f_C) \text{ and not } I_i(f_F) \text{ and } p_i \geq \theta\big]$.

As noted previously, impact and popularity can be thought of as measuring a similar quantity, namely individual-level harms, from two alternate perspectives. Popularity examines individual-level harms from a model-centric perspective, and measures the fraction of the population which will *not* be negatively impacted by a change in model (in our case, changing from f_C to f_F). On the other-hand, impact (particularly realized impact) examines individual-level harms from an individual-centric perspective, and can be thought of as capturing the confidence that an decrease to a given individual's utility was the result of switching to a fair classifier, rather than simply switching to any other type of classifier. In realized impact, the parameter θ captures this confidence, namely in that an individual must have probability at least θ to receive to the resource under any reasonable alternative of conventional classifier.

We further contextualize impact and popularity in relationship to individual fairness [35]; both metrics can be viewed as complimentary to individual fairness. Individual fairness asks “are similar individuals treated similarly under the same model?”, whereas impact and popularity both ask “is the same individual treated differently under different models?” That is, individual fairness considers different individuals interacting with the same model, while impact and popularity consider the same individual interacting with different models. As alluded to previously, the parameter θ provides those wishing to measure impact a means of tuning the sensitivity to background variation. Specifically, when a practitioner sets the value of θ, they are imposing that an individual must receive a resource under θ-fraction of alternate baseline classifiers (the possible f_C') in order to be impacted. For example setting $\theta = 0.75$ implies that an individual must receive a resource 75% of the time under baseline classifiers in order to be impacted.

To better understand the components of impact, we present a simple example of how this definition is opperationalized in practice.

Example 6.3.4. *Imagine that a company is selecting among applicants to fill a position. The company, which previously used conventional models to make hiring decisions, decides to switch to a fair model, f_F. Suppose that candidate i is not selected to receive the position under f_F, but would have received it under one possible conventional model f_C. That candidate might perceive the impact of moving to a fair model as changing the outcome from a 1 (the candidate receives the position) to a 0 (the candidate is rejected), and therefore claim to be negatively impacted by the company's decision to use f_F rather than f_C.*

However, this perception of harm fails to accurately capture the effect of moving from a conventional to a fair model, because it fails to account for alternative outcomes under other conventional classifiers. Suppose, for example, that f_C is only one of number of reasonable choices of conventional models, say $f_C^{(1)}, f_C^{(2)}, f_C^{(3)}$. If under each of these alternative conventional classifiers candidate i would also not be hired, then the negative effect caused by switching from f_C to f_F is in a sense independent of the fact that f_F is a fair classifier, since switching to any *other classifier appears to result in candidate i not being hired. To account for this, our* realized impact *measure takes into account the probability that candidate i actually receives the position under alternative conventional classifiers, given in this example* *by* $p_i = 1/4$*. Thus the realized impact for candidate* i *is* $\mathbb{I}[1 \text{ and not } 0 \text{ and } 1/4 \geq \theta]$*. Thus, if* $\theta \leq 1/4$ *the realized impact for candidate* 1 *is* 1*; otherwise it is* 0*.*

6.4 Popularity of Existing Fair-Learning Schemes

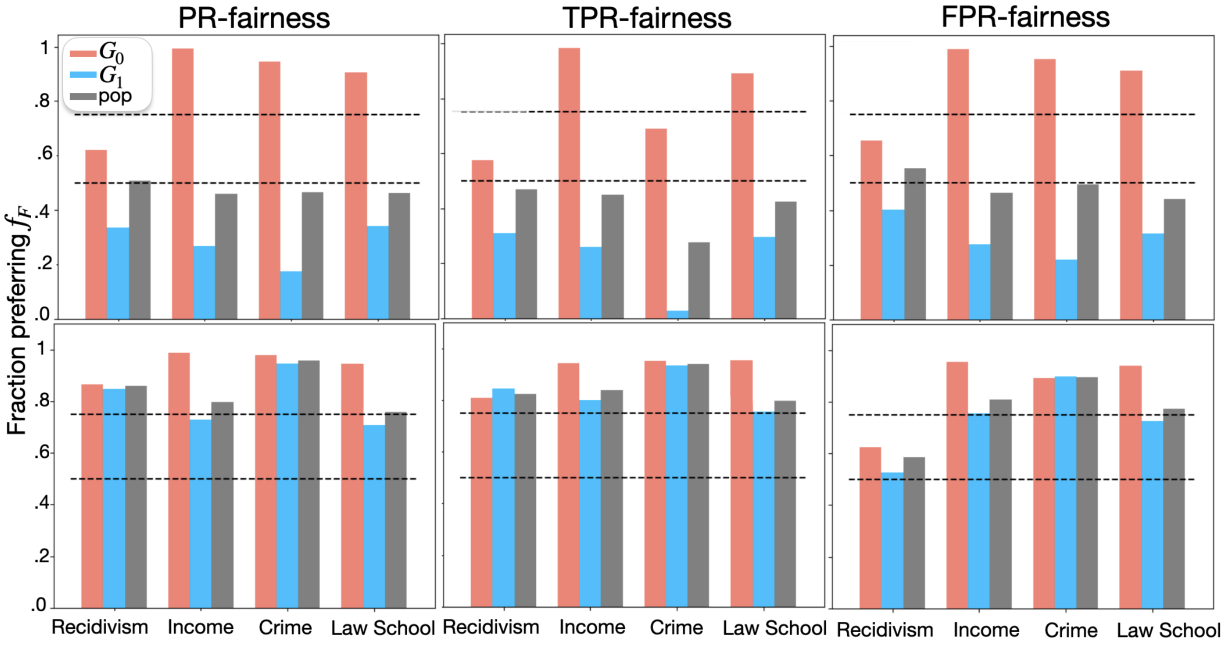

Figure 6.1: Fraction of each population or group preferring f_F over f_C for randomized classifiers (top) and deterministic classifiers (bottom), when f_F is learned via the Reductions algorithm.

We begin by empirically investigating the relationship between popularity and fairness, and evaluate the efficacy of the proposed postprocessing algorithms. Each experiment is conducted on four data sets outlined in Chapter 4. Group membership is defined by race for Community Crime and Law School, and by gender for Recidivism and Income; either feature is assumed to be binary. All other sensitive features, such as age, are removed from the dataset. We consider three baselines of fair learning schemes: Reductions, CalEqOdds, and KDE. Results for the latter two are qualitatively similar and provided in Section B of the Appendix.

The fractions of the overall population, and subgroup population, which prefer the fair classifier are shown in Figure 6.1, where fairness is achieved using the Reductions method.

Not surprisingly, we see that in all instances the disadvantaged group G_0 prefers the fair classifier f_F at far higher rates than G_1. With the exception of the CalEqOdds algorithm (which achieves fairness via group specific score shifts, resulting in far stronger group-level preference over classifiers), results for other methods are qualitatively similar; these are provided in the Appendix. Overall, randomized fair classifiers frequently have popularity of less than 50%. On the other hand, fair deterministic classifiers are relatively popular in most cases.

To understand why the discrepancy in group preferences, consider the case of equalizing PR-based fairness in which f_C has a high positive rate on G_1(e.g. 70%) and a low positive rate on G_0 (e.g. 30%). For sake of example, suppose that f_C is accuracy maximizing. In this case the *unfairness* of f_C is $.7 - .3 = .4$. Suppose that a fair model f_F is learned with the constraint that the positive rate between groups be no greater than .2. Then, there are two ways that f_F can improve the unfairness of f_C: 1) reduce the positive rate on G_1, and 2) increase the positive rate on G_0. Since f_C is accuracy maximizing, it is unlikely that f_F would correctly classify an example when deviating from the prediction of f_C. As such, disparate decisions between the two models are likely to occur only as the result of f_F attempting to achieving a specified level of fairness. Therefore, we would expect that the only score differences between the two models would be increases to G_0 and decreases to G_1. It is precisely this dynamic that gives rise to the disparate model preferences between groups.

These results indicate that for each type of fairness (PR, TPR, and FPR), f_F is achieving fairness, at least in part, by decrease the positive rate on some agents from G_1 and increasing the positive rate on some agents from G_0. Due to the fact that in each dataset G_1 is not only advantaged, but also the majority group, the population-level popularity is skewed towards that of G_1. In later chapters we will see that despite this imbalance, achieving high levels of population-level popularity can be achieved with minimal degradation to fairness and performance, indicating that it need not be the case that fairness is attained at the cost harming the advantaged group.

6.4.1 Individual preferences over fairness

In order to better understand the relationship between varying levels of fairness and individual welfare we also examine the aggregate group-preference for fairness. In particular we look at what level of fairness each individual in the population would prefer, i.e., if each individual wished to maximize their probability of being positively classified, but could only choose the level of fairness of the classifier.

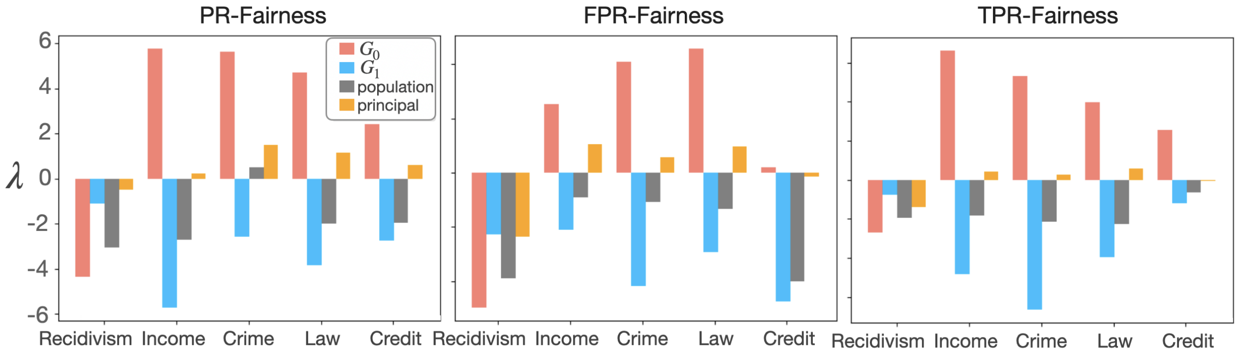

Figure 6.2: Average preferred λ of the population, or each group, for three choices of fairness metrics, all using Logistic Regression. For each dataset and fairness metric β is selected to $\frac{1}{4}$ the unfairness of an accuracy maximizing classifier. For "G_0", "G_1" and "population", λ is the average preferred Lagrangian. For "principal" λ is the Lagrangian for which model accuracy is maximized while being β-fair (i.e., the optimal Lagrangian penalty).

A common method for learning β-fair classifiers is the so called Lagrangian penalty method associated penalty $\lambda \in \mathbb{R}$, i.e.,

$$f_{F_\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{L}(f, X, Y) + \lambda\big(U(f, \mathcal{X}, Y, G) - \beta\big)$$

Here λ gives the relative "importance" of fairness. When $\lambda = 0$ the objective of the conventional classifier is recovered. As such, an agent's most preferred λ, i.e. the λ which maximizes their likelihood of being positively classified can be interpreted as a measurement of much that agent prefers a conventional model. To understand the preference of agents over fair and conventional models, we can also look at the the relative preference for fairness among each group. In particular, suppose that each agent prefers the value of λ yielding the highest expected outcome, i.e. an agent with features $\mathbf{x}$ prefers $\lambda^* = \arg\max_\lambda f_{F_\lambda}(\mathbf{x})$. Then the average of these preferred λ^* across each group gives the groups relative preference for fairness: higher values of λ corresponds to a stronger preference for fairness.

In Figure 6.2 we see the average preferred λ of each group and as well as the total population. In each combination of hypothesis class and dataset, the advantaged group G_1 in aggregate prefers smaller λ while the disadvantaged group G_0 prefers larger λ (larger than the principals choice in-fact). While somewhat unsurprising on its own, this relationship between aggregate group utility (probability of positive classification) and levels of fairness is precisely what leads to fair classifiers being "unpopular". When the advantaged group is also the majority group (as is often the case), the relative size of G_1 implies that the total population on average also prefers smaller values of λ, i.e., less fair models. As such, there is a trade-off between aggregate individual utility and model fairness, later in Chapter 8 we will aim to balance this trade-off by developing models which are both fair and preserve high levels of individual welfare.

6.5 Individual Impact

Next we investigate the individual-level impacts which arise when imposing group fairness in the case of scarce resource allocation. The experimental setup for this chapter is the same as that of the previous chapter with the exception of switching from classification to scarce resource allocation. In scarce resource allocation, a set of k homogeneous resource are allocated to the population of n individuals using a score function $f(\mathbf{x})$ (which produces real-valued outputs), the top-k scoring agents receive a resource. Recall that perceived and realized impact are binary measures (either an agent is impacted or they are not) given in Definition 6.3.3. Our experiments examine how likely an agent is to be impacted (or in how many different worlds it is impacted) under perceived or realized impact. To simulate these different worlds, we use the use the following pipeline to generate f_C, f_F and their associated allocation decisions:

1. Split the dataset into two portions: 80% for training and 20% for testing.
2. Subsample 80% of the training set 50 times with replacement, generating 50 *random* training subsets, which we refer to as *subsamples*.
3. On each of the 50 subsamples, train one of each conventional and fair model, after a 5-fold hyper-parameter search.
4. On each subsample t, assign to each agent i in the test dataset, with a corresponding feature vector $\mathbf{x}_i$, a score $f^{(t)}(\mathbf{x}_i)$, used to compute the resource allocation decision as well as impact measures.

All results are reported as a 5-fold average of the above pipeline. That is, we partition the dataset into 5 disjoint sets (each containing 20% of the data) and run the above pipeline 5 times, across each of the 5 partitions, (reporting the average impact on the 20% of agents designated in the testing set).

When computing realized impact, the randomness of $p_i = \mathbb{P}_{f'_C}\big(I_i(f'_C) = 1\big)$ may come in two forms: 1) uncertainty about the training data, and 2) (additional) uncertainty about the model type. Experimentally, we account for randomness in the training data by randomly subsampling an initial training dataset, and training a model on each of the subsamples.

6.5.1 Perceived and Realized Impact

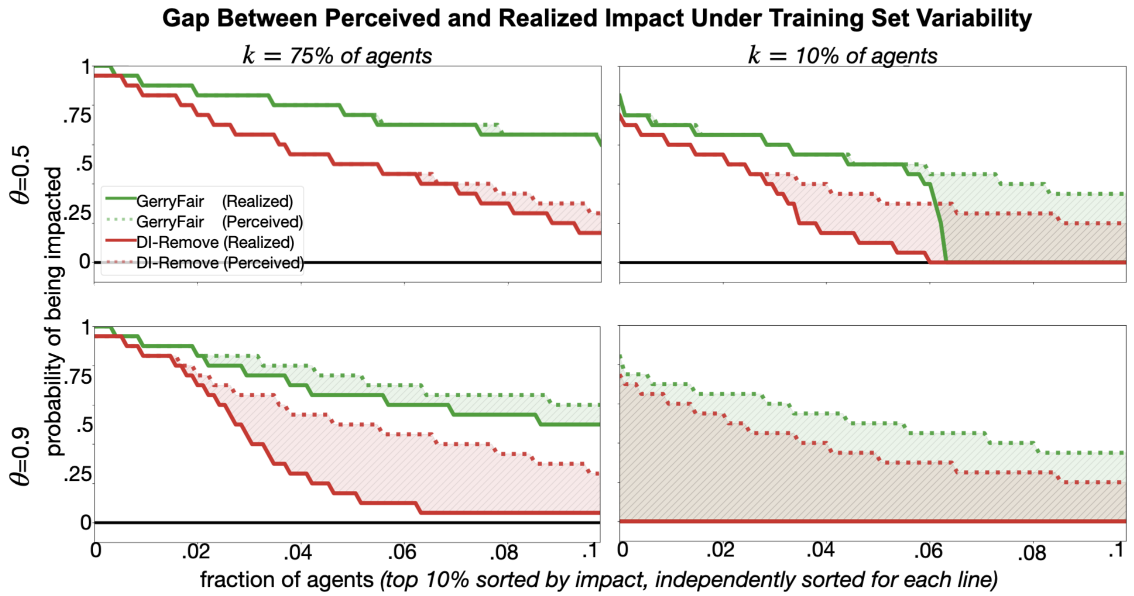

Figure 6.3: Impact measured on Homelessness dataset. Comparison between perceived impact (solid lines) and realized impact (dotted lines) when background variation only accounts for randomness in training data (captured by subsampling). Plots compare Logistic Regression to both GerryFair($\gamma = 0.001$) and DI-Remove. Each graph shows the fraction of the times (y-axis) that at least a certain percentage of individuals (x-axis) suffer perceived or realized impact. This can also be interpreted as the probability that the individual at the x^{th} percentile (on the x-axis) is impacted. Note that perceived impact (dotted) is not completely independent of θ due to occasional ties in scores. The shaded region indicates the degree to which perceived impact overestimates realized impact.

Our definition of *impact* above was in the context of a *particular* individual challenging a "group fair" approach for obtaining a scoring function used to allocate scarce resources. We now step back and consider how perceived and realized impact measures differ *on average, over a population of individuals* if we view them from a policy perspective, that is, over a large collection of such hypothetical cases. Specifically, we investigate the extent to which natural variation in learned models affects how much smaller realized impact is, on average, compared to perceived impact—that is, to what extent perceived impact may simply be due to natural variation rather than the choice of group-fair strategy.

Our definition of *impact* above was in the context of a *particular* individual challenging a "group fair" approach for obtaining a scoring function used to allocate scarce resources. We now step back and consider how perceived and realized impact measures differ *on average,*

over a population of individuals if we view them from a policy perspective, that is, over a large collection of such hypothetical cases. Specifically, we investigate the extent to which natural variation in learned models affects how much smaller realized impact is, on average, compared to perceived impact—that is, to what extent perceived impact may simply be due to natural variation rather than the choice of group-fair strategy.

We consider first the natural variation that occurs simply as a result of sampling the training data. As mentioned above, we first randomly select 20% of each dataset as the *test set*. We consider these the individuals among whom the resources will be allocated. The remaining 80% (the *training universe*) are then subsampled multiple times (again, as described above, we sample 80% of the training universe with replacement 50 different times), each subsample creating a *resource allocation scenario*, as it determines the model f used for resource allocation. Throughout, we quantify scarcity as a percentage k of individuals (in the test set) who can be allocated a resource. Taking into account only the variation caused by sampling the data, we have the following main observations:

Observation O1: *Realized impact is, in most settings, much lower than perceived impact.*

Observation O2: *The gap between realized and perceived impact increases as scarcity increases (i.e. as k decreases).*

Observation O3: *Realized impact is relatively rare. This is particularly true when either k is small (high scarcity) or when θ is high (i.e., individuals must receive the resource with high probability when using the conventional classifier to be considered impacted).*

Figure 6.3 provides a representative illustration of the observations above using homelessness data; in Appendix A we provide additional results for the remaining datasets, as well as for different conventional and fair classifiers. Each plot shows the percentage of individuals who experienced at least a certain level of impact according to our definitions. The x-axis shows the percentile label, while the y-axis is the fraction of subsamples in which the given measure of impact is observed. The solid lines represent realized impact and the dotted lines perceived impact. In this figure, we compare conventional logistic regression with either *GerryFair* (green lines) or *DI-Remove* (red lines), and natural variation occurs solely due to the repeated subsampling of the training data.

For example, take the upper right plot in Figure 6.3, which corresponds to $k = 10\%$ and $\theta = 0.5$, and consider the *GerryFair* group-fair approach, which corresponds to the green lines. At $x = 0.08$, the *perceived* impact (dotted green line) is approximately 0.4, which means that 8% of individuals appear to be impacted in over 40% of resource allocation scenarios. In contrast, the 8th percentile value of *realized* impact (solid green line) is 0—we do not observe it in *any* allocation scenarios, which means that fewer than 8 percent of individuals are impacted at this level. The *gap* between the perceived and realized impact in this case is the green shaded area (the area between the dotted and solid green lines).

While the perceived impact (dotted lines in Figure 6.3) will always be above realized impact (solid lines) by construction, the gap is often substantial (**Observation O1**), demonstrating the importance of accounting for natural variation in the learned model in measuring individual impact. The starkest illustration is the plot on the lower-right ($\theta = 0.9$ and $k = 10\%$), where perceived impact is considerable for at least 10% of individuals, but almost entirely due to natural variation (realized impact is essentially zero for all individuals).

To illustrate **Observation O2**, note from Figure 6.3 that the gap between perceived and realized impact (shaded areas) is considerably larger in the right pair of plots (for $k = 10\%$, high scarcity) than in the left pair ($k = 75\%$, low scarcity); thus, greater scarcity leads to a greater gap between perceived and realized impact. Furthermore, both decreasing the percentage of available resources (smaller k) or being more conservative about what we consider to be realized impact (higher θ) tends to make realized impact considerably less common (**Observation O3**), as we see in Figure 6.3 when comparing different values of θ (top and bottom) as well as the availability of resources (left and right). For example, when $k = 10\%$, fewer than 8% of individuals see any realized impact, and if we conservatively choose $\theta = 0.9$, once we account for natural variation, impact is exceedingly rare for *all* individuals. However, in several cases the fraction of individuals with realized impact is not negligible. For example, when $\theta = 0.5$ and $k = 75\%$, we see substantial realized impact for as many as 20% of the individuals. However, it is worth noting that $\theta = 0.5$ is a very weak threshold indeed – the individual's chance of getting the resource is a coin flip. We present it as an upper bound on what fraction might possibly be considered impacted.

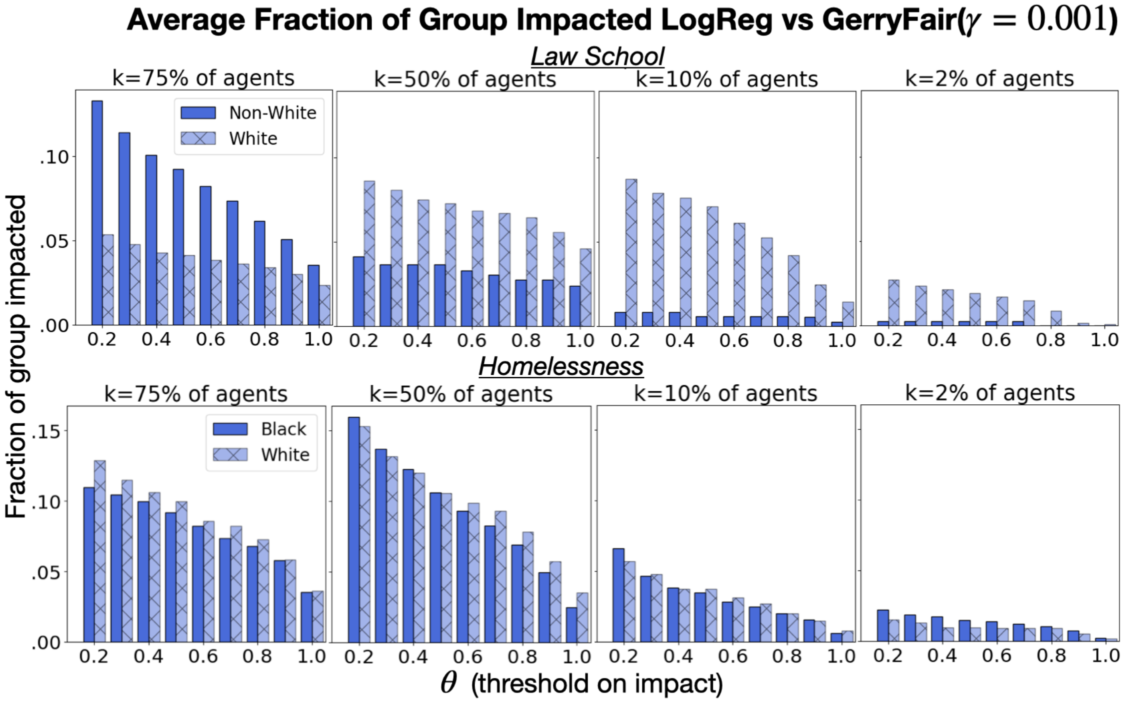

Figure 6.4: Average percentage of individuals from each group who are impacted (realized impact, with parameter θ) when switching from Logistic Regression to Gerryfair. Results are shown on the Law School (top row) and homelessness (bottom row) data sets.

6.5.2 Individual Impacts on Disadvantaged Groups

Recall that we defined the *disadvantaged group* as the group with lower positive rate (PR) and/or true positive rate (TPR) when conventional learning is used; we refer to the group with higher PR and/or TPR under conventional learning as the *advantaged group.*

Remark 6.5.1. *Realized impact can disproportionately negatively affect members of the disadvantaged group.*

This observation can be seen in Figures 6.4 and 6.5, in which we examine for two racial groups the *average* (over the 50 data subsamples) percentage of individuals (y-axis) suffering realized impact for various impact thresholds θ (x-axis). Results for other models are provided in Section B of the Appendix. In Figure 6.4, this is done for the GerryFair approach to group-fair learning, while 6.5 shows analogous results for DI-Remove. Yet again, we observe that fair learning approaches, much as they may aspire towards similar goals, can yield qualitatively different results from each other. First, consider GerryFair. On the law school dataset

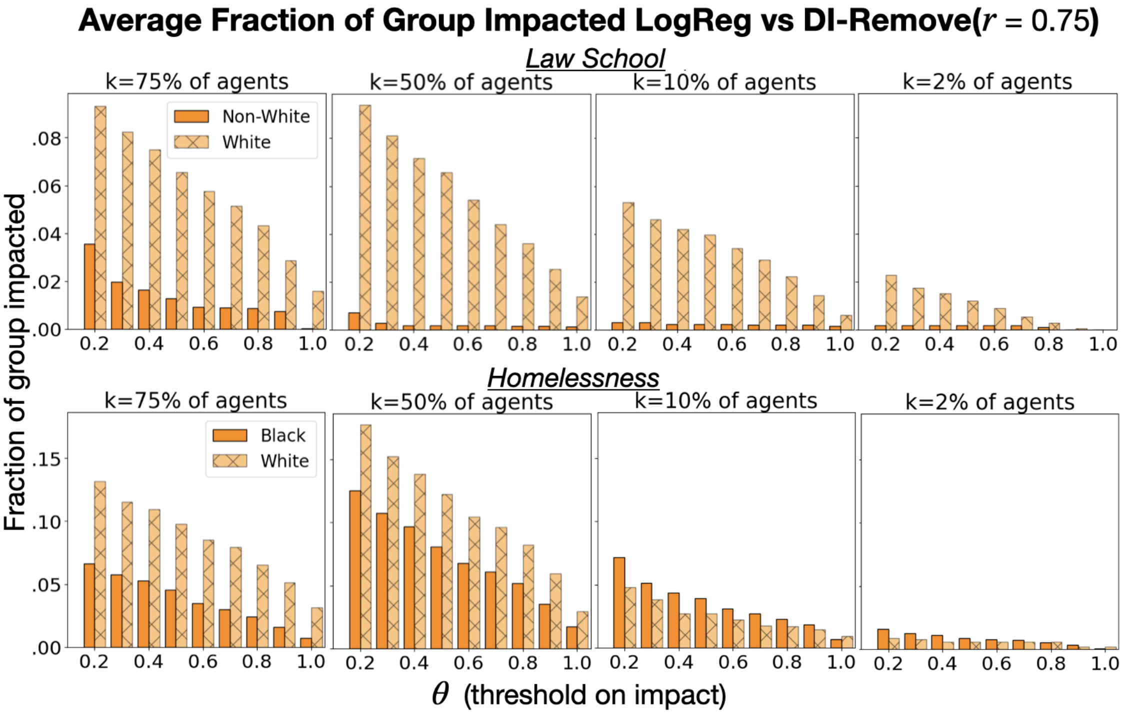

Figure 6.5: Average percentage of individuals from each group who are impacted (realized impact, with parameter θ) when switching from Logistic Regression to DI-Remove. Results are shown on the Law School (top row) and homelessness (bottom row) data sets.

(top row), as the plots range from low scarcity (left) to high scarcity (right), we observe a startling phenomenon: when scarcity is low, it is the Non-White individuals who exhibit a disproportionate share of the negative impact, but this changes with greater scarcity, where negative impact becomes more concentrated among White individuals. On the homelessness data (bottom row), the impact is roughly equal between groups with a slight shift towards the Black individuals as resources becomes more scarce.

On the Law School dataset, the negatively impacted individuals are almost entirely from the White demographic in the case of DI-Remove. However, the negative impact of DI-Remove falls disproportionately on the Black group in the Homelessness dataset when the resources are more abundant. Nevertheless, as we have already observed above, with sufficient scarcity we see *very few* individuals from either group who are meaningfully impacted. These examples serve to illustrate that group fairness, a concept mostly considered in terms of averages and expected values, may at times have unintended effects on individual members of subpopulations which are obscured by the nature of averages.

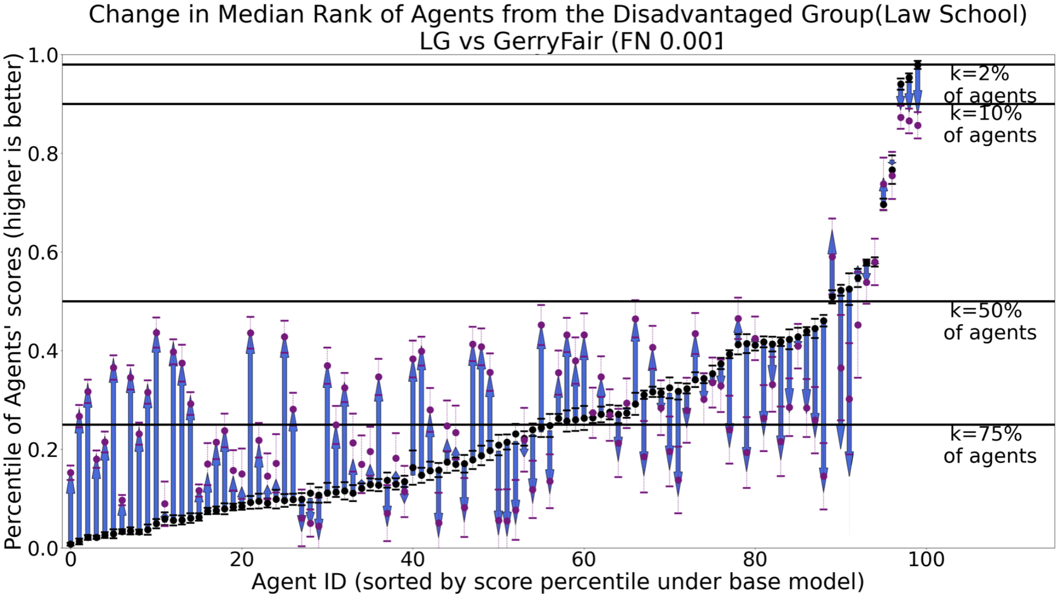

Figure 6.6: Change in individuals' median score percentile when changing from Logistic Regression to GerryFair. Arrows represent the change in median rank (base of each arrow is the rank using the conventional classification score and tip is using the fair classification score). Black (purple) gives the average and standard deviation of an individual's rank under the conventional (fair) model. For ease of display we have uniformly selected 100 individuals from the Non-White group.

6.5.3 Causes of Disparities in Individual Impact

As shown previously, to a considerable extent the (negative) impact may devolve on the individuals in the disadvantaged group may. Similar to our investigation into popularity (Section 6.4), we are interested in why adverse effects to individual welfare may fall disproportionately on one group. In particular we are interested in why members of the disadvantaged group can share the majority of impact, especially when resources are abundant. Central in our investigation is the consideration of the relative ranks of individuals (in terms of what percentile their scores fall), and how these shift when switching from a conventional classifier to a fair classifier.

Consider Figures 6.6 and 6.7 for the Law School dataset; in the former, f_F is GerryFair, while in the latter f_F is DI-Remove. In these figures, each arrow corresponds to an individual from the disadvantaged group. The base of an arrow is the individual's average percentile rank in terms of the conventional classification score, while the tip is this individual's average percentile in terms of the fair classification score. Thus, if the fair classifier improves the

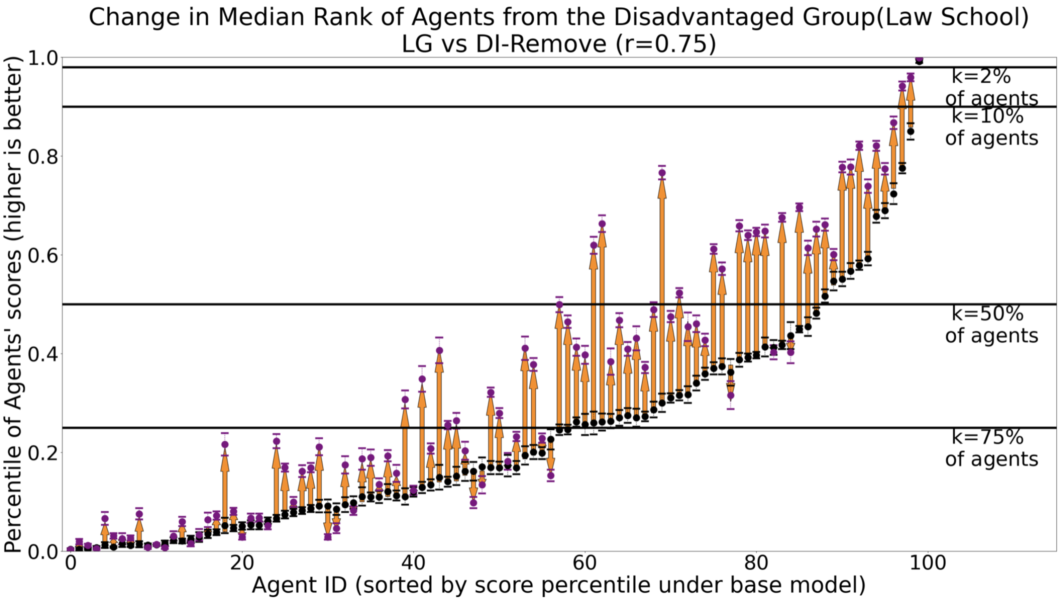

Figure 6.7: Change in individuals' median score percentile when changing from Logistic Regression to DI-Remove. Arrows represent the change in median rank (base of each arrow is the rank using the conventional classification score and tip is using the fair classification score).

individual's average rank, the arrow points up, whereas it points down if the average rank drops as we switch from conventional to fair. In both figures, conventional classifier is the logistic regression, and for legibility we present 100 randomly chosen individuals from the Law School dataset.

Remark 6.5.2. *The disadvantaged group is unlikely to have scores from the conventional model, which fall with in the top* 2% *and* 10% *of scores. As such, there are few members of the disadvantaged group capable of losing allocation when resources are more scarce (i.e.* $k = 2\%, 10\%$*).*

The first key observation is that we observe many arrows that point down—that is, there are many cases in which the fair classifier actually reduces the average percentile rank of an individual from the disadvantaged group. This is particularly common in Figure 6.6, where f_F is GerryFair, but can also be observed, albeit far less frequently, in Figure 6.7, where f_F is DI-Remove. This difference between the effects of two different fair classifiers tracks with our observations of the extent of impact on the members of disadvantaged groups in Figures 6.4 and 6.5 above for the Law School dataset.

Moreover, Figure 6.6 offers an additional insight into our observations above. Note that this this figure, the positive effects from fair classification on the disadvantaged group (arrows pointing up) are far more prevalent for individuals who are ranked particularly low by the conventional classifier. However, much as fair classification helps increase relative classification scores for these individuals, this is of little consequence if resources are scarce, since even with the percentile boost, they can only receive the resource if it's quite abundant. On the other hand, as we make the resource more abundant, we also accumulate many individuals from this group whose ranking drops due to fair classification, as especially many of them are in the intermediate $k \in [0.5, 0.75]$ range.

Considering, in contrast, DI-Remove (Figure 6.7), the distribution of percentile changes is fundamentally different from GerryFair. For example, in this case the vast majority of benefit (arrows facing up) among individuals whose conventional classification score percentile is relatively high, and the boost in this score from DI-Remove is often substantial. In this random sample of individuals, we see no instances of reduced ranking (arrows facing down) for individuals in the disadvantaged group ranked in the top 50th percentile, and relatively few in the $k \in [0.5, 0.75]$ interval.

6.6 Discussion

Algorithms for fair learning have emerged as a response to a number of demonstrations that conventional machine learning algorithms can lead to inequalities in prediction between historically advantaged and disadvantaged groups. The concern is serious: as algorithms are increasingly used to support decisions that have a direct impact on people, such as lending and employment decisions, such algorithmic inequalities can perpetuate historical injustices. However, to the extent that fair learning approaches involve explicitly taking into account race and other protected characteristics, they may themselves raise ethical and legal concerns. In particular, group-fair learning approaches necessarily shift how resources are distributed, but does that cause unjustifiable harm to some individuals? Answering that question entails normative and policy judgments, but making those decisions requires an accurate understanding of who is actually impacted and how. Our study focuses on properly characterizing those individual impacts by systematically examining which individuals are impacted, and to what degree, when a decision-maker chooses to implement group-fairness learning approaches.

Our results challenge the assumption that group-fair classification inherently harms individuals who are not members of the protected group. To the contrary, we find that once we properly account for the fact that machine learning-based decisions will naturally vary due to a host of factors, typically only a small fraction of individuals (from both advantaged and disadvantaged groups) is impacted in a meaningful way.

The claim that many individuals are harmed by group-fairness rests on what we call *perceived impact*—the view that an individual denied a resource under a group-fair model has been harmed because that person could have received the resource under one possible version of a conventional model. Defining individual harms in this way ignores the uncertainty of outcomes under conventional models and misconceives the extent to which a negative outcome is attributable to the fairness constraints. By ignoring the natural variation that arises under even conventional machine learning, perceived impact will often overestimate, sometimes vastly so, as our results show, the extent to which a particular individual has actually been negatively impacted by the choice of a group fairness model. Indeed, in many contexts where group-fair machine learning might be used, few individuals would be guaranteed to receive a positive outcome across plausible baseline (non-fairness constrained) models. We show that even random draws of the training dataset produce a certain amount of "natural" variation in outcome for each individual that must be taken into account. The choice of learning algorithm among plausible conventional models introduces additional variation.

Part III

Mitigating Impacts of Group-Fair Learning

Chapter 7

Auditing to Mitigate Strategic Manipulation

In Part II we saw that strategic agent behavior can undermine the fairness and performance of group-fair models, so much so that they becomes less fair than fairness-agnostic models, i.e., the phenomenon of fairness reversals. We saw that the root cause of these fairness reversals was classifier selectivity, i.e. the fair model negatively classifies numerous individuals who would otherwise received positive classified if a fairness-agnostic model were used. In Chapter 6 the notion of selectivity was further expanded into the concept of individual impact, both in terms of real- and perceived-impact as well as individual welfare (captured by popularity). Lastly we discussed several postprocessing techniques which achieve high levels of popularity as well as group-fairness. While these techniques can be used to to prevent fairness reversals, as discussed in Section 8.4, they do not directly mitigate strategic behavior. To this end, we discuss auditing as a means of disincentivizing strategic behavior. Recall that auditing constitutes a verification that an agent's reported data (e.g., the information supplied on a lending application). Agents found to be misreporting their data are subject to denial (e.g., no longer granted a loan) and may be subject to a fine. As audits are often costly, we study cases in which the principal can only audit a limited number of agents in a population.

Our goal is to design audit policies which either prevent strategic behavior from impacting proprieties of the model, such as performance, or reduce the incentivizes for agents to behave strategically. More specifically, we study three types of objectives: *incentivize-minimization* in which the principal aims to develop a policy which minimizes the maximum incentive that any agent has to manipulate, *recourse-maximization* in which the principal aims to develop a policy which results in the largest number of agents electing to perform recourse (true feature changes), and *utility-maximization* in which the principal aims to develop a policy which maximizes their own utility (e.g., total profit for re-payed loans).

Additionally, we investigate the use of subsidies as a deterrent to manipulation. In particular, agents are frequently capable of both manipulations and *true* feature changes (such as those studied in the field of algorithmic recourse). When agents are posses both abilities, it may no longer be optimal to use punitive measures (audits) in isolation. We study the case when the principal is capable of allotting a portion of their audit budget to help agents make true feature changes. We find that in a large number of cases, it is optimal for the principal to allot a nonzero fraction of their audit budget to subsidies.

7.1 Summary of Results

We first introduce a framework that unifies manipulation and recourse (true feature changes), and obtain several consequential results in this model. We show that computing an optimal audit policy is tractable for both a utility-maximizing principal and a principal who simply wishes to maximize the number of agents choosing recourse (recourse-maximizing). This is true both when the costs of failing an audit are exogeneously specified and when the costs are chosen by the principal. We prove that when fines are exogeneously specified, the objectives of a recourse- and utility-maximizing principal are aligned for any distribution of agents, features, and cost of recourse.

We then turn our attention to studying a model of *subsidies*, where the principal can choose to devote allot part of their audit budget to instead *subsidize agents to choose recourse*. We derive necessary and sufficient conditions for the principal to use a nonzero portion of their audit budget on subsidies. We show that even with subsidies, the objectives of a recourse- and utility-maximizing principal are again aligned when agents value positive classifications equally. We then characterize the relationship between auditing/subsidies, the total amount of fines or cost of recourse imposed on a population, and the fraction of the population preferring recourse to manipulation.

7.2 Related Work

Here we provide further discussion on works which pertain specifically to auditing and recourse.

Auditing Theory examines problems in which a system (e.g., a bank) possesses the ability to verify (audit) information reported by an individual (e.g., a loan applicant). Auditing carries a negative consequence, such as a fine, when the reported information is found to be inauthentic. The work of [14] formulates auditing as a game between a defender attempting disincentivize manipulations, and an attacker attempting to avoid detection while obtaining a desired outcome (similar to a Stackelberg security game). Other works have studied auditing in the context of multiple individuals attempt to manipulate a classification or allocation system in order to gain a desired resource [80, 37]. Auditing in the context of Strategic Classification remains relatively underexplored with the primary work being [37] which examines auditing as a means of inducing incentive compatibility (i.e. all agents truthfully report), but does not examine model robustness outside of this narrow lens. Works in this domain do not consider the ability for agents to perform recourse and are typically agnostic to system utility.

Recourse focuses on providing agents receiving undesirable outcomes from a machine learning model, with the ability to contest or improve their outcome via a modification to their attributes in a *genuine* manner (paying off debt to increase creditworthiness) [110, 55, 57, 107, 39, 111]. The concept of recourse in machine learning was first introduced in [110] where an integer programming solution was developed to offer actionable recourse to agents who are rejected by a linear classifier. Our work in this chapter makes use of the general formulation of recourse proposed in [110], which frames recourse as an optimization problem of finding *minimum* cost feature modifications which an agent can feasibly make in order to obtain a desired outcome. Within this framework, we explore the role of auditing as a means of incentivizing recourse over manipulation.

7.3 Preliminaries

We begin with a motivating example. A bank aims to maximize their expected profit by issuing fixed-rate credit cards (with set spending limits and interest rates). Because of the high volume compared with, say, corporate loans, credit cards are a major area where banks use algorithmic decision-making [17]. Each applicant (with application $\mathbf{x}$) is approved for a fixed-rate card if the bank's model predicts that an applicant will offer a positive profit. While the profitability comes from different channels, e.g. building a relationship with a client who will then use the bank for other services versus actual interest payments, the main risk in issuing a card is that the customer will default after running up a balance [65], so banks want to filter out those applicants. IF the bank denies an application, the bank's utility is 0 as no money is exchanged. The bank may offer denied applicants access to recourse, i.e., a plan for making the applicant more creditworthy, such as paying off outstanding debt or increasing income. However, when applicants have knowledge of recourse actions, they may report that they have taken such actions in order to get approved, without actually taking the actions (e.g. hiding debt or inflating income). The bank could audit applicants by verifying information in their applications. However, since this is costly, the audit budget is limited.

We now present our formal model of auditing and recourse. Let $\mathcal{D}$ be a distribution over features $\mathcal{X} \subset \mathbb{R}^d$ with probability measure p. Consider a principal who aims to make a binary decision $\hat{y}(\mathbf{x}) \in \{0, 1\}$ for each input feature vector $\mathbf{x}$, for example, to approve or deny a loan. We refer to the decision $\hat{y}(\mathbf{x}) = 1$ as *selection*, with $\hat{y}(\mathbf{x}) = 0$ corresponding to $\mathbf{x}$ not being selected. For any *actual* feature vector $\mathbf{x}$ (to distinguish from manipulated features we discuss below), the principal receives a utility of $u_p(\mathbf{x})$ whenever $\hat{y}(\mathbf{x}) = 1$ (e.g., expected profit from a loan) and utility of 0 otherwise; in other words, the principal's utility is $u_p(\mathbf{x})\hat{y}(\mathbf{x})$.

Prediction function We assume that the principal's utility from selecting $\mathbf{x}$ is based on an objective measure, such as loan repayment rate, that is not known directly, but can be estimated from data. Thus, let $f : \mathcal{X} \to \mathbb{R}$ be a model learned from data that predicts $u_p(\mathbf{x})$. For example, f can predict the probability that a loan is repaid, multiplied by expected profits conditional on repayment. Importantly, we assume that f is fixed and common-knowledge, and is applied to the *reported* features. The application of f is thus mechanistic and not an action under the control of the principal in the game-theoretic sense. This is consistent with our use cases – bank regulators, for example typically require that a model is demonstrably a valid predictor and that it should be used consistently across the entire population of applicants for a period of time. Thus, f is simply used to select all $\mathbf{x}$ that yield a predicted utility above a given threshold θ:

$$\hat{y}(\mathbf{x}) = \mathbb{I}\big[f(\mathbf{x}) \geq \theta\big].$$

If we set $\theta = 0$, this has the natural interpretation in the context of loans that all applications with positive expected utility (based on the reported features) are approved.

Principal's Actions: Auditing Agents can misreport their feature vectors. The principal's main tool to disincentivize such misrepresentation is the use of audits. When the principal audits an agent reporting features $\mathbf{x}'$, the agent's true features $\mathbf{x}$ are revealed to the principal. Failing an audit, i.e., being audited when $\mathbf{x}' \neq \mathbf{x}$ will result in the agent paying a fine; we follow the models of auditing in [14, 37] and assume agents pay a constant fine C when they are caught manipulating, in addition to not being selected. Before agents report their features, the principal publicly declares its *audit policy*.

Definition 7.3.1. (Audit Policy) *Given a set of n agents with true features $\mathbf{X}$ and reported features $\mathbf{X}'$, an* audit policy *is a mapping $\alpha : \mathcal{X}^{n+1} \to [0,1]$ where $\alpha(\mathbf{x}';\mathbf{X}')$ corresponds to the probability that an agent reporting features $\mathbf{x}'$ is audited, given the set of reports $\mathbf{X}'$ for the n agents. The principal is limited B audits on average, i.e., $\mathbb{E}\big[\sum_{\mathbf{x}' \in \mathbf{X}'} \alpha(\mathbf{x}';\mathbf{X}')\big] \leq B$.*

An audit of a particular agent is a check whether $\mathbb{I}[\mathbf{x}' \neq \mathbf{x}]$, which we assume to be reliable. Agents caught misreporting their features are subject to a fine $C \in \mathbb{R}_{\geq 0}$.

Agents An agent with true features $\mathbf{x}$ gains utility $u_a(\mathbf{x})$ when approved by the principal, and 0 otherwise. When reporting features $\mathbf{x}'$, and not being caught by an audit, the agent then obtains utility $u_a(\mathbf{x})\hat{y}(\mathbf{x}')$. In addition to the general case, we also consider a special case where the utility of being selected is a constant, i.e., $u_a(\mathbf{x}) = \bar{u}_a$ for all $\mathbf{x}$. This special case has received most attention in prior literature, particularly in the context of recourse [110].

Agents' Actions: Recourse and Manipulation Formally, n agents arrive i.i.d. with features $\mathbf{x} \sim \mathcal{D}$; we assume that $\mathcal{D}$ is common knowledge. We use $\mathbf{X} \sim \mathcal{D}$ to indicate a collection of n feature vectors thereby generated. Each agent has an action space comprised of two qualitatively distinct types of actions: recourse and manipulation. We allow arbitrary composition of these, although prove below that such compositions are dominated by a choice of manipulation, recourse, or neither (reporting true initial features $\mathbf{x}$). Let $\mathbf{z}$ denote a recourse choice, which we restrict to be in the set $A(\mathbf{x})$ that defines what is *actionable* [110]. The agent always has the option to do nothing, i.e. $\mathbf{x} \in A(\mathbf{x})$, and if the agent elects this do-nothing action (which carries no cost), then $\mathbf{z} = \mathbf{x}$. The cost of a recourse action $\mathbf{z}$ for an agent with initial features $\mathbf{x}$ is denoted by $c_R(\mathbf{x}, \mathbf{z})$. We use $\mathbf{z}'$ to denote reported (potentially manipulated) features.

While selection decisions $\hat{y}$ are implemented independently for each reported feature vector $\mathbf{z}'$, the audit policy $\alpha(\mathbf{z}'; \mathbf{Z}')$ depends on the full collection of n *reported* feature vectors of all agents, namely $\mathbf{Z}'$. Let $g(\mathbf{x})$ be the strategy of an agent with true features $\mathbf{x}$ in the choice of both recourse $\mathbf{z}$ and reported (and possibly untruthful) features $\mathbf{z}'$. We restrict attention to symmetric pure strategies, so that g deterministically returns a pair $(\mathbf{z}, \mathbf{z}')$. Given a symmetric strategy profile g and an agent who reports a feature vector $\mathbf{z}'$, the probability of this agent being audited is $\mathbb{E}\big[\alpha(\mathbf{z}'; g(\mathbf{X}))\big]$, where the expectation is with respect to $\mathbf{X} \sim \mathcal{D}$ (here, it is only the final reports induced by g that matter). We define the expected cost of manipulation for an agent with true features $\mathbf{z}$ (possibly after recourse) and reported features $\mathbf{z}'$, when all other agents jointly follow strategy g as

$$c_A(\mathbf{z}, \mathbf{z}'; g) = \mathbb{E}\big[\alpha(\mathbf{z}'; g(\mathbf{X}))\big]\mathbb{I}\big[\mathbf{z}' \neq \mathbf{z}\big]\big(u_a(\mathbf{z})\hat{y}(\mathbf{z}') + C\big).$$

Putting everything together, the expected utility of an agent with initial features $\mathbf{x}$, recourse $\mathbf{z}$, and reported features $\mathbf{z}'$, given a symmetric strategy profile g followed by all others, is

$$U_a(\mathbf{z}, \mathbf{z}', g; \mathbf{x}) = u_a(\mathbf{z})\hat{y}(\mathbf{z}') - c_R(\mathbf{x}, \mathbf{z}) - c_A(\mathbf{z}, \mathbf{z}'; g). \tag{7.1}$$

When all agents follow g, we simply write $U_a(g; \mathbf{x})$, as $(\mathbf{z}, \mathbf{z}') = g(\mathbf{x})$. Our solution concept for agent strategies is a (pure-strategy symmetric) Bayes-Nash equilibrium.

Definition 7.3.2. *A symmetric pure-strategy strategy profile g is a* Bayes-Nash equilibrium *(BNE) if for all agents i with initial features $\mathbf{x}_i$, the action $g(\mathbf{x}_i)$ is a best response, i.e., $U_a(g; \mathbf{x}_i) \geq U_a(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}'_i, g; \mathbf{x}_i)$ for all $\bar{\mathbf{z}}_i \in A(\mathbf{x}_i)$ and $\bar{\mathbf{z}}'_i$. We denote the BNE profile with the maximum number of manipulations as $g_{\max}$.*

7.4 Auditing and the Principal's Objective

In this section we investigate the audit polices of both a recourse-maximizing principal and a utility-maximizing principal. We begin by characterizing some key facts about agents' best responses given the principal's audit policy.

Lemma 7.4.1. *It is never a best response for an agent to perform both recourse and manipulation i.e. either* $\mathbf{z} = \mathbf{x}$ *or* $\mathbf{z} = \mathbf{z}'$.

Proof. This result follows from the fact that agent utility is independent of the report $\mathbf{z}'$ whenever $\hat{y}(\mathbf{z}') = 1$ □

Next we examine the best response of each agent $\mathbf{x}$, with recourse action $\mathbf{z}$ ($\mathbf{z} = \mathbf{x}$ if no recourse occurs), given prediction function f, decision making scheme $\hat{y}$, audit policy α, and fine C. For any strategy g by other agents, the optimal manipulation and recourse are given respectively by,

$$\mathbf{x}_M = \arg\max_{\mathbf{z}' \neq \mathbf{x}} \; u_a(\mathbf{x}) - c_A(\mathbf{x}, \mathbf{z}'; g) \qquad \text{s.t.} \quad \hat{y}(\mathbf{z}') = 1 \tag{7.2}$$

$$\mathbf{x}_R = \arg\max_{\mathbf{z} \in A(\mathbf{x})} \; u_a(\mathbf{z}) - c_R(\mathbf{x}, \mathbf{z}) \qquad \text{s.t.} \quad \hat{y}(\mathbf{z}) = 1. \tag{7.3}$$

For an agent $\mathbf{x}$, let $U_{a,R}(\mathbf{x}) = u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)$ and $U_{a,M}(\mathbf{x}) = u_a(\mathbf{x}) - c_A(\mathbf{x}, \mathbf{x}_M; g)$, i.e. the agent's respective utility gain from recourse or manipulation. The next lemma characterizes the structure of agent best response actions in terms of their expected utility gain.

Lemma 7.4.2. *The best response of an agent with features* $\mathbf{x}$ *has the following form:*

$$\mathbf{z}^* = \begin{cases} \mathbf{x} & \text{if } \; \hat{y}(\mathbf{x}) = 1 & \text{(7.4a)} \\ \mathbf{x}_R & \text{if } \; U_{a,R}(\mathbf{x}) \geq \max\left(0, U_{a,M}(\mathbf{x})\right) & \text{(7.4b)} \\ \mathbf{x}_M & \text{if } \; U_{a,M}(\mathbf{x}) \geq \max\left(0, U_{a,R}(\mathbf{x})\right) & \text{(7.4c)} \\ \mathbf{x} & \text{otherwise} & \text{(7.4d)} \end{cases}$$

where Equations 7.4a, 7.4d correspond to truthful reporting, Equation 7.4b corresponds to recourse, and Equation 7.4c corresponds to manipulation.

Lemma 7.4.2 follows directly from each action's definition.

Next we formalize the objective of the principal. We consider three types of principals: risk-averse principal who aims to minimize the maximum incentive that any agent has to manipulate (dubbed *incentive-minimizing*), a population-oriented principal who aims to maximize the proportion of agents that prefer recourse to manipulation (dubbed *recourse-maximizing*) and a principal who aims to maximize the total utility gain of the decisions made by $\hat{y}$

(dubbed *utility-maximizing*). The latter two objectives respectively represent a principal who is socially-oriented (we treat recourse as a kind of social good, as it benefits participants), or solely self interested.

Definition 7.4.3. ***(Incentive-Minimizing Principal)****: A principal is* incentive-minimizing *if their objective is to select an audit policy α which minimizes the maximum incentive for agents to manipulate, i.e. select α such that truthful reporting is a ε-Bayes-Nash Equilibrium, for the minimum possible ε. In this case, the model is said to be ε-incentive compatible.*

$$\begin{aligned} \alpha^* = \arg\min_{\alpha} \max_{\mathbf{x}\in\mathcal{X}} U_{a,M}(\mathbf{x}) \\ s.t.\ \mathbb{E}_{\alpha}\big[\sum_{\mathbf{z}'\in\mathbf{Z}'} \alpha(\mathbf{z}';\mathbf{Z}')|\mathbf{Z}'\big] \leq B \quad \forall\ \mathbf{Z}' \end{aligned} \tag{7.5}$$

Definition 7.4.4. ***(Recourse-Maximizing Principal)****: A principal is* recourse-maximizing *if their objective is to select an audit policy α which maximizes the proportion of agents who prefer recourse over manipulation:*

$$\begin{aligned} \alpha^* = \arg\max_{\alpha} \mathbb{P}_{\mathbf{X}}\big(U_{a,R}(\mathbf{x}) \geq U_{a,M}(\mathbf{x})\big) \\ s.t.\ \mathbb{E}_{\alpha}\big[\sum_{\mathbf{z}'\in\mathbf{Z}'} \alpha(\mathbf{z}';\mathbf{Z}')|\mathbf{Z}'\big] \leq B \quad \forall\ \mathbf{Z}' \end{aligned} \tag{7.6}$$

Definition 7.4.5. ***(Utility Maximizing Principal)****: A principal is* utility maximizing *if their objective is to select an audit policy α which maximizes the principal's utility. For an agent with true features $\mathbf{x}$, let $\mathbf{z} = \mathbf{x}_R$ if the agent performs recourse and $\mathbf{z} = \mathbf{x}$ otherwise, and let $\mathbf{z}'$ be the agent's report. This objective can be framed as,*

$$\begin{aligned} \alpha^* = \arg\max_{\alpha} \mathbb{E}\big[\hat{y}(\mathbf{z}')f(\mathbf{z})\big(\alpha(\mathbf{z}';\mathbf{Z}')\mathbb{I}\left[\mathbf{z}\neq\mathbf{z}'\right] + \mathbb{I}[\mathbf{z}=\mathbf{z}']\big)\big] \\ s.t.\ \mathbb{E}_{\alpha}\big[\sum_{\mathbf{z}'\in\mathbf{Z}'} \alpha(\mathbf{z}';\mathbf{Z}')|\mathbf{Z}'\big] \leq B \quad \forall\ \mathbf{Z}' \end{aligned} \tag{7.7}$$

Remark 7.4.6. *Note that the objective of incentive minimization (Definition 7.4.3) is independent of agent's ability to perform recourse, while the latter two objectives depend on recourse. Recourse may be infeasible in some domains, which can be captured by setting $c_R = \infty$, while in other domains the ability of agents to perform recourse, or the associated cost of such actions may be unknown. In these cases, note that the incentive to manipulate*

ε, can never be increased (but may be decreased) if the principal misspecifies agents ability to perform recourse.

Prior to characterizing the optimal auditing strategies for each type of principal objective, we first demonstrate that for recourse- and utility-maximization the principal need only consider the equilibrium strategy profile of agents in which the maximum number of agents manipulate. That is, if an audit policy α is recourse- or utility-maximizing when the maximum number of agents manipulate, then that policy is also recourse- or utility-maximizing for *any* other equilibrium strategy profile of agents. Note that this not consequential to the objective of incentive-minimization.

Theorem 7.4.7. *Let $g_{\max}$ be the BNE profile which has the maximum number manipulations. If an audit policy α is recourse (or utility) maximizing with respect to $g_{\max}$, it is recourse (or utility) maximizing for any other BNE profile g.*

Henceforth, we leverage this result to only consider the principal's objective with respect to $g_{\max}$.

Proof. This result follows directly from the characterization of optimal policies when the induced BNE strategy profile of agents is $g_{\max}$ given later in Theorem 7.5.1 and 7.5.2. From the proofs of these theorems, any policy which does not audit positively classified agents uniformly in expectation is suboptimal when agents follow $g_{\max}$. For any strategy profile g induced by the audit policy α, the condition that an agent prefers recourse to manipulation if given as

$$\frac{c_R(\mathbf{x}, \mathbf{x}_R) + u_a(\mathbf{x}) - u_a(\mathbf{x}_R)}{u_a(\mathbf{x}) + C} \leq \min_{\mathbf{x}' \in \mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))$$

any optimal policy will maximize the fraction of agents for which this condition holds (both for recourse and utility maximization). By a similar line of reasoning to Theorems 7.5.1 and 7.5.2, maximizing the minimum value of $\alpha(\mathbf{z}'; g(\mathbf{X}))$ maximizes recourse and system utility. Such an audit policy is precisely the uniform audit policy provided in Theorem 7.5.1. Thus uniform auditing (i.e. the only optimal policy when agent follow $g_{\max}$) is also an optimal policy under any other induced g. □

7.5 Optimal Auditing

We begin by examining the computation of optimal audit policies for each of the three objectives. In particular, we demonstrate that for each objective, the optimal policy can be computed efficiently.

First we examine the recourse- and utility-maximizing principals, in which recourse actions are a salient consideration. Once the optimal policy for these objectives is formulated, we will see that it is straightforward to find a policy which minimizes incentives, even independent of recourse actions.

Theorem 7.5.1. *For any recourse cost function $c_R(\mathbf{x}, \mathbf{z})$, agent utility function $u_a(\mathbf{x})$, feature distribution $\mathcal{D}$, a recourse maximizing principal with budget B and fine C has optimal policy*

$$\alpha(\mathbf{z}'; \mathbf{Z}') = B/|\mathbf{Z}'^{(1)}|, \quad \forall\ \mathbf{z}' \in \mathbf{Z}'^{(1)}, \forall\ \mathbf{X}$$

when agent reports $\mathbf{Z}'$ are induced the BNE profile $g_{\max}$; $\mathbf{Z}'^{(1)}$ is the set of all reports $\mathbf{z}'$ with $\hat{y}(\mathbf{z}') = 1$.

Proof. The objective of a recourse maximizing principal is to maximize the expected number of agents, over realizations $\mathbf{X}$ of agents' true features, who prefer recourse to manipulation, i.e.,

$$\begin{aligned} \alpha^* = \arg\max_{\alpha}\ & \mathbb{E}_{\mathbf{X}}\left[|\mathbf{X}_R^{(0)}|\right] \\ \text{s.t. } & \mathbb{E}_{\alpha}\Big[\sum_{\mathbf{x}' \in g(\mathbf{X})} \alpha(\mathbf{x}' : g(\mathbf{X}))\Big|\ g(\mathbf{X})\Big] \leq B \qquad\qquad \forall\ g(\mathbf{X}) \end{aligned}$$

where

$$\mathbf{X}_R^{(0)} = \big\{\mathbf{x} \in \mathbf{X}^{(0)} :\ u_a(\mathbf{x}) - \alpha(\mathbf{x}_M; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \leq\ u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)\big\}$$

i.e. $\mathbf{X}_R^{(0)} \subset \mathcal{X}^{(0)}$ is the set of agents in $\mathbf{X}$ with true features which are negatively classified by $\hat{y}$. Recall that g is the strategy profile of agents induced by α, i.e. $g = g_{\max}$, and $g(\mathbf{X})$ is the reports given by the agents in $\mathbf{X}$. Since each report is $\mathbf{x}'$, we also refer to $g(\mathbf{X})$ as simply $\mathbf{X}'$ to avoid overly cumbersome notation.

To show that the policy of uniform auditing is optimal in the case of recourse maximization, we examine each possible realization of true features $\mathbf{X}$ independently and show that uniform auditing maximizes the number of agents which perform recourse in each set $\mathbf{X}$. For each agent $\mathbf{x} \in \mathbf{X}$, the optimal manipulation is,

$$\begin{aligned} \mathbf{x}_M =& \arg \max_{\mathbf{x}' \in \mathcal{X}^{(1)}} u_a(\mathbf{x})\big(1 - \alpha(\mathbf{x}'; g(\mathbf{X}))\big) - C\alpha(\mathbf{x}') \\ =& \arg \max_{\mathbf{x}' \in \mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{x}' : g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \end{aligned}$$

The optimal recourse action for $\mathbf{x}$ is

$$\mathbf{x}_R = \arg \max_{\mathbf{z} \in \mathcal{X}^{(1)}} u_a(\mathbf{z}) - c_R(\mathbf{x}, \mathbf{z})$$

Since $\mathbf{x}_R$ is independent of the choice of α, the agent will always choose $\mathbf{x}_R$ if recourse is the optimal action.

Agent $\mathbf{x}$ will choose recourse over manipulation if

$$\begin{aligned} &\max_{\mathbf{x}' \in \mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{x}'; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \\ &\qquad \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R) \\ \iff& u_a(\mathbf{x}) - \alpha(\mathbf{x}_M; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \\ &\qquad \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R) \end{aligned} \tag{7.8}$$

Thus, for realization $\mathbf{X}$, the principal's objective w.r.t to $\mathbf{X}$ is, to select α such that

$$\begin{aligned} \max_{\alpha} \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{I}\Big[& u_a(\mathbf{x}) - \alpha(\mathbf{x}_M; g(\mathbf{X}))(u_a(\mathbf{x}) + C) \\ & \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)\Big] \\ \text{s.t.} \quad & \mathbb{E}\big[\sum_{\mathbf{x}' \in g(\mathbf{X})} \alpha(\mathbf{x}') | g(\mathbf{X})\big] \leq B \end{aligned} \tag{7.9}$$

That is, w.r.t. to $\mathbf{X}$, the principal aims to select α such that Inequality 7.8 holds for the largest fraction of agents in $\mathbf{X}$. In order to select this α, we can rewrite Inequality 7.8 as,

$$\begin{aligned}
&\max_{\mathbf{x}'\in\mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{x}'; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big)\\
&\qquad\qquad \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)\\
\iff\ & u_a(\mathbf{x}) - \min_{\mathbf{x}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big)\\
&\qquad\qquad \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)\\
\iff\ & u_a(\mathbf{x}) - \big(u_a(\mathbf{x}) + C\big) \min_{\mathbf{x}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\\
&\qquad\qquad \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)\\
\iff\ & \frac{c_R(\mathbf{x}, \mathbf{x}_R) + u_a(\mathbf{x}) - u_a(\mathbf{x}_R)}{u_a(\mathbf{x}) + C} \leq \min_{\mathbf{x}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))
\end{aligned}$$

The principal's objective of ensuring that this condition holds for the maximum number of $\mathbf{x} \in \mathbf{X}$ can thus be expressed as,

$$\begin{aligned}
&\max_{\alpha} \min_{\mathbf{x}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\\
&\qquad \text{s.t. } \ \mathbb{E}\big[\sum_{\mathbf{x}'\in g(\mathbf{X})} \alpha(\mathbf{x}')|g(\mathbf{X})\big] \leq B
\end{aligned}$$

Therefore the principal aims to maximize the minimum audit probability $\alpha(\mathbf{x}'; g(\mathbf{X})$ over the entire feature space $\mathcal{X}^{(1)}$. Max-min optimization with homogeneous budget constraint is known to have a uniform solution, namely

$$\alpha^*(\mathbf{x}'; g(\mathbf{X})) = \begin{cases} \frac{B}{|\mathbf{X}'^{(1)}|} & \text{if } \ \mathbf{x}' \in \mathcal{X}^{(1)} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, uniform auditing maximizes the number of agents in $\mathbf{X}$ which prefer recourse over manipulation. Since this holds for any such realization $\mathbf{X}$ and each realization is independent of other realizations, it must be the case that uniform auditing maximizes the number of agents in expectation (expectation over possible realizations $\mathbf{X}$) which prefer recourse over manipulation. □

Theorem 7.5.2. *For any recourse cost function $c_R(\mathbf{x}, \mathbf{z})$, agent utility function $u_a(\mathbf{x})$, feature distribution $\mathcal{D}$, the policy in Theorem 7.5.1 (uniform auditing) is a utility maximizing policy, when the induced BNE profile of agents is $g_{\max}$.*

Proof of Theorem 7.5.2. To prove this theorem we will break the principal's objective into two components (utility gained by incetivizing recourse, and utility gain from successfully auditing manipulations), and show that the policy outlined in Theorem 7.5.1 maximizes both terms. The objective of a utility maximizing principal is given as

$$\alpha^* = \arg\max_{\alpha} \mathbb{E}\Big[\hat{y}(\mathbf{z}')f(\mathbf{z})\big(1-\alpha(\mathbf{z}'; g(\mathbf{X}))\big)\mathbb{I}\left[\mathbf{z} \neq \mathbf{z}'\right] \\ + \hat{y}(\mathbf{z}')f(\mathbf{z})\mathbb{I}[\mathbf{z} = \mathbf{z}']\Big] \\ \text{s.t. } \mathbb{E}_{\alpha}\big[\sum_{\mathbf{z}' \in g(\mathbf{X})} \alpha(\mathbf{z}'; g(\mathbf{X}))|g(\mathbf{X})\big] \leq B \quad \forall\, g(\mathbf{X})$$

For any agent $\mathbf{x}$ let $\mathbf{z}$ be the agents recourse action (i.e. $\mathbf{z} = \mathbf{x}$ if no recourse occurs) and $\mathbf{z}'$ be the agents report (i.e. $\mathbf{z}' \neq \mathbf{z}$ if the agent manipulates). When using audit policy α, which induces BNE profile g for agents, the principal's utility w.r.t. to $\mathbf{x}$ is

$$\begin{aligned} U(\mathbf{x}; \alpha, g) &= \hat{y}(\mathbf{z}')f(\mathbf{z})\big(1-\alpha(\mathbf{z}'; g(\mathbf{X}))\big)\mathbb{I}\left[\mathbf{z} \neq \mathbf{z}'\right] \\ &\quad + \hat{y}(\mathbf{z}')f(\mathbf{z})\mathbb{I}[\mathbf{z} = \mathbf{z}'] \\ &= \hat{y}(\mathbf{z}')f(\mathbf{z})\Big(\big(1-\alpha(\mathbf{z}'; g(\mathbf{X}))\big)\mathbb{I}\left[\mathbf{z} \neq \mathbf{z}'\right] + \big(1-\mathbb{I}\left[\mathbf{z}' \neq \mathbf{z}\right]\big)\Big) \\ &= \hat{y}(\mathbf{z}')f(\mathbf{z})\big(1-\alpha(\mathbf{z}'; g(\mathbf{X}))\mathbb{I}\left[\mathbf{z} \neq \mathbf{z}'\right]\big) \end{aligned}$$

Note that any agent $\mathbf{x}$ with $\mathbf{x} \in \mathcal{X}^{(1)}$ (i.e. $\hat{y}(\mathbf{x}) = 1$) has no incentive to perform any action other than to truthfully report, i.e. $\mathbf{x} \in \mathcal{X}^{(1)} \implies \mathbf{x} = \mathbf{z} = \mathbf{z}'$ for any α and g. Therefore, we need only consider the agents $\mathbf{x} \in \mathcal{X}^{(0)}$ (i.e. $\hat{y}(\mathbf{x}) = 0$) when computing the principal's optimal α. Since $\mathbf{x} \in \mathcal{X}^{(0)} \iff f(\mathbf{x}) < 0$, we can decompose the principal's utility into two cases

$$U(\mathbf{x}; \alpha, g) = \begin{cases} \hat{y}(\mathbf{z}')f(\mathbf{z})\big(1-\alpha(\mathbf{z}'; g(\mathbf{X}))\big) & \text{if } \; f(\mathbf{z}) < 0 \\ \hat{y}(\mathbf{z}')f(\mathbf{z}) & \text{if } \; f(\mathbf{z}) \geq 0 \end{cases}$$

Given this formulation of the principal's utility w.r.t. to $\mathbf{x} \in \mathcal{X}^{(0)}$, consider the three possible actions of $\mathbf{x}$: 1.) $\mathbf{x}$ performs recourse to $\mathbf{x}_R$, 2.) $\mathbf{x}$ misreports $\mathbf{x}'$, and 3.) $\mathbf{x}$ truthfully reports $\mathbf{x}$ (i.e. do-nothing). In case (1), the principal gains utility $f(\mathbf{x}_R) \geq 0$, in case (2) the principal gets utility $f(\mathbf{x})\big(1 - \alpha(\mathbf{x}'; g(\mathbf{X}))\big) \leq 0$, and in case (3) the principal gets utility 0. There are two important observations to make with respect to theses cases. First note that only in case (1) can the actions of $\mathbf{x}$ yield the principal positive utility. Second, an agent's preference between actions of case (3) and case (1) are independent of the audit policy α. To see this, note that $u_a(\mathbf{x}_R)$ and $c_R(\mathbf{x}, \mathbf{x}_R)$ (the values which determine the costs of (1) and (2)), are independent of α. Therefore for each $\mathbf{x} \in \mathcal{X}^{(0)}$ there are only two cases that the principal need consider: either the agent manipulates, i.e. case (2), or the agent selects (1) or (3), the selection of which is uniquely determined by $\mathbf{x}$.

We need only consider two possible actions for both types of agents in $\mathcal{X}^{(0)}$: a) recourse or manipulation when $c_R(\mathbf{x}, \mathbf{x}_R) < u_a(\mathbf{x}_R)$, and b) do-nothing or manipulation when $c_R(\mathbf{x}, \mathbf{x}_R) \geq u_a(\mathbf{x}_R)$. For an agent $\mathbf{x}$ with $c_R(\mathbf{x}, \mathbf{x}_R) < u_a(\mathbf{x}_R)$, the agent will choose recourse over manipulation if and only if

$$\begin{gathered}\max_{\mathbf{x}' \in \mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{x}'; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \\ \leq u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R) \\ \iff \frac{c_R(\mathbf{x}, \mathbf{x}_R) + u_a(\mathbf{x}) - u_a(\mathbf{x}_R)}{u_a(\mathbf{x}) + C} \leq \min_{\mathbf{x}' \in \mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\end{gathered}$$

which is precisely the condition of the recourse maximizing principal found in the proof of Theorem 7.5.1. Similarly, for an agent $\mathbf{x}$ with $c_R(\mathbf{x}, \mathbf{x}_R) \geq u_a(\mathbf{x}_R)$, the agent will choose do-nothing over manipulation if and only if

$$\begin{gathered}\max_{\mathbf{x}' \in \mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{x}'; g(\mathbf{X}))\big(u_a(\mathbf{x}) + C\big) \leq 0 \\ \iff \frac{u_a(\mathbf{x})}{u_a(\mathbf{x}) + C} \leq \min_{\mathbf{x}' \in \mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\end{gathered}$$

In both cases, the principal can prevent the agent from manipulation by sufficiently increasing the minimum value of α. The condition that any agent $\mathbf{x}$ will not manipulate can then be written succinctly as,

$$\begin{aligned}&\min\left(\frac{u_a(\mathbf{x})}{u_a(\mathbf{x})+C}, \frac{c_R(\mathbf{x},\mathbf{x}_R)+u_a(\mathbf{x})-u_a(\mathbf{x}_R)}{u_a(\mathbf{x})+C}\right)\\ &\leq \min_{\mathbf{x}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{x}'; g(\mathbf{X}))\end{aligned}$$

The principals utility is monotonically increasing in the number of agents $\mathbf{x}$ for which the above condition holds. Thus, momentarily ignoring the principal's ability to increase their utility by successfully catching (and subsequently denying) manipulated reports, the principal's objective reduces to maximize the minimum value of α across all features in $\mathcal{X}^{(1)}$, which is precisely the objective of a recourse maximizing principal.

Thus it remains only to show that uniform auditing is also the policy which results in the largest utility gain from successfully catching manipulated reports. To see this, consider any agent $\mathbf{x} \in \mathcal{X}^{(0)}$. The optimal manipulation for this agent is

$$\begin{aligned}\mathbf{x}_M &= \arg\max_{\mathbf{z}'\in\mathcal{X}^{(1)}} u_a(\mathbf{x}) - \alpha(\mathbf{z}'; g(\mathbf{X}))\big(u_a(\mathbf{x})+C\big)\\ &= \arg\min_{\mathbf{z}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{z}'; g(\mathbf{X}))\big(u_a(\mathbf{x})+C\big)\\ &= \arg\min_{\mathbf{z}'\in\mathcal{X}^{(1)}} \alpha(\mathbf{z}'; g(\mathbf{X}))\end{aligned}$$

That is, any agent choosing manipulation, will select the report $\mathbf{z}'$ corresponding to the feature in $\mathcal{X}^{(1)}$ with the lowest probability of being audited in expectation. Let $\mathbf{z}_{\min} \in \mathcal{X}^{(1)}$ be the feature with the lowest probability of being audited for an audit policy α. Then $\mathbf{x}_M = \mathbf{z}_{\min}$ for every $\mathbf{x}$, i.e. it is always optimal to misreport $\mathbf{z}_{\min}$. In the best possible case for the principal, no agents would preform recourse to $\mathbf{z}_{\min}$, in which case all reports $\mathbf{z}' = \mathbf{z}_{\min}$ are manipulations. Then, for any particular agent reporting $\mathbf{z}_{\min}$, the probability of that agent being audited is strictly lower than the probability that a truthful report $\mathbf{z}' \neq \mathbf{z}_{\min}$ being audited (due to the fact that $\mathbf{z}_{\min}$ has the lowest probability of being audited and α is not uniform). Under uniform auditing, every positively classified report has equal probability of being audited. Hence, the expected number of audit manipulations under this non-uniform policy α must be strictly less than the expected number of audited manipulations under uniform auditing. Therefore, uniform auditing is the policy which catches the maximum amount of manipulations.

Combining this with the fact that uniform auditing also maximizes the number of agents not manipulating (i.e. maximizing the number of agents which yield the principal nonnegative utility), uniform auditing is optimal. □

Remark 7.5.3. *Theorems 7.5.4, 7.5.1, and 7.5.2 show an equivalence between an incentive-minimizing, a recourse-maximizing, and a utility-maximizing principal. The equivalence of the latter two is significant for three primary reasons: (1) the actions of a self-interested (utility-maximizing) principal are as beneficial to the population as the actions of a recourse-maximizing principal directly trying to maximize for population benefit, (2) self-interested auditing decreases the percentage of agents which engage in "risky" and potentially socially detrimental behavior (manipulation), and (3) optimal auditing* does not *require any knowledge of dynamics of agents recourse actions (e.g. solving Program 7.3, or even knowing c_R).*

Lastly we examine the an incentive-minimizing principal. Note that the objective of the incentive-minimizing principal is independent of recourse actions.

Theorem 7.5.4. *For any agent utility function $u_a(\mathbf{x})$, feature distribution $\mathcal{D}$, an incentive-minimizing principal with budget B and fine C has optimal policy*

$$\alpha(\mathbf{z}'; \mathbf{Z}') = B/|\mathbf{Z}'^{(1)}|, \quad \forall\ \mathbf{z}' \in \mathbf{Z}'^{(1)}, \forall\ \mathbf{X}$$

when agent reports $\mathbf{Z}'$ are induced the ε-BNE profile of truthful reporting; $\mathbf{Z}'^{(1)}$ is the set of all reports $\mathbf{z}'$ with $\hat{y}(\mathbf{z}') = 1$.

Proof. This proof follows from a similar line of reason to the case of recourse- and utility-maximization. Since agent utility depends only on the agent's true feature $\mathbf{x}$, all manipulations which result in positive classification are equal to the agent (barring auditing). Thus, the only difference in expected utility from reporting any $\mathbf{z}' \in \mathcal{X}^{(1)}$ stems from the principal's audit policy. Moreover, the difference in utility is precisely the difference in audit probability, i.e., reporting $\mathbf{z}'$ yields

$$u_a(\mathbf{x}) - \alpha(\mathbf{z}'))\big(u_a(\mathbf{x} - C\big).$$

All terms are constant in $\mathbf{z}'$ with the exception of $\alpha(\mathbf{z}')$, and thus the agent's optimal manipulation is again

$$\mathbf{x}_M = \arg\min_{\mathbf{z}'} \alpha(\mathbf{z}')$$

indicating that if any feature $\mathbf{z}'$ was audited with a probability lower than other features, then the optimal strategy for agents is to misreport $\mathbf{z}'$, independent of the other agents actions. As such, an optimal audit policy must audit all positively classified features with equal probability. □

7.6 Auditing With Subsides

Audits provide a punitive measure for incentivizing recourse over manipulation. Another natural option is to offer subsidies that make recourse cheaper to implement for agents. Here we investigate how the principal optimally splits the limited budget between auditing and subsidies. For example, a bank may choose to allocate a fraction of their budget from application verification to the development of educational material to help increase financial literacy. Our key result is that in the important special case of constant utilities, both recourse-maximizing and (own) utility-maximizing principals choose the same fraction of budget for subsidies. Moreover, we show that despite the complex interdependencies of the problem, when agent utilities are constant, the objective of both principals can be formulated as a single-dimensional optimization problem, depending only on the impact of subsidies on the cost of recourse and audit budget. We begin by formalizing subsidies in our model.

Definition 7.6.1. *A* subsidy function $s : [0, B] \to [0, 1]$ *yields a multiplicative decrease in the cost of recourse, such that for a subsidy budget b, the cost of recourse becomes $s(b)c_R(\mathbf{x}, \mathbf{z})$, and the remaining budget $B-b$ is then used for auditing. Subsidy functions $s(b)$ are decreasing in b and $s(0) = 1$ (allocating no subsidies recovers the original recourse cost).*

Remark: *For any subsidy trade-off b^* with $s(b^*) = 0$, the cost of recourse is $s(b^*)c_R(\mathbf{x}, \mathbf{z}) = 0$ for all $\mathbf{x}, \mathbf{z}$. When such a trade-off exists, it is always optimal for the principal to select b^* as their subsidy allocation (i.e., their objective reduces to univariate root finding of $s(b)$). Consequently, we henceforth assume that $s(b) > 0$.*

Next, we present our key result showing that when agent utilities are constant, optimal subsidy characterization is identical for either recourse- or utility-maximizing principal, and amounts to solving a one-dimensional optimization problem.

Theorem 7.6.2. *Suppose that agent utilities are constant, i.e., $u_a(\mathbf{x}_i) = \bar{u}_a$, and the induced BNE profile of agents is $g_{\max}$. Then, for both a recourse-maximizing and utility-maximizing* *principal, the optimal subsidy is given by*

$$b^* = \arg\max_{b\in[0,B]} \frac{B-b}{s(b)} \tag{7.10}$$

Proof. When the principal chooses a subsidy trade-off of b, their resulting audit budget is $B-b$, and the cost of recourse is $s(b)c_R(\mathbf{x},\mathbf{z})$. For a given subsidy trade-off b, each agent with true feature $\mathbf{x}$ has optimal recourse action

$$\mathbf{x}_R = \arg\max_{\mathbf{z}\in\mathcal{X}^{(1)}} u_a(\mathbf{z}) - s(b)c_R(\mathbf{x},\mathbf{z})$$

Let b_1, b_2 be any two subsidy trade-offs and $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{X}^{(1)}$ be any two recourse actions for an agent with true features $\mathbf{x}$. Then,

$$\begin{aligned} & u_a(\mathbf{z}_1) - s(b_1)c_R(\mathbf{x},\mathbf{z}_1) \leq u_a(\mathbf{z}_2) - s(b_1)c_R(\mathbf{x},\mathbf{z}_2) \\ \implies & s(b_1)c_R(\mathbf{x},\mathbf{z}_1) \geq s(b_1)c_R(\mathbf{x},\mathbf{z}_2) \\ \implies & s(b_2)c_R(\mathbf{x},\mathbf{z}_1) \geq s(b_2)c_R(\mathbf{x},\mathbf{z}_2) \\ \implies & u_a(\mathbf{z}_1) - s(b_2)c_R(\mathbf{x},\mathbf{z}_1) \leq u_a(\mathbf{z}_2) - s(b_2)c_R(\mathbf{x},\mathbf{z}_2) \end{aligned}$$

implying that the optimal recourse action $\mathbf{x}_R$ is invariant w.r.t. to the subsidy trade-off b. Therefore, for any fixed b, the problem of optimal auditing with either objective reduces down to exactly the cases studied in Section 7.4 for which uniform auditing is optimal. An agent with features $\mathbf{x}$ will prefer recourse over manipulation when

$$\begin{aligned} & \max_{\mathbf{x}'\in\mathcal{X}^{(1)}} u_a(\mathbf{x}) \text{ - } \alpha(\mathbf{x}')\,(u_a(\mathbf{x}){+}C) \\ & \leq u_a(\mathbf{x}_R) \text{ - } s(b)c_R(\mathbf{x},\mathbf{x}_R) \\ \iff & s(b)c_R(\mathbf{x},\mathbf{x}_R) \leq \mathbb{E}\left[\frac{B-b}{|\mathbf{X}'^{(1)}|}\right](u_a(\mathbf{x}){+}C) \\ \iff & \frac{c_R(\mathbf{x},\mathbf{x}_R)}{(u_a(\mathbf{x})+C)} \leq \mathbb{E}\left[\frac{1}{|\mathbf{X}'^{(1)}|}\right]\frac{B-b}{s(b)} \end{aligned} \tag{7.11}$$

For a recourse maximizing principal it is straightforward to see that maximizing the number of features for which Inequality 7.11 holds is equivalent to maximizing recourse. In the case of a utility maximizing principal, the argument is identical to that of the proof of Theorem 7.5.2, in which maximizing the number of agents performing recourse is shown to also maximize the principals expected utility.

For any set of true features $\mathbf{X}$, each of the reports in $\mathbf{X}'$ falls under one of the following categories: 1.) the reported feature $\mathbf{x}'$ is truthful and positively classified, i.e. $\mathbf{x}' = \mathbf{x}$ and $\hat{y}(\mathbf{x}) = 1$, 2.) $\mathbf{x}'$ is a manipulation which is positively classified, i.e. $\mathbf{x}' \neq \mathbf{x}$ and $\hat{y}(\mathbf{x}') = 1$, 3.) $\mathbf{x}'$ is the result of recourse and is positively classified, i.e. $\mathbf{x}' = \mathbf{x}_R$ and $\hat{y}(\mathbf{x}_R) = 1$, and 4.) $\mathbf{x}'$ is truthful and negatively classified, i.e. $\mathbf{x}' = \mathbf{x}$ and $\hat{y}(\mathbf{x}') = 0$. Note that $\mathbf{X}'^{(1)}$ is comprised entirely of reports of type (1), (2), and (3). Moreover, the value of b will not change reports of type (1) and occur with probability $\mathbb{P}\big(\hat{y}(\mathbf{x}) = 1\big)$.

For reports of type (4) the agent is negatively classified and does not perform recourse or manipulation, implying that both are too costly, i.e. the following holds for feature $\mathbf{x}$,

$$u_a(\mathbf{x}) \leq \alpha(\mathbf{x}')(u_a(\mathbf{x}) + C) \;\text{ and }\; u_a(\mathbf{x}_R) \leq s(b)c_R(\mathbf{x}, \mathbf{x}_R)$$

The value of $s(b)c_R(\mathbf{x}, \mathbf{x}_R)$ is monotonically decreasing in b, therefore increasing b could not push a report of type (3) to type (4). Similarly the value of $\alpha(\mathbf{z}'; g(\mathbf{X}))$ is monotonically decreasing in b (since $B - b$ is monotonically decreasing in b), and therefore increasing b could not push a report of type of (2) to type (4). Thus the number of reports of type (4) is monotonically decreasing in b.

Given a realization of agent features $\mathbf{X}$, let $\mathbf{X}_0'^{(1)}$ be the set of positively classified reports when $b = 0$, i.e. when the principal audits without subsides. Then for any $b > 0$, let $\mathbf{X}_b'^{(1)}$ be the set of reports which are positively classified when the principal uses subsidies. Then $\mathbf{X}_0'^{(1)} \subset \mathbf{X}_b'^{(1)}$, implying that

$$\mathbb{E}\left[\frac{1}{|\mathbf{X}_b'^{(1)}|}\right] \leq \mathbb{E}\left[\frac{1}{|\mathbf{X}_0'^{(1)}|}\right],$$

and

$$\mathbb{E}\left[\frac{1}{|\mathbf{X}_b'^{(1)}|}\right]\frac{B-b}{s(b)} \leq \mathbb{E}\left[\frac{1}{|\mathbf{X}_0'^{(1)}|}\right]\frac{B-b}{s(b)}$$

thus, for any $b > 0$ Inequality 7.11 is also satisfied when

$$\frac{c_R(\mathbf{x}, \mathbf{x}_R)}{(u_a(\mathbf{x}){+}C)} \leq \mathbb{E}\left[\frac{1}{|\mathbf{X}_b'^{(1)}|}\right]\frac{B-b}{s(b)} \leq \mathbb{E}\left[\frac{1}{|\mathbf{X}_0'^{(1)}|}\right]\frac{B-b}{s(b)}$$

Therefore the principal's objective can be framed as maximizing $\mathbb{E}\left[\frac{1}{|\mathbf{X}_0'^{(1)}|}\right]\frac{B-b}{s(b)}$. Since $\mathbb{E}\left[\frac{1}{|\mathbf{X}_0'^{(1)}|}\right]$ is independent of b, the principal's objective is thus

$$\max_{b\in[0,B]} \frac{B-b}{s(b)}$$

Therefore the subsidy trade given in the in Theorem 7.6.2 is both recourse maximizing and utility maximizing. □

Illustration: *To gain some intuition into the result of Theorem 7.6.2, consider $s(b) = \frac{1}{b+1}$, where the impact of subsidies on recourse costs exhibits diminishing returns in the subsidy allocation. In this case, the objective can be solved analytically, obtaining the optimal subsidy $b^* = \frac{B-1}{2}$. Thus, the principal, whether maximizing overall welfare or their own utility, would allocate nearly half of the audit budget to subsidies. The reason is that even a self-interested principal actually benefits from providing subsidies and thereby incentivizing recourse, as such actions also increase the principal's profits, whereas manipulation results in an expected loss.*

Corollary 7.6.3. *When agent utility is constant, both a recourse-maximizing and utility-maximizing principal will allot a nonzero portion of their budget to subsides if and only if there exists some b s.t. $s(b) \leq 1 - b/B$, i.e. s has better than linear scaling for at least one value of b.*

In contrast to the case of constant agent utilities, however, optimal subsidy becomes non-trivial for general agent utilities. Moreover, the alignment between recourse- and utility-maximizing principal no longer obtains.

Theorem 7.6.4. *For general agent utilities, recourse maximization and utility maximization are no longer aligned.*

Proof. We present a counter-example showing a case (or more broadly a family of cases) in which the two objectives are not aligned. Let $B = 3$ be the principal's total budget for auditing and subsidies, let the manipulation fine be $C = 2$, and we will specify the subsidy function $0 < s(b) \leq 1$ for each case below.

Consider the case when both agent's utility u_a and the principal's utility u_p are *heterogeneous*, namely $u_a(x) \neq u_a(x'), u_p(x) \neq u_p(x')$ for $x \neq x'$. Suppose there are 5 different features:

x_0, x_1, x_2, x_3, x_4, where x_0, x_1, x_2 are negatively classified and x_3, x_4 are positively classified. Assume four agents, each with feature x_0, x_1, x_2, x_3 accordingly. Thus the first three agents get denied by the classifier, while the last agent gets approved. All the agent's utility and the principal's utility are specified as follows:

$$\begin{aligned} u_a(x_0) = 0.5, u_a(x_1) = 1, u_a(x_2) = 1, \\ u_a(x_3) = \frac{91}{128}, u_a(x_4) = \frac{75}{64} \\ u_p(x_0) = u_p(x_1) = u_p(x_2) = 0, \\ u_p(x_3) = 0.1, u_p(x_4) = 10 \end{aligned}$$

The costs of recourse for each negatively classified feature are specified as follows:

$$\begin{aligned} c_R(x_0, x_3) = \frac{1}{2}, c_R(x_0, x_4) = 1 \\ c_R(x_1, x_3) = \frac{1}{4}, c_R(x_1, x_4) = 1 \\ c_R(x_2, x_3) = \frac{1}{2}, c_R(x_2, x_4) = 1 \end{aligned}$$

Consider the optimal policies for the two principals.

For a recourse maximizing principal, the optimal subsidy trade-off is to set $b_r^* = 0$, which implies that $s(b_r^*) = 1$, namely they will spend all budget $B = 3$ on auditing and none on subsidies is optimal. The corresponding audit audit policy is $\alpha_r = \frac{B - b_r^*}{|X'^{(1)}|} = \frac{3}{4}$. With the above specification, the optimal actions for all three agents are to perform recourse to feature x_3. To see this, we provide one example using the agent with feature x_0:

$$\begin{aligned} &\text{(cost of manipulation): } u_a(x_0) - \alpha_r(u_a(x_0) + C) < 0 \\ &\text{(cost of recourse to } x_3\text{): } u_a(x_3) - s(b_r^*)c_R(x_0, x_3) = \frac{27}{128} \\ &\text{(cost of recourse to } x_4\text{): } \ u_a(x_4) - s(b_r^*)c_R(x_0, x_4) = \frac{11}{64} \end{aligned}$$

In this case, the recourse maximizing principal successfully incentivizes all three negatively classified agents to perform recourse, achieving recourse maximization. Meanwhile, the total principal utility is $U_p^r = \sum_{i=1}^{4} u_p(x_i) = 0.4$.

In contrast to the recourse maximizing principal, a utility maximizing principal has subsidy trade-off $b_u^* = 2$ and thus the subsidy function is $s(b_u^* = 2) = \frac{27}{32}$. We can compute the optimal audit policy as $\alpha_u = \frac{B - b_u^*}{|X'^{(1)}|} = \frac{3-2}{4} = \frac{1}{4}$. In this case, the agent with feature x_0's optimal action is to perform recourse to x_4, while agents with feature x_1 and x_2 will perform manipulation. To see this, we provide the calculation for the agent with feature x_0:

$$\begin{aligned}
&\text{(cost of manipulation): } u_a(x_0) - \alpha_u(u_a(x_0) + C) < 0 \\
&\text{(cost of recourse to } x_3\text{): } u_a(x_3) - s(b_u^*)c_R(x_0, x_3) = \frac{37}{128} \\
&\text{(cost of recourse to } x_4\text{): } \;\; u_a(x_4) - s(b_u^*)c_R(x_0, x_4) = \frac{11}{64}
\end{aligned}$$

For the agent with feature x_1:

$$\begin{aligned}
&\text{(cost of manipulation): } u_a(x_1) - \alpha_u(u_a(x_1) + C) = \frac{11}{16} \\
&\text{(cost of recourse to } x_3\text{): } u_a(x_3) - s(b_u^*)c_R(x_1, x_3) = \frac{1}{2} \\
&\text{(cost of recourse to } x_4\text{): } \;\; u_a(x_4) - s(b_u^*)c_R(x_1, x_4) = \frac{21}{64}
\end{aligned}$$

In this case, the principal's total utility is $\sum_{i=1}^{4} u_p(x_i) = 10.1$, which is larger than the utility from the recourse maximizing principle's utility even though the total number of agents who performs recourse is only 1.

The above example shows that the two principals' objectives are not aligned when agent utility is non-constant. □

7.7 Costs of Auditing to the Population

In domains where recourse is a salient consideration, it is natural to examine the average cost suffered by a population when performing recourse [110]. With the introduction of auditing and subsides into such domains, it becomes imperative to consider costs/fines imposed on the population as both a function of auditing and subsides.

We first describe the differences between the impact on the utility of the principal and that of the agents. In particular, as the auditing budget B and fine C increase, the principal's utility gain is monotonically increasing, while the agents' utility gain is monotonically decreasing.

Theorem 7.7.1. *Average agent utility is monotonically decreasing in B and C. In contrast, the principal's expected utility is monotonically increasing in B and C.*

Proof. The expected utility of and agent $\mathbf{x}$ misreporting $\mathbf{z}'$ can be expressed as

$$u_a(\mathbf{x}) - \mathbb{E}\big[B/|\mathbf{Z}'^{(1)}|\big]\,(u_a(\mathbf{x}) + C)$$

which is monotonically decreasing in both B and C. The utility of recourse is invariant w.r.t. B and C. Agents only perform recourse if manipulation yields lower utility gain, thus agent utility gain is monotone decreasing in B and C. A symmetric argument can be made for the principal's utility. □

Theorem 7.7.2. *When agent utility is constant, the expected number of agents who either choose to perform recourse or truthfully report is $nF_R\left(\min\left(\bar{u}_a, \frac{B(C+\bar{u}_a)}{n}\right)\right)$.*

Proof. This follows directly from Theorem 7.5.1. □

Lastly, we bound the fines paid by agents when the principal has budget B and the fine is C.

Theorem 7.7.3. *Let $F_R(k) = \mathbb{P}\big(c_R(\mathbf{x}, \mathbf{x}_R) \le k\big)$ (CDF of c_R). Suppose agent utility is constant, define $C' \equiv C + \bar{u}_a$, and let A_M be the expected fines paid by agents. Then,*

$$BC\big(1 - F_R(2BC'/n)\big) \le A_M \le BC2\big(1 - F_R(BC'/n)\big)$$

Theorem 7.7.3 can be interpreted as quantifying the fines paid by the population in terms of how costly recourse is (i.e., the growth rate of F_R). If the principal audits B manipulating agents, the population pays $C \cdot B$. The terms $1 - F_R(2BC'/n)$ and $2(1 - F_R(BC'/n))$, in turn, approximate the probability that a given audit was conducted on a manipulating agent.

These bounds also express the parabolic nature of the fines paid by agents. For small B and C, the fines paid by agents are small (even if all agents manipulate). For large B and C, the cost of manipulation is sufficiently high that few agents will manipulate, and thus, average fines are small. It is the *intermediate* range of values of B and C for which both BC (fines paid when all audits are successful) and $1 - F_R(BC'/n)$ (probability of a successful audit) are large.

Proof. Given a set of n reports $\mathbf{X}'$, let n_M be the number of reports which are manipulations and $|\mathbf{X}'^{(1)}|$ be the number of reports which are approved prior to auditing. Each report in $\mathbf{X}'^{(1)}$ has an equal probability of being audited, namely $\frac{B}{|\mathbf{X}'^{(1)}|}$. For any set of reports $\mathbf{X}'$ the expected number of caught manipulations is,

$$\sum_{\mathbf{z}' \in \mathbf{X}'^{(1)}} \mathbb{E}\big[\mathbb{I}[\mathbf{z}' \neq \mathbf{z}_i]\alpha(\mathbf{z}'; g(\mathbf{X}))\big] = \frac{Bn_M}{|\mathbf{X}'^{(1)}|}$$

Since B is constants it remains only to bound the value of $\frac{n_M}{|\mathbf{X}'^{(1)}|}$, which can be done by examining its expectation with respect to agent reports,

$$\mathbb{E}\bigg[\frac{n_M}{n}\bigg] \leq \mathbb{E}\bigg[\frac{n_M}{|\mathbf{X}'^{(1)}|}\bigg] \leq \mathbb{E}\bigg[\frac{n_m}{|\mathbf{X}^{(1)}| + 1}\bigg] \tag{7.12}$$

Where the left-hand side is due to the fact that the number of approved reports $|\mathbf{X}'^{(1)}|$ cannot exceed the number of agents n, i.e. $|\mathbf{X}'^{(1)}| \leq n$. The right-hand side of the inequality is due to the fact that the number approved reports $|\mathbf{X}'^{(1)}|$ must be greater than the number of agents whose true features would be approved $|\mathbf{X}^{(1)}|$, and if $|\mathbf{X}'^{(1)}| = |\mathbf{X}'^{(1)}|$, then no agents manipulated and $n_M = 0$, in such a case the fines paid by agents is 0 and thus if any fines are paid, it must be the case that $|\mathbf{X}^{(1)}| + 1 \leq |\mathbf{X}'^{(1)}|$.

Defining $C' \equiv C + \bar{u}_a$, we can examine the left-hand side of Inequality 7.12 as,

$$\begin{aligned}
&\mathbb{E}\bigg[\frac{n_M}{|\mathbf{X}'^{(1)}|}\bigg] \geq \mathbb{E}\bigg[\frac{n_M}{n}\bigg] \\
&\geq \frac{n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x}, \mathbf{x}_R) \geq \alpha(\mathbf{x}_M; g(\mathbf{X}))C'\big)}{n} \\
&\geq \frac{n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x}, \mathbf{x}_R) \geq \alpha(\mathbf{x}_M; g(\mathbf{X}))C'\big)}{n} \\
&\geq \frac{n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x}, \mathbf{x}_R) \geq 2BC'/n\big)}{n} \\
&= 1 - F_R\big(2BC'/n\big)
\end{aligned}$$

Examining the right-hand Inequality 7.12 yields,

$$
\begin{aligned}
\mathbb{E}\left[\frac{n_M}{|\mathbf{X}'^{(1)}|}\right] &\le \mathbb{E}\left[\frac{n_m}{|\mathbf{X}^{(1)}|+1}\right] \\
&\le n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x},\mathbf{x}_R) \ge \alpha(\mathbf{x}_M; g(\mathbf{X}))C'\big)\left(\frac{1-\big(1-\mathbb{P}(\mathbf{x}\in\mathcal{X}^{(1)})\big)^{(n+1)}}{(n+1)\mathbb{P}(\mathbf{x}\in\mathcal{X}^{(1)})}\right) \\
&\le n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x},\mathbf{x}_R) \ge BC'/n\big)\left(\frac{1-\big(1-1/2\big)^{(n+1)}}{(n+1)1/2}\right) \\
&\le n\ \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x},\mathbf{x}_R) \ge BC'/n\big)\frac{1}{n/2} \\
&= 2\big(1-F_R(BC'/n)\big)
\end{aligned}
$$

Therefore, we can rewrite Inequality 7.12 as,

$$
1-F_R\big(2BC'/n\big) \le \mathbb{E}\left[\frac{n_M}{|\mathbf{X}'^{(1)}|}\right] \le 2\big(1-F_R(BC'/n)\big)
$$

Thus, the fines paid by agents can be bounded by

$$
BC\big(1-F_R\big(2BC/n\big)\big) \le A_M \le BC\ 2\big(1-F_R(BC/n)\big)
$$

□

7.8 Social Burden and Auditing

Next, we discuss the relationship between auditing and traditional methods for achieving robustness, focusing on their impact on populations. As discussed in the previous section, auditing imposes a non-trivial amount of fines on the population. However, these fines are imposed only on agents that choose to manipulate. Approaches to designing models robust to strategic behavior typically involve modifications to the classifier itself [41, 31, 87]. These methods are similar to adversary simulation approaches found in adversarial machine learning [122, 102, 113]. They attempt to find a pseudo-equilibrium between a model designer and an attacker (or a population of strategic agents in our case). However, these types of retraining methods can have undesirable consequences in the context of strategic classification.

Auditing offers several unique advantages over traditional robustness approaches in strategic classification, such as retraining [73] and *boundary smoothing* [41]. As pointed out by [87], the aforementioned methods almost exclusively result in classifiers with lower qualification rates. This, in turn, necessitates that agents engage in strategic behavior to maintain approval. The cost incurred by such agents is defined as *social burden*, which has been shown to disproportionately affect disadvantaged groups [87]. In the case of our auditing schemes, we demonstrate that optimal audit policies impose fines equitably between groups. Specifically, among all agents who elect to manipulate, each group faces an equal expected average fine. Since our auditing schemes are model-agnostic, they can be applied in both group-fair learning and group-agnostic learning scenarios.

Note that in the case of strategic agents and binary classification, manipulations are unidirectional; agents will only manipulate to achieve positive classification. Thus if f is to be robust to an agent with features $\mathbf{x}$, cost function c, and budget B, then it must be the case that $c(\mathbf{x}, \mathbf{x}') > B$ for all $\mathbf{x}'$ with $f(\mathbf{x}') = 1$. In such a case, if any agent $\mathbf{z} \in \{\mathbf{x}' \in \mathcal{X} : c(\mathbf{x}, \mathbf{x}') \leq B\}$ was positively classified prior to robust training, they would lose positive classification after robust training.

This outcome can be undesirable for several reasons, as discussed previously. These include issues related to selectivity and fairness reversals as well as those concerning individual welfare. Moreover, this type of retraining necessitates that agents engage in strategic behavior to maintain positive classification. For example, the aforementioned agent with features $\mathbf{z}$ would need to engage in strategic behavior to maintain a positive classification. This concept is best captured through the term *social burden*, as first discussed in [87]. This term measures the average cost that agents must pay to maintain their positive classification. We formalize this observation with the following *self-evident* theorem.

Theorem 7.8.1. *For any fine C, cost of recourse c_R, audit budget B, auditing imposes no social-burden. That is, the cost that any agent must pay to maintain positive classification when auditing is deployed, compared to truthfully reporting when auditing is not deployed, is always 0.*

7.9 Tractability of Auditing

Prior to presenting experimental results for auditing we first note an interesting relationship between the tractability of executing an optimal audit policy and computing the efficacy of that policy. For each of the three possible objectives (incentive-minimization, recourse-maximization, and utility-maximization), we saw that uniformly auditing all positively classified agents was optimal, implying that computing optimal audit policies is linear in the number of agents. However, despite the ease at which such policies can be computed and executed, it is intractable for the principal to verify their success. More specifically, in the case of incentivize-minimization, computing the minimum value of ε for which the the model is ε-IC is NP-hard. Similarly, in the cases of recourse-maximization and utility-maximization it is intractable for the principal to compute their expected objective value when deploying the optimal audit policy (fraction of agents choosing recourse and average system utility respectively).

Theorem 7.9.1. *For each type of principal (incentive-minimization, recourse-maximization, and utility-maximization) it is NP-hard to compute the optimal value of the principal's objective, i.e., Equations 7.5, 7.6, and 7.7.*

Since the optimal audit policy uniformly audits all positively classified agents, the probability of any specific positively classified agent being audited is monotonically decreasing in the expected number of positively classified individuals. Thus, determining the expected number of positively classified individuals is key in being able to compute the value of each objective (as these objectives depend on the number of agents electing to manipulate). However, in general computing the expected positive rate of a classifier is NP-hard; a trivial example being the case of binary features and a classifier corresponding to a Boolean formula. A full proof is provided in Section B of the Appendix.

7.10 Experiments

We conduct experiments using the datasets outlined in Chapter 4. Our experimental setup remains the same as previous sections, with the exception of needing to define agent and system utility. In the Adult Income and Law School datasets, agents have constant utility

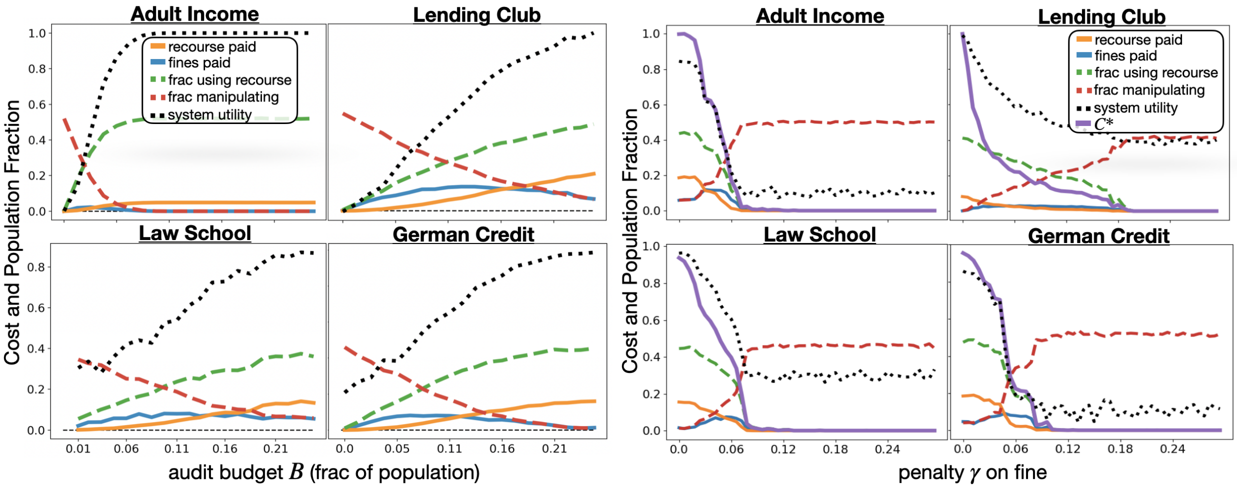

Figure 7.1: Fraction of agents choosing recourse or manipulation (green and red), average cost paid for each action (orange and blue), and system utility (black), for a fixed fine of $C = 1$ (left) or designed fines with audit budget $B = n/10$ (right). ’ To estimate utility the principal uses Logistic Regression (top row) and 2-layer Neural Networks (bottom row).

over approved features, i.e., the conventional recourse setting where $u_a(\mathbf{x}) = 1$ for all $\mathbf{x}$; the principal (system) has utility $u_p(\mathbf{x}) = 1$ when $y = 1$ and $u_p(\mathbf{x}) = -1$ when $y = 0$. In the German Credit and Lending Club datasets, agents have utility which is inversely proportional to their income and savings (credit is more valuable to those with lower existing capital); the principal's utility is equal to the total repayment of approved agents. The cost of recourse is $c_R(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|_2$.

We measure the fraction of the population performing recourse or manipulation, as well as the average cost incurred by agents for either action (Figure 7.1). In this figure three interesting phenomena occur. First, the average fines paid by agents is roughly parabolic in the audit budget B (Figure 7.1 left), and in the penalty γ which controls the size of the fine C^* (Figure 7.1 right). Thus, it is the intermediate values of B and C for which agents are most heavily fined. In these cases, B and C are not large enough to effectively dissuade manipulations, but are large enough to frequently catch and fine agents manipulating. This parabolic relationship is anticipated by Theorem 7.7.3. Second, the maximum cost spent on recourse exceeds the maximum fines paid. This is due to the fact that agents will only select recourse once the cost of manipulation is sufficiently high. Third, as the number of agents choosing recourse increases, so to does system utility. When an agent performs recourse, their true qualification improves (e.g., greater loan repayment), thus increasing the principal's utility when approving that agent.

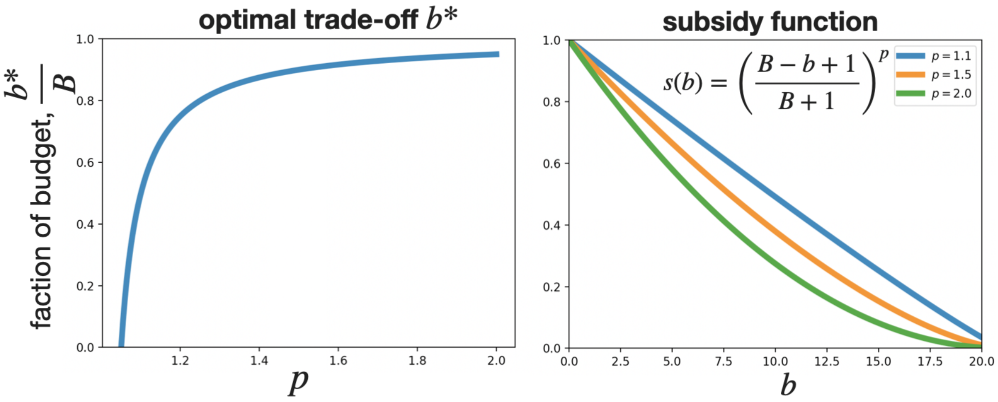

Figure 7.2: For subsidy trade-off function $s(b) = \left(\frac{B-b+1}{B+1}\right)^p$ with $B = 20$, the optimal subsidy trade-off b^* as a function of p (left), and $s(b)$ for different values of p (left). The value of p can be interpreted as how rapidly the cost of recourse decreases with greater allotments of b.

Additionally, in Figure 7.2, we measure the fraction of the audit budget which the principal allocates to subsides for varying subsidy functions. As predicted by Theorem 7.6.2, we observe that when the subsidy function more effectively decrease recourse costs, both the allocation of subsides and the principal's utility increases. Thus, settings in which the cost of recourse is more easily offset give rise to a mutual benefit for both the system and individuals.

7.11 Discussion

We investigated the relationship between manipulation and recourse when the principal possesses the ability to audit agent reports. We demonstrated that auditing can be used as an effective tool in preventing agent manipulation while still allowing the principal to offer recourse and maintain their desired classifier $\hat{y}$. For both a recourse-maximizing and utility-maximizing principal, the optimal audit policy is straightforward to execute, despite the seemingly complex nature of the problem. In particular, given a set of report $\mathbf{X}'$ the principal's best strategy is to uniformly audit all positively classified reports.

Additionally we studied subsides, which allow the principal to allot a portion of their audit budget in order to decrease the cost of recourse. In this case, we find that when agent utility is constant, both objectives of recourse maximization and utility maximization are aligned; however, this is not the case for general agent utilities. Moreover, when agent utility is constant, the principal is guaranteed to spend a nonzero fraction of their audit budget on subsides, so long as the subsidy function $s(b)$ has better than linear scaling in b. We examined this problem from an empirical perspective and found that auditing can successful induce recourse as well as maximize system utility in practice.

Chapter 8

Popularizing Fairness: Group Fairness and Individual Welfare

Previously in Chapter 6 we examined settings in which a principal is considering a change from a conventional, and potentially biased, prediction model to a group-fair approach. We saw that these changes in the prediction model frequently result in widespread decreases to the individual welfare among the population. To capture these individual level harms we proposed the notion of *popularity* which measures the fraction of a population which is made worse off (in terms of utility) when switching from the conventional model to the fair model. While an extensive array of previous works has focused the development of models which have both high performative efficacy and fairness [63, 1, 96, 42, 26, 85, 8, 6, 35], our observations of the individual-level impacts caused by such models necessitates an investigation into whether it is feasible to limit the amount of individual impacts, while maintaining high performance and fairness. To this end, we seek to develop fair-learning schemes which are popular, and have competitive fairness and performance with other stat-of-the-art approaches.

We model the principal's problem as a comparison between a conventional model f_C and a group-fair model f_F, with the principal considering a switch from the former to the latter. Both models select a subset of individuals from a target population to obtain a particular desirable outcome (e.g., a resource, such as admission to a college). We examine popularity in this context through the lens of preferences of individuals in a target population over selection outcomes (which we can encode as positive outcomes of binary classification): an individual *weakly* prefers f_F to f_C if the probability of being selected is not lower under the former than under the latter. Popularity of a group-fair approach f_F then amounts to ensuring that a given fraction (e.g., majority) of a target population prefers f_F to f_C.

Given that group-fair approaches have significant motivation and momentum behind them, instead of designing an entirely new approach to finding popular and fair classifiers, we ask whether it is possible to *minimally postprocess* the output of a group-fair classifier in order to achieve some target popularity while maintaining a high level of fairness. We answer this question in the affirmative. Specifically, we describe two approaches to efficiently postprocess the outputs from a given group-fair classifier in order to boost its popularity. The first approach, called Direct Outcome Shift (DOS), formalizes the problem as a minimal change of outcome probabilities over the target population to guarantee a target level of fairness and popularity. This postprocessing scheme runs in polynomial time. Our second approach, called k-Quantile Lottery Shift (k-QLS), involves a form of regularized empirical risk minimization with fairness and popularity constraints. This approach relies on partitioning prediction scores into a set of quantiles, and we show that, in general, the problem is strongly NP-Hard. However, we also show that if the number of quantiles k is constant, this problem can be solved in polynomial time. Our methods are applicable in both the classification and scarce resource allocation settings, and allow a model designer to directly control the level of popularity and fairness. Moreover, these approaches can be used to postprocesses both deterministic as well as stochastic models.

8.1 Summary of Chapter Results

We introduce two postprocessing algorithms which allow a principal to directly control the popularity of a given fair model, while maintaining good fairness properties. The first postprocessing technique, dubbed Direct Outcome Shift (DOS), is polynomial time solvable for both deterministic and randomized classifiers, and can also be applied to the scarce resource allocation setting. The second technique, k-Quantile Lottery Shift (k-QLS), works by grouping agents into k quantiles (where k is chosen by the model designer), and running lotteries on each quantile. k-QLS is polynomial time solvable for deterministic classifiers. While we show that k-QLS is NP-hard in the randomized case, it becomes tractable for constant k, as would be standard in practice. We empirically demonstrate that the proposed postprocessing techniques can achieve high levels of popularity and fairness with minimal impact on prediction accuracy.

8.2 Additional Related Work

As noted previously, fair learning schemes can be partitioned into three families: preprocessing, inprocessing, and postprocessing. Our proposed algorithms work through postprocessing, and operate in a capacity similar to that of [96, 42, 54, 21, 78, 49]. In these works, the scores or decisions of a conventional classifier are modified in order to achieve fairness. Most post processing techniques for fairness work through "inclusion/exclusion" systems where a potentially randomized procedure is uniformly applied across groups, e.g. random selection of group-specific thresholds [42, 49], or randomly selecting agents from one group to receive positive classification with constant probability [96]. Our postprocessing techniques, while concerned not exclusively with fairness, follow a similar inclusion/exclusion principal.

Our postprocessing approaches also apply to randomized prediction methods which are are common in prior literature. In some cases, randomization is inherently desirable, for example, to explore or correct existing bias in domains such as hiring [11, 105, 44] or lending [59, 60]. In other settings, the aim is to increase model robustness [94, 101], or to achieve better tradeoffs between model performance and fairness, as is common in many group-fair classification approaches [1, 63, 96]. Our model also allows randomness in model decisions to stem from uncertainty as to the exact nature of the model. This uncertainty can stem proprietary or unpublished data [18, 119, 5] or frequent retraining as new data is collected overtime [124, 84, 123].

8.3 Preliminaries

In this chapter we focus the broad family of *additive* fairness metrics, which covers most metrics commonly found across the fair learning literature.

Definition 8.3.1. ***(Additive Efficacy Metric)****: An efficacy metric $\mathcal{M}$ is* additive *if for any population $(\mathcal{X}, Y, G)$,*

$$\mathcal{M}\big(f(\mathcal{X}), Y; g\big) = \sum_{y\in\{0,1\}} \sum_{\substack{i\in G_g: \\ y_i=y}} f(\mathbf{x}_i)c^{(g)}_{y,1} + (1\text{-}f(\mathbf{x}_i))c^{(g)}_{y,0}$$

for some $c_{y,0}^{(g)}, c_{y,1}^{(g)} \in [0,1]$*. In the case of scarce resources* $f(\mathbf{x}_i)$ *is interchangeable with* $I_i(\mathcal{X}, h, k)$*. In the case of randomized models,* $f(\mathbf{x}_i)$ *is replaced with* $\mathbb{E}[f(\mathbf{x}_i)]$ *or* $\mathbb{E}[I_i(\mathcal{X}, h, k)]$*.*

In an additive efficacy metric, the coefficients $c_{y,0}^{(g)}, c_{y,1}^{(g)}$ give the respective "costs" of classifying an example from group G_g, with true label y, as negative or positive, respectively. Thus, unfairness $\mathcal{U}$ is given as the difference in the total efficacy cost between groups. Additive metrics are widely studied in the literature and include metrics such as error rate (ER), positive (or selection) rate (PR), false positive rate (FPR), and true positive rate (TPR). As an example, in the case of PR fairness $c_{y,1}^{(g)} = {}^{1}/_{|G_g|}$ and $c_{y,0}^{(g)} = 0$ for each $y, g \in \{0,1\}$.

In this chapter, we also increase the scope of the *conventional* classifier, in that it is no longer required to be an accuracy maximizing classifier (as was the case when investigating fairness reversals). Instead, the conventional model f_C can be an arbitrary model resulting from the minimizing of a given loss function $\mathcal{L}_C$, that is

$$f_C \in \arg\min_{f \in \mathcal{H}_C} \ \mathcal{L}_C\big(f, \mathcal{X}, Y\big)$$

8.4 Popularity and Fairness Reversals

Prior to our main results, we first outline the connections between individual welfare and classifier selectivity (recall that classifier selectivity was the driving force behind the fairness-reversal phenomenon investigated in Part II). In particular we demonstrate a correspondence between popularity and classifier selectivity, namely that classifiers with higher high popularity are guaranteed to be less selective. Thus, the postprcessing techniques (discussed in detail later) have the added benefit of reducing both the likelihood and severity of fairness reversals.

We begin with the hard-selectivity, and then extend our observations to soft-selectivity. Recall that for hard-selectivity, a classifier f_F is said to be more selective than its conventional counterpart f_C if

$$\big\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\big\} \subset \big\{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\big\},$$

that is, all examples $\mathbf{x}$ which are positively classified by f_F are also positive classified by f_C, and there exists at least one example which positively classified by f_C, but not by f_F.

Theorem 8.4.1. *For deterministic classifiers f_C and f_F, postprocessing with $\gamma = 1$ ensures that f_F is* not *more selective than f_C.*

Proof. In the case of deterministic classification, γ-popularity of f_F is defined as

$$\frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{I}\big[f_C(\mathbf{x}) \leq f_F(\mathbf{x})\big] \geq \gamma.$$

Setting $\gamma = 1$, implies that $f_C(\mathbf{x}) \leq f_F(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{X}$. Thus, any $\mathbf{x}$ with $f_C(\mathbf{x}) = 1$ will also have $f_F(\mathbf{x}) = 1$, implying that $\big\{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\big\} \subseteq \big\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\big\}$ and f_F is not more selective than f_C. □

Recall that in the less restrictive case selectivity is defined as

$$S(f_C, f_F) = \mathbb{P}\big(f_C(\mathbf{x}) = 1 \neq f_F(\mathbf{x})\big) - \mathbb{P}\big(f_C(\mathbf{x}) = 0 \neq f_F(\mathbf{x})\big)$$

The selectivity of f_F can be bounded in terms of the popularity of f_F.

Theorem 8.4.2. *For deterministic classifiers f_C and f_F, if f_F is γ-popular, then f_F is no more than $(1 - \gamma)$ selective, i.e. $S(f_C, f_F) \leq 1 - \gamma$.*

Proof. If f_F is γ-popular, then $\mathbb{P}\big(f_C(\mathbf{x}) < f_F(\mathbf{x})\big) \leq 1 - \gamma$. Thus

$$S(f_C, f_F) \leq 1 - \gamma - \mathbb{P}\big(f_C(\mathbf{x}) = 0 \neq f_F(\mathbf{x})\big)$$

In the worst case, $\mathbb{P}\big(f_C(\mathbf{x}) = 0 \neq f_F(\mathbf{x})\big)$ can be 0 which yields $S(f_C, f_F) \leq 1 - \gamma$. □

In addition to demonstrating the use of our postprocessing methods to prevent fairness reversals, Theorems 8.4.1 and 8.4.2 imply a fundamental connection between fairness reversals and individual welfare. Classifiers which result in a decrease to individual welfare are self-evidently undesirable, especially when they can be easily avoided, however, the fact that such classifiers will also result in fairness reversals when agents are strategic is equally undesirable, but less obvious a-priori.

	Deterministic Models	Randomized Models	Requires Labels	Scarce Resource Allocation
DOS	✓	✓	–	✓
k-QLS	✓	✓	✓	–

Table 8.1: Applicability of each method. Checkmarks indicate situations in which the method is applicable.

8.5 Improving Popularity through Postprocessing

We consider two approaches to minimally postprocess a β-fair scheme f_F such that the resulting decisions also become γ-popular, for exogenously specified β and γ: 1) *direct outcome shift (DOS)* and 2) *k-quantile lottery shift (k-QLS)*. Postprocessing is performed in a transductive setting, in which the populations' features $(\mathcal{X}, G)$ (and possibly also labels Y) are known in advance. Throughout, we use f_P to refer to either approach we propose that combines both popularity and group fairness. Prior to discussing both approaches in detail, we first outline the settings in which each technique is applicable.

Direct Outcome Shift (DOS) DOS-based postprocessing arises from solving the problem of finding a minimal perturbation to the agents' outcomes that achieves both fairness and popularity, e.g. Program 8.15 for randomized classification. For a target population with feature vectors $\mathcal{X}$, we shift individuals' outcomes $f_F(\mathcal{X})$ or expected outcomes $\mathbb{E}[f_F(\mathcal{X})]$ by a *perturbation* vector $\mathbf{p}$. For deterministic decisions, $\mathbf{p} \in \{-1, 0, 1\}^n$, while for stochastic decisions $\mathbf{p} \in [-1, 1]^n$. The optimization goal in either case is to minimize $\|\mathbf{p}\|_q$ for some ℓ_q-norm ($q \in \{1, 2, \infty\}$) such that the final decisions, whether they involve predictions ($f_F(\mathcal{X}) + \mathbf{p}$, or $\mathbb{E}[f_F(\mathcal{X})] + \mathbf{p}$) or allocations ($I(\mathcal{X}, h, k) + \mathbf{p}$, or $\mathbb{E}\big[I(\mathcal{X}, h, k)\big] + \mathbf{p}$) are both β-fair and γ-popular. Since DOS does not use knowledge of true labels Y, it can be applied directly at prediction time to a population of individuals. However, this also means that it can only be applied when the measure of fairness is independent of the true labels Y (for example, ensuring equality of positive rates).

***k*-Quantile Lottery Shift (*k*-QLS)** Another option for creating popular and fair classifiers is to directly minimize a loss function regularized by the distance of the fair-and-popular classifier from the fair classifier (distance is measured on predictions at training time), e.g. Program 8.23 for randomized classifiers. k-QLS-based postprocessing achieves this goal by partitioning scores $h_F(\mathcal{X})$ for a population $\mathcal{X}$ into k bins (based on quantiles). The goal is then to compute probabilities $p_\ell^{(g)}$ for each bin ℓ and group g, which minimize empirical risk and change to each agent's outcome, while achieving γ-popularity and β-fairness. This is done at training time. Then at prediction time, we take all agents in group g with scores in bin ℓ and run a lottery, where each agent is classified as 1 with probability $p_\ell^{(g)}$, and 0 otherwise. Since k-QLS is applied on the training dataset, it also allows us to use fairness metrics that depend on labels Y; for this reason k-QLS is not used in allocation, where Y is typically unknown.

k-QLS is motivated by works such as [42, 96, 54, 21, 78] which aim to postprocess a conventional model to achieve β-fairness by running an "inclusion/exclusion" lottery on groups of agents. However, k-QLS differs from these approaches: shifting all outcomes of a group, even in a randomized manner, is too granular to achieve γ-popularity, and thus we shift outcomes within k quantiles.

Remark 8.5.1. *Achieving γ-popularity and β-fairness may be infeasible in general. However, for common efficacy metrics (e.g., PR, FPR, and TPR), doing so is always possible. Both DOS and k-QLS have a feasible solution for any level of γ-popularity and β-fairness, for both randomized and deterministic models.*

8.6 Postprocessing for Popularity and Fairness

When the conventional model f_C, and β-fair model f_F are deterministic, the optimization problems defined for both the DOS approach and the k-QLS approach can be efficiently solved for any $\mathcal{U}$ defined by an additive efficacy metric $\mathcal{M}$. In both cases, since model decisions are binary, post processing amounts to finding some set of agents negatively classified by f_C, which minimally impact loss while not violating fairness, when positively classified. We first investigate this in the case of DOS and then later in case of k-QLS; both postprocessing paradigms can be solved in polynomial time.

8.6.1 DOS for Deterministic Models

Recall that DOS post processing, given conventional model f_C, β-fair model f_F and population $(\mathcal{X}, G)$, aims to select a vector $\mathbf{p} \in \{-1, 0, 1\}^n$ such that the classifier $f_P(\mathbf{x}_i) = f_F(\mathbf{x}_i) + p_i$

is both γ-popular and β-fair, while minimizing $\|\mathbf{p}\|_q$. For deterministic DOS we study $q = 1$ as each $0 \leq q < \infty$ are equivalent, namely in that each yields the Hamming distance between f_P and f_F, and $q = \infty$ is simply an indicator of $f_P \neq f_F$. For deterministic classifiers DOS can be formulated as

$$\min_{\mathbf{p}\in\{-1,0,1\}^n} \ \|\mathbf{p}\|_q \tag{8.1}$$

$$\text{s.t.} \ \ \mathcal{U}\big(f_F + \mathbf{p},\ D\big) \leq \beta \tag{8.2}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big[f_C(\mathbf{x}_i) \leq f_F(\mathbf{x}_i) + p_i\big] \geq \gamma \tag{8.3}$$

$$0 \leq f_F(\mathbf{x}_i) + p_i \leq 1 \tag{8.4}$$

Objective 8.1 can be solved by Algorithm 1.

Since decisions are binary, DOS is effectively selecting some minimum number of decisions from $f_F(\mathcal{X})$ to flip. In the deterministic case, DOS postprocessing is not technically difficult, but is illustrative of some key ideas used in other, more complex, cases. Specifically, the selection of which agents to flip decisions for is made straightforward by two observations. First, popularity increases only when flipping the decisions of agents with $f_F(\mathbf{x}) = 0$ and $f_C(\mathbf{x}) = 1$. Second, agents within a group are exchangeable with respect to fairness in the sense that for $i, j \in G_g$, either setting of $f_P(\mathbf{x}_i) = 1 - f_P(\mathbf{x}_j)$ results in identical fairness since $\mathcal{U}$ is derived from an additive metric. Combining these observations implies that no optimal solution can have $p_i = 1$ and $p_j = -1$ for $i, j \in G_g$. Moreover, an optimal solution will only choose to flip agents to negative classification, i.e. $p_i = -1$, if doing so is required to rebalance fairness. Thus, with respect to flipping decisions, agents are equivalent up to group membership, $\mathbb{I}[f_F(\mathbf{x}) < f_C(\mathbf{x})]$, and $\mathbb{I}[f_F(\mathbf{x}) = 1]$; implying DOS reduces to deciding whether to increase or decrease positive classifications on each G_g.

Using these facts, it is straightforward to alternate between groups and either positively classify an agent from $S_1 = \{i : f_F(\mathbf{x}_i) < f_C(\mathbf{x}_i)\}$, or negatively classify an agent from $S_2 = \{i : f_F(\mathbf{x}_i) = 1\}$. When positively classifying two agents from different groups has a cancellation-like affect on unfairness (e.g. PR), DOS will never negatively classify an agent with $f_F(\mathbf{x}) = 1$. In such cases f_P is a Pareto-impairment from f_F with respect to agent preference.

Algorithm 1 (Deterministic DOS) Postprocessing technique, applied directly at prediction time, for converting a deterministic β-fair model f_F into γ-popular β-fair model f_P.

```
inputpopulation: $(\mathcal{X}, G)$, $\beta$-fair model: $f_F$,  conventional model: $f_C$, popularity: $\gamma$ result:Weight vector $\mathbf{p} \in \{0,1\}^n$ s.t. $f_P = f_F + \mathbf{p}$ is $\gamma$-popular and $\beta$-fair
$\mathbf{p} := \mathbf{0}$
/* positively classifying agents from different groups has a cancelling effect with respect to unfairness */
if sign$(c_1^{(1)} - c_0^{(1)})$ = sign$(c_0^{(0)} - c_0^{(0)})$ then
    $S_g := \{i \in G_g : f_F(\mathbf{x}_i) < f_C(\mathbf{x}_i)\}$
    $a :=$ # of agents that prefer $f_F$
    /* less than $\gamma n$ agent prefer $f_F$ or unfairness is violated */
    while $a < \gamma n$ or $\mathcal{U}\big(f_F(\mathcal{X}) + \mathbf{p}, G\big) > \beta$ do
        $i_0, i_1 := S_0[0], S_1[0]$
        /* positively classify the agents resulting in the lowest increase to unfairness */
        $g := \mathbb{I}\big[$setting $\mathbf{p}[i_0] := 1$ increases unfairness less than $\mathbf{p}[i_1] := 1\big]$
        $\mathbf{p}[i_g] := 1$
        $S_g$.delete$(i_g)$
        $a$ += 1
    end while
    return p
    /* positively classifying agents from different groups has a monotonic effect on unfairness */
else if sign$(c_1^{(1)} - c_0^{(1)}) = -$sign$(c_1^{(0)} - c_0^{(0)})$ then
    for $g' \in \{0,1\}$ do
        $S_{g'} := \{i \in G_g : f_F(\mathbf{x}_i) < f_C(\mathbf{x}_i)\}$/* all agents from group $G_{g'}$ who prefer $f_C$ */
        /* all agents in $G_{(1-g')}$ positively classified under $f_F$, sorted by $f_C$ */
        $A_{(1-g')} := \{i \in G_{(1-g')} : f_F(\mathbf{x}_i) == 1\}$    s.t. $f_C(\mathbf{x}_i) > f_C(\mathbf{x}_{i+1})$
        $a :=$ # of agents preferring $f_F$
        /* less than $\gamma n$ agents prefer $f_F$ or unfairness is violated */
        while $a < \gamma n$ or $\mathcal{U}\big(f_F(\mathcal{X}) + \mathbf{p}, G\big) > \beta$ do
            /* if unfairness is violated, attempt to fix it */
            if $\mathcal{U}(f_F(\mathcal{X}) + \mathbf{p}, G) > \beta$ then
                $i, j = S_{g'}[0], A_{(1-g')}[0]$
                if $\mathbf{p}^{(g')}[i] := 1$ decreases unfairness then $\mathbf{p}^{(g')}[i] := 1$ else $\mathbf{p}^{(1-g')}[j] := -1$;
            else
                $\mathbf{p}^{(g')}[i] := 1$ /* if unfairness is not violated, increase the positive rate on $G_{g'}$ */
            end if
            update $a$, (+1 or -1)
        end while
    end for
end ifreturn p
```

Theorem 8.6.1. *Let f_C and f_F be a conventional and β-fair classifier respectively, both of which are deterministic. Let U be derived from an additive efficacy metric $\mathcal{M}$ which is independent of Y (e.g., PR). Then DOS, given by Program 8.1, returns a γ-popular β-fair model f_P and can be solved by Algorithm 1 in time $\Theta(n)$.*

Proof. Let $m = \lceil \gamma n \rceil$, i.e., m is the number of popularity constraints that must be satisfied. Each such constraint involves a single variable $p_i \in \{-1, 0, 1\}$ and thus is independent from any other popularity constraint. Moreover since unfairness $\mathcal{U}$ is additive, it can be expressed as

$$\begin{aligned}
&\mathcal{U}\big(f_F(\mathcal{X}) + \mathbf{p}, G\big) = \big|\mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p} : g = 1\big) - \mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p} : g = 0\big)\big| \\
=&| \sum_{i \in G_1} c_1^{(1)}\big(f_F(\mathbf{x}_i) + p_i\big) + c_0^{(1)}\big(1 - \big(f_F(\mathbf{x}_i) + p_i\big)\big) \\
&\qquad - \sum_{j \in G_0} c_1^{(0)}\big(f_F(\mathbf{x}_j) + p_j\big) + c_0^{(0)}\big(1 - \big(f_F(\mathbf{x}_j) + p_j\big)\big)| \\
=&\Bigg|\Bigg(\sum_{i \in G_1} (c_1^{(1)} - c_0^{(1)})p_i - \sum_{j \in G_0} (c_1^{(0)} - c_0^{(0)})p_j\Bigg) \\
&\qquad + \Bigg(\sum_{i \in G_1} c_1^{(1)} f_F(\mathbf{x}_i) + c_0^{(1)}\big(1 - f(\mathbf{x}_i)\big) - \sum_{j \in G_0} c_1^{(0)} f_F(\mathbf{x}_j) - c_0^{(0)}\big(1 - f_F(\mathbf{x}_j)\big)\Bigg)\Bigg|
\end{aligned}$$

for scalars $c_1^{(g)}, c_0^{(g)}$ which give the respective cost of positively or negatively classifying an agent from group G_g. Note that

$$u := \sum_{i \in G_1} c_1^{(1)} f_F(\mathbf{x}_i) + c_0^{(1)}\big(1 - f(\mathbf{x}_i)\big) - \sum_{j \in G_0} c_1^{(0)} f_F(\mathbf{x}_j) - c_0^{(0)}\big(1 - f_F(\mathbf{x}_j)\big)$$

is constant for fixed f_F and $(\mathcal{X}, Y, G)$. Then the fairness constraint can be expressed as

$$\begin{aligned}
&\mathcal{U}\big(f_F(\mathcal{X}) + \mathbf{p}, G\big) \leq \beta \\
\iff& -\beta - u \leq \sum_{i \in G_1} (c_1^{(1)} - c_0^{(1)})p_i - \sum_{j \in G_0} (c_1^{(0)} - c_0^{(0)})p_j \leq \beta - u \qquad (8.5)
\end{aligned}$$

Due to the additive nature of this fairness term, each member of group g is exchangeable, meaning that for any two agents $i_1, i_2 \in G_g$, fairness is invariant under any alteration to p_{i_1}, p_{i_2} which preserves the value of $p_{i_1} + p_{i_2}$. More specifically, for any $i_1, i_2 \in G_g$ and any feasible solution $\mathbf{p}$ with $p_{i_1} = -1$ and $p_{i_2} = 1$, let $\mathbf{p}'$ be defined by $p'_k = p_k$ for all $k \neq i_1, i_2$

and $p'_{i_1} = p'_{i_2} = 0$. Then $\mathbf{p}'$ is both a feasible solution and has $\|\mathbf{p}'\| \leq \|\mathbf{p}\|$. The latter part of which is straightforward; to see the former we need only consider the popularity constraints since fairness is satisfied by the feasibility of $\mathbf{p}$ and $p_{i_1} + p_{i_2} = p'_{i_1} + p'_{i_2} = 0$. Although it may be the case that $f_C(\mathbf{x}_i) > f_F(\mathbf{x}_i) + p'_i = f_F(\mathbf{x}_i) + p_i - 1$, i.e., agent i no longer prefers f_F, it must be the case that $f_C(\mathbf{x}_j) \leq 1 \leq f_F(\mathbf{x}_i) + p_j + 1 = f_F(\mathbf{x}_i) + p'_j$, i.e., agent j prefers f_F.

Since agents from the same group are exchangeable and no optimal solution has both $p_{i_1} = -1$ and $p_{i_2} = 1$ for $i_1, i_2 \in G_g$, the optimal score shift $\mathbf{p}$ can be found by alternating between groups and greedily assigning either $p_i = 1$ or $p_i = -1$, as outlined by Algorithm 1. To see the optimality of this greedy selection procedure, let

$$U(f_F(\mathcal{X}) + \mathbf{p}, G) = \mathcal{M}(f_F(\mathcal{X}) + \mathbf{p} : g = 1) - \mathcal{M}(f_F(\mathcal{X}) + \mathbf{p} : g = 1),$$

i.e. the function U is equivalent to $\mathcal{U}$ without absolute value. With respect to greedy selection, only two cases need be considered: 1.) $\text{sign}(c_1^{(1)} - c_0^{(1)}) = \text{sign}(c_1^{(0)} - c_0^{(0)})$ and 2.) $\text{sign}(c_1^{(1)} - c_0^{(1)}) = -\text{sign}(c_1^{(0)} - c_0^{(0)})$.

In case (1) if choosing to positively classify an agent from group G_1 increases (decreases) the value of $U(f_F(\mathcal{X}) + \mathbf{p}, G)$ then positively classifying an agent from group G_0 decreases (increases) the value of $U(f_F(\mathcal{X}) + \mathbf{p}, G)$. Thus if increasing the number of positive classifications on G_0, or on G_1, violates unfairness, the only way to resatisfy fairness is to increase the number of positive classifications on the other group. In case (2) if choosing to positively classify an agent from group G_1 increases (decreases) the value of $U(f_F(\mathcal{X}) + \mathbf{p}, G)$ then positively classifying an agent from group G_0 also increases (decreases) the value of $U(f_F(\mathcal{X}) + \mathbf{p}, G)$. Thus if increasing the number of positive classifications on G_0, or on G_1, violates unfairness, the only way to resatisfy fairness is to decrease the number of positive classifications on the other group.

The selection process examines at most n agents, and each decision on an agent takes constant time. Thus DOS can be solved in time $\Theta(n)$ for deterministic classifiers. □

8.6.2 k-QLS for Deterministic Classification

Recall that k-QLS is a postprocessing technique which, given a conventional model f_C, a β-fair model f_F, and a training set $D = (\mathcal{X}, Y, G)$, postprocesses the predictions of f_F such that they are γ-popular and β-fair. This is achieved by running a lottery on k quantiles defined by the scores of f_F, namely $h_F(\mathcal{X})$, the resulting model after postprocessing is refereed to as f_P. Specifically, the scores $h_F(\mathcal{X})$ are partitioned into k intervals in the following manner: let ρ_ℓ be the maximum score associated with quantile ℓ of $h_F(\mathcal{X})$, and let $I_\ell = [\rho_{\ell-1}, \rho_\ell]$ with the understanding that $\rho_{-1} = 0$ and $\rho_k = 1$. The resulting classifier f_P makes predictions $f_P(\mathbf{x}) = p_\ell^{(g')}$ for $h_F(\mathbf{x}) \in I_\ell$ with $g = g'$. In the case of deterministic models, $p_\ell^{(g')} \in \{0, 1\}$. The optimal β-fair γ-popular model can be found by solving:

$$\min_{\mathbf{p}^{(0)},\mathbf{p}^{(1)} \in \{0,1\}^{2k}} \mathcal{L}\big(f_P, D\big) + \lambda \| f_P(\mathcal{X}) - f_F(\mathcal{X}) \|_q^q \tag{8.6}$$

$$\text{s.t.} \quad \mathcal{U}\big(f_P, D\big) \leq \beta \tag{8.7}$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\big[f_C(\mathbf{x}_i) \leq f_P(\mathbf{x}_i, g_i)\big] \geq \gamma \tag{8.8}$$

where $\mathcal{L}$ is balanced accuracy. Unlike DOS, k-QLS does not admit a straightforward solution, but is still polynomial time solvable. The key difference between these two techniques is that k-QLS makes decisions over sets of agents, rather than individual agents, and each interval may contain any number of agents with any combination of true labels and predicted outcomes under both f_C and f_F. Thus much of the symmetry from the DOS case is lost, however enough symmetry remains that a dynamic programming solution can produce the optimal f_P in polynomial time.

Theorem 8.6.2. *Let f_C and f_F be a conventional and β-fair classifier respectively, both of which are deterministic. Let U be derived from an additive efficacy metric $\mathcal{M}$ (e.g. FPR). Then k-QLS, given by Program 8.6, returns a γ-popular and β-fair model f_P time $\Theta\big(\gamma k n^6\big)$ via dynamic programming. Moreover, when $\mathcal{M}$ is given by FPR, TPR, PR, or ERR, f_P can be found in time $\Theta\big(\gamma k n^4\big)$.*

Before proving this theorem we first mention that while k-QLS admits a polynomial time solution, k-QLS can also be transformed into a MILP and solvers such as CPLEX may be more efficient in practice since k, the number of variables, will typically be constant (e.g., breaking scores into 10 intervals) and the program contains only two constraints (one for popularity and one for fairness).

Proof of Theorem 8.6.2. When post processing with k-QLS the model designer creates k intervals based on the quantiles of $h_F(\mathbf{x})$ and aims to shift the scores of agents in each

Algorithm 2 (Deterministic k-QLS) Postprocessing technique, learned at training time and later applied at prediction time, for converting a deterministic β-fair model f_F into γ-popular β-fair model f_P.

input: population: $(\mathcal{X}, G)$, β-fair model: f_F, score function of f_F: h_F, conventional model: f_C, popularity: γ, quantiles k **result:** Weight $\mathbf{p} \in \{0,1\}^{2k}$ of the γ-popular β-fair f_P

```
$\rho_\ell :=$ maximum score in quantile $\ell$ of $h_F(\mathcal{X})$   $\forall \ell \in [k]$ /* partition scores $h_F(\mathcal{X})$ in $k$ intervals based on quantile */
$I_\ell = [\rho_{\ell-1}, \rho_\ell]$   $\forall \ell \in [k]$
$\mathbf{p}^{(0)}, \mathbf{p}^{(1)} := \mathbf{0}$
/* parameters indicating the effects of setting $p_\ell^{(g)} := 1$ */
$N_\ell^{(g)} := \#$ of agents in $G_g$ with $h_f(\mathbf{x}) \in I_\ell$ and $f_C(\mathbf{x}) = 1$
$C_\ell^{(g)} :=$ increase to unfairness (without absolute value)
$L_\ell^{(g)} :=$ increase to loss
/* partition each group and interval according according to effect on fairness */
$S_+ := \{(g, \ell) : 0 \leq C_\ell^{(g)}\}$
$S_- := \{(g, \ell) : C_\ell^{(g)} < 0\}$
/* loss independent on each $I_\ell$, and thus on $S_+$ and $S_-$ */
build a knapsack-like problem over each $S$ using weights $C_\ell^{(g)}$, $N_\ell^{(g)}$, and values $L_\ell^{(g)}$
/* $p_\ell^{(g)}$ corresponds to selecting item $(g, \ell)$ polynomial number of possibilities for each */
$\mathbf{m}_+, \mathbf{m}_- :=$ all possible $\#$ of agents preferring $f_F$ corresponding to solution from $S_+, S_-$

$\mathbf{u}_+, \mathbf{u}_- :=$ all possible values of unfairness corresponding to solution from $S_+, S_-$
for each $m_- \in \mathbf{m}_-$ and each $u_- \in \mathbf{u}_-$ do
    dynamically compute optimal solution from $S_-$ using exactly $m_-$ agents and $u_-$ unfairness
    for each $m_+ \in \mathbf{m}_+$ and $u_+ \in \mathbf{u}_+$ do
        dynamically compute optimal solution from $S_+$ using exactly $m_+$ agents and $u_+$ unfairness.
        if solution from $S_-$ and $S_+$ is feasible then
            save the combined solution
        end if
    end for
end forreturn $\mathbf{p}$ corresponding to solution with the lowest loss
```

interval such that γ-popularity and β-fairness are achieved. Let ρ_ℓ be the maximum score associated with quantile ℓ of $h_F(\mathcal{X})$, and let $I_\ell = [\rho_{\ell-1}, \rho_\ell]$ with the understanding that $\rho_{-1} = 0$.

Thus k-QLS aims to find binary vectors $\mathbf{p}^{(g)} \in \{0,1\}^k$ for each group G_g, such that the model $f_P(\mathbf{x}) = p^{(g)}_\ell$ for $h_F(\mathbf{x}) \in I_\ell$, in group g is γ-popular and β-fair. Since unfairness $\mathcal{U}$ is given in terms of an additive efficacy metric $\mathcal{M}$, the unfairness of f_P over population $D = (\mathcal{X}, Y, G)$ can be expressed as

$$\begin{aligned}
&\mathcal{U}\big(f_P, D\big) = \big|\mathcal{M}\big(f_P(\mathcal{X}, g), Y : g = 1\big) - \mathcal{M}\big(f_P(\mathcal{X}, g), Y : g = 0\big)\big| \\
=&\Bigg|\sum_{\ell=1}^{k} \sum_{y \in \{0,1\}} \Bigg(\sum_{\substack{i \in G_1 \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(1)}_{y,1} p^{(1)}_\ell \big(1 - |y - y_i|\big) + c^{(1)}_{y,0}\big(1 - p^{(1)}_\ell\big)\big(1 - |y - y_i|\big) \\
&\qquad - \sum_{\substack{j \in G_0 \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(0)}_{y,1} p^{(0)}_\ell \big(1 - |y - y_i|\big) + c^{(0)}_{y,0}\big(1 - p^{(0)}_\ell\big)\big(1 - |y - y_i|\big)\Bigg)\Bigg|
\end{aligned}$$

$$\begin{aligned}
=&\Bigg|\sum_{\ell=1}^{k} \Bigg(p^{(1)}_\ell \sum_{y \in \{0,1\}} (c^{(1)}_{y,1} - c^{(1)}_{y,0}) \sum_{\substack{i \in G_1 \\ h_F(\mathbf{x}_i) \in I_\ell}} \big(1 - |y - y_i|\big) \\
&\quad - p^{(0)}_\ell \sum_{y \in \{0,1\}} (c^{(0)}_{y,1} - c^{(0)}_{y,0}) \sum_{\substack{i \in G_0 \\ h_F(\mathbf{x}_i) \in I_\ell}} \big(1 - |y - y_i|\big) \\
&\quad + \sum_{y \in \{0,1\}} \sum_{\substack{i \in G_1: \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(1)}_{y,0}\big(1 - |y - y_i|\big) - \sum_{y \in \{0,1\}} \sum_{\substack{i \in G_0: \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(0)}_{y,0}\big(1 - |y - y_i|\big)\Bigg)\Bigg|
\end{aligned}$$

for scalar costs $c^{(g)}_{y,1}, c^{(g)}_{y,0}$. Note that

$$u := \sum_{\ell=1}^{k} \sum_{y \in \{0,1\}} \sum_{\substack{i \in G_1: \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(1)}_{y,0}\big(1 - |y - y_i|\big) - \sum_{y \in \{0,1\}} \sum_{\substack{i \in G_0: \\ h_F(\mathbf{x}_i) \in I_\ell}} c^{(0)}_{y,0}\big(1 - |y - y_i|\big)$$

and each

$$C^{(g)}_\ell := (1 - 2g) \sum_{y \in \{0,1\}} (c^{(g)}_{y,1} - c^{(g)}_{y,0}) \sum_{\substack{i \in G_g \\ h_F(\mathbf{x}_i) \in I_\ell}} \big(1 - |y - y_i|\big)$$

are constants. Thus the fairness constraint on f_P can be expressed as

$$\begin{aligned}
&\mathcal{U}\big(f_P, D) \leq \beta \\
\iff\ -\beta - u \leq &\sum_{\ell=1}^{k} \big(2(1)-1\big)p_\ell^{(1)}C_\ell^{(1)} - \big(2(0)-1\big)p_\ell^{(0)}C_\ell^{(0)} \leq \beta - u \qquad (8.9) \\
\iff\ -\beta - u \leq &\sum_{\ell=1}^{k} p_\ell^{(1)}C_\ell^{(1)} + p_\ell^{(0)}C_\ell^{(0)} \leq \beta - u \qquad (8.10)
\end{aligned}$$

Thus, unfairness of f_P is given by a linear constraint on the vectors $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(0)}$.

Similar to the unfairness term, the optimization objective

$$\begin{aligned}
&\mathcal{L}(f_P, \mathcal{X}, Y, G) + \lambda \| f_F(\mathcal{X}) - f_P(\mathcal{X}) \|_q^q \\
=&\sum_{\ell}^{k} \sum_{g \in \{0,1\}} \Bigg(\sum_{\substack{i \in G_g \\ h_F(\mathbf{x}_i) \in I_\ell}} (1 - y_i)p_\ell^{(g)} + y_i(1 - p_\ell^{(g)}) \Bigg) + \sum_{g \in \{0,1\}} \sum_{\substack{i \in G_g \\ h_F(\mathbf{x}_i) \in I_\ell}} \lambda \big| f_F(\mathbf{x}_i) - p_\ell^{(g)} \big|^q
\end{aligned}$$

can, by shifting and rescaling, be equivalently expressed as

$$\sum_{\ell}^{k} \sum_{g \in \{0,1\}} p_\ell^{(g)} \Bigg(\sum_{\substack{i \in G_g \\ h_F(\mathbf{x}_i) \in I_\ell}} (1 - y_i - \lambda f_F(\mathbf{x}_i)) \Bigg) \qquad (8.11)$$

due to the fact that f_F, y, and $\mathbf{p}$ are binary; each term

$$L_\ell^{(g)} := \sum_{\substack{i \in G_g \\ h_F(\mathbf{x}_i) \in I_\ell}} (1 - y_i - \lambda f_F(\mathbf{x}_i))$$

is constant. Lastly, let

$$N_\ell^{(g)} = \big| \big\{ i \in G_g : f_C(\mathbf{x}_i) = 0 \ \text{ and } \ h_F(\mathbf{x}_i) \in I_\ell \big\} \big|$$

Thus the optimization of k-QLS can be is equivalently formulated as,

$$\min_{\mathbf{p}^{(0)},\mathbf{p}^{(1)}\in\{0,1\}^k} \sum_{\ell=1}^{k}\sum_{g\in\{0,1\}} p_\ell^{(g)} L_\ell^{(g)} \tag{8.12}$$

$$\text{s.t.} \quad -\beta - u \leq \sum_{\ell=1}^{k} p_\ell^{(1)} C_\ell^{(1)} + p_\ell^{(0)} C_\ell^{(0)} \leq \beta - u \tag{8.13}$$

$$\sum_{\ell=1}^{k}\sum_{g\in\{0,1\}} p_\ell^{(g)} N_\ell^{(g)} \leq \big\lfloor (1-\gamma)\big(1-\mathrm{PR}(f_C)\big)n \big\rfloor \tag{8.14}$$

The popularity term $\sum_{\ell=1}^{k}\sum_{g\in\{0,1\}} p_\ell^{(g)} N_\ell^{(g)}$ can take on at most $\lceil n(1-\mathrm{PR}(f_C))\rceil$ different values. In the fairness constraint, each term $\sum_{\ell=1}^{k} p_\ell^{(g)} C_\ell^{(g)}$ can take on at most ${}^1\!/\!{}_2|G_g|\big(|G_g|+1\big)$ unique values since each can be written as

$$\sum_{\ell=1}^{k} p_\ell^{(g)} C_\ell^{(g)} = \sum_{y\in\{0,1\}} a_y(c_{y,1}^{(g)} - c_{y,0}^{(g)}) \quad \text{for some } \; a_0, a_1 \in \mathbb{N} \;\text{ with }\; a_0 + a_1 \leq |G_g|$$

Next we create two index sets which keep track of which groups g and intervals ℓ have positive and negative coefficients $C_\ell^{(g)}$. Let $S_+ = \{(g,\ell) : C_\ell^{(g)} \geq 0\}$ and $S_- = \{(g,\ell) : C_\ell^{(g)} \geq 1\}$. Thus S_+ and S_- indicate whether $p_\ell^{(g)}$ will increase or decrease the value of Equation 8.13. Specifically, suppose that for each $(g,\ell) \in S_-$, $r_\ell^{(g)}$ is a solution. Let $R = \sum_{(g,\ell)\in S_-} r_\ell(g) p_\ell^{(g)}$, and $\theta = \big\lfloor (1-\gamma)\big(1-\mathrm{PR}(f_C)\big)n \big\rfloor - \sum_{(g,\ell)\in S_-} r_\ell^{(g)} N_\ell^{(g)}$. Then the problem reduces to solving

$$\begin{aligned} \min \sum_{g,\ell\in S_+} & p_\ell^{(g)} L_\ell^{(g)} \\ & \text{s.t.} \; -\beta - u - R \leq \sum_{g,\ell\in S_+} p_\ell^{(g)} C_\ell^{(g)} \leq \beta - u - R \\ & \sum_{g,\ell} p_\ell^{(g)} N_\ell^{(g)} \leq \theta \end{aligned}$$

which yields a knapsack problem with two constraints, with weights $C_\ell^{(g)}$ and $N_\ell^{(g)}$. Since there are at most k decision variables, $\sum_{g,\ell} p_\ell^{(g)} N_\ell^{(g)}$ can take on at most γn unique feasible values, and $\sum_{g,\ell\in S_+} p_\ell^{(g)} C_\ell^{(g)}$ can take on at most n^2 unique values. This problem is therefore solvable in $\Theta(k\gamma n^3)$ time. Moreover, since any solution set generated from S_- can produce at most n^2 values of R and n values of θ, any configuration of variables from S_- produce γn^3 unique subproblems, each of which can be solved in time $\Theta(k\gamma n^3)$. Thus, Algorithm 2 solves

k-QLS in time $\Theta\big(\gamma k n^6\big)$ for general additive metrics. Moreover for PR, TPR, FPR, and ER, each $\sum_{g,\ell\in S_+} p_\ell^{(g)} C_\ell^{(g)}$ can take on at most n unique values (rather than n^2), implying there are only n^2 unique subproblems, each requiring $\Theta(k\gamma n^2)$ time to solve, thus k-QLS is solvable in time $\Theta\big(\gamma k n^4\big)$. □

8.6.3 DOS for Randomized Models

Next we investigate popularity as it relates to randomized classifiers. Recall that in the case of randomized classifiers DOS aims to minimally shift the expected outcomes of f_F on a population $(\mathcal{X}, G)$, with unknown true labels Y, to produce the γ-popular β-fair model, which we denote by f_P, where $\mathbb{E}\big[f_P(\mathbf{x}_i)\big] = \mathbb{E}\big[f_F(\mathbf{x}_i)\big] + p_i$, and $0 \le \mathbb{E}\big[f_P(\mathbf{x}_i)\big] \le 1$. Thus, DOS aims to solve the following optimization problem:

$$\min_{\mathbf{p}\in[-1,1]^n} \;\|\mathbf{p}\|_q \tag{8.15}$$

$$\text{s.t.}\;\; \mathcal{U}\Big(\mathbb{E}\big[f_F(\mathcal{X})\big] + \mathbf{p},\; G\Big) \le \beta \tag{8.16}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\Big[\mathbb{E}\big[f_C(\mathbf{x}_i)\big] \le \mathbb{E}\big[f_F(\mathbf{x}_i)\big] + p_i\Big] \ge \gamma \tag{8.17}$$

for $q \in \{1, 2, \infty\}$. A key challenge is that the popularity constraint (8.17) is discrete and non-convex, amounting to a combinatorial problem of identifying a subset of $\gamma|\mathcal{X}|$ individuals who prefer the f_P to its conventional counterpart f_C. Nevertheless, this problem can be solved in polynomial time.

Theorem 8.6.3. *Let f_C and f_F be respectively a conventional and β-fair randomized classifier. Let $\mathcal{U}$ be derived from an additive efficacy metric $\mathcal{M}$ which is independent of Y (e.g., PR). Then for $q \in \{1, 2, \infty\}$ Program 8.15 can be solved in time $\Theta(\gamma n T)$ (where $\Theta(T)$ is the time required to solve a linear program or semi-definite program, as appropriate) by Algorithm 3, which returns a γ-popular, β-fair model f_P.*

Proof. To prove this claim, we make use of the fact that fairness is given by an additive metric. That is, fairness (or rather unfairness) is linear with respect to perturbations in agents' scores. Moreover any two agents with the same group and label, are exchangeable, i.e., swapping the scores of the two agents has no effect on fairness. Using this as our key intuition prove this theorem in two parts; first, that Algorithm 3 produces an optimal solution assuming that each Program P_i in this algorithm can be solved optimally, and second, that Algorithm 4 does in fact provide an optimal solution to P_i.

(Optimality of Algorithm 3): Recall that $\mathbb{E}\big[f_F(\mathbf{x})\big] = h_F(\mathbf{x})$ and $\mathbb{E}\big[f_C(\mathbf{x})\big] = h_C(\mathbf{x})$, i.e., the expected outcome of each classifier is given by is respective score function. For notational convince we use we use h_F and h_C throughout this proof. Program 8.15 is non-convex with respect to $\mathbf{p}$ due to the constraint that γ-fraction of the population needs to prefer f_F over f_C, namely that $m = \gamma n$ of the constraints

$$h_C(\mathbf{x}_i) \leq h_F(\mathbf{x}_i) + p_i$$

need to be satisfied. However, note that if instead of needing to satisfy *any* m constraints, we needed to satisfy a specific set of m constraints, say

$$S = \big\{h_C(\mathbf{x}_{i_1}) \leq h_F(\mathbf{x}_{i_1}) + p_{i_1}, \ldots, h_C(\mathbf{x}_{i_m}) \leq h_F(\mathbf{x}_{i_m}) + p_{i_m}\big\},$$

then the resulting program would be trivial to solve as it amounts to ℓ_q-norm minimization subject to linear constraints. Thus, if the optimal set of popularity constraints can be found efficiently, the problem is polynomial time solvable.

Algorithm 3 (Randomized DOS) Postprocessing technique for converting a β-fair model f_F into a γ-popular β-fair model f_P.

```
Input: population: $(\mathcal{X}, Y, G)$, $\beta$-fair model: $f_F$, conventional model: $f_C$, popularity: $\gamma$
Result: weights $\mathbf{p}$ s.t. $f_P = f_F + \mathbf{p}$ is $\gamma$-popular and $\beta$-fair
$G_g := \big\{i : g_i = g\big\}$
/* each group is sorted by their loss in utility when switching to $f_F$ */
Sort each $G_g$ s.t. $\mathbb{E}\big[f_C(\mathbf{x}_i)\big] - \mathbb{E}\big[f_F(\mathbf{x}_i)\big] \leq \mathbb{E}\big[f_C(\mathbf{x}_{i+1})\big] - \mathbb{E}\big[f_F(\mathbf{x}_{i+1})\big]$
$m := \lceil \gamma n \rceil$
for $i = 1$ to $m$ do
    /* $m$ popularity constraints from each group */
    $S_i = \big\{\mathbb{E}\big[f_C(\mathbf{x}_j)\big] \leq \mathbb{E}\big[f_F(\mathbf{x}_j)\big] + p_j : j \in G_1[: i]\big\}$
        $\cup \big\{\mathbb{E}\big[f_C(\mathbf{x}_j)\big] \leq \mathbb{E}\big[f_F(\mathbf{x}_j)\big] + p_j : j \in G_0[: m - i]\big\}$
    /* add popularity constraints $S_i$ for the $m$ “easiest” agents between groups */
    $P_i$ = Program 8.15 with Constraint 8.17 replaced by $S_i$;
    solve the program via Algorithm 4 or off-the-shelf solvers;
    $\mathbf{p}_i$ = solution to the modified program
end for
return $\mathbf{p}^* = \arg\min_i \|\mathbf{p}_i\|$
```

Algorithm 4 Algorithm to solve programs associated with DOS in the randomized classification setting, when given S, a specific set of γn that must prefer f_P to f_C.

```
input: population: $(\mathcal{X}, Y, G)$, $\beta$-fair model: $f_F$, conventional model: $f_C$, $\gamma n : S$
result: Weight vector $\mathbf{p}$ s.t. $f_P = f_F + \mathbf{p}$
p = 0
$m := \gamma n$
$\boldsymbol{\delta} := h_F(\mathcal{X})$ /* lower bound on perturbation to agents' scores */
/* min score increase and $p_i$ for $i$ to prefer $f_P$ */
$(p_i, \delta_i) := \max\big(0, h_C(\mathbf{x}_i) - h_F(\mathbf{x}_i)\big), \quad h_C(\mathbf{x}_i) - h_F(\mathbf{x}_i);$
/* check "direction" in which fairness is violated */
if $\big|\mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p}, Y; g = 1\big) - \mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p}, Y; g = 0\big) < -\beta$ then
    /* $s_g$=1 ($s_g$=-1) indicates increasing (decreasing) scores on $G_g$ */
    $s_g := \mathrm{sign}\big((1-2g)(c_1^{(g)} - c_0^{(g)})\big)$ for $g \in \{0,1\}$
else if $\big|\mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p}, Y; g = 1\big) - \mathcal{M}\big(f_F(\mathcal{X}) + \mathbf{p}, Y; g = 0\big) > \beta$ then
    $s_g := \mathrm{sign}\big((1-2g)(c_0^{(g)} - c_1^{(g)})\big)$ for $g \in \{0,1\}$
end if
while $\mathcal{U}\big(f_F + \mathbf{p}, \mathcal{D}\big) > \beta$ do
    /* agents from each group whose scores can still be perturbed */
    if $s_g = 1$ then
        $G_g := \{i \in G_g : h_F(\mathbf{x}_i) + p_i < 1\}$
        /* maximum additional perturbation to $G_g$ which is feasible */
        $\varepsilon_g := \min\Big(\big\{1 - h_F(\mathbf{x}_i) : i \in G_g\big\} \cup \big\{\frac{\beta - \mathcal{U}(f_F + \mathbf{p}, \mathcal{D})}{(c_1^{(g)} - c_0^{(g)})|G_g|}\big\}\Big)$
    else
        $G_g := \{i \in G_g : \delta_i < h_F(\mathbf{x}_i) + p_i\}$
        $\varepsilon_g := \min\Big(\big\{\delta_i : i \in G_g\big\} \cup \big\{\frac{\beta - \mathcal{U}(f_F + \mathbf{p}, \mathcal{D})}{(c_0^{(g)} - c_1^{(g)})|G_g|}\big\}\Big)$
    end if
    /* check if increasing scores by $\varepsilon_0, \varepsilon_1$ would fix fairness */
    $\mathbf{p}' := \mathbf{p}$
    $\mathbf{p}'[G_g] += \varepsilon_g$ for $g \in \{0,1\}$
    if $\mathcal{U}\big(f_F + \mathbf{p}', \mathcal{D}\big) < \beta$ then
        /* in the case that fairness is achieved, $\varepsilon_g$ may be too large */
        decrease $\varepsilon_g$ s.t. $\mathcal{U}\big(f_F + \mathbf{p}', \mathcal{D}\big) = \beta$ and $\sum_{i \in G_0} \big|p_i + \varepsilon_0\big|^q = \sum_{i \in G_1} \big|p_i + \varepsilon_1\big|^q$
        return $\mathbf{p}'$
    end if
    $\mathbf{p}_g := \mathbf{p}[G_g] + \varepsilon_g \quad$ for $g \in \{0,1\}$
    /* ratio of "fairness repair" to increase in loss */
    $g^* := \arg\min_{g \in \{0,1\}} \Big(\frac{\varepsilon_g |G_g| s_g (c_1^{(g)} - c_0^{(g)})}{\|\mathbf{p}_g\|_q - \|\mathbf{p}\|_q}\Big)$ /* where $\frac{a}{0} := \infty$ for $a \neq 0$ and $\frac{0}{0} := 0$ */
    $\mathbf{p} := \mathbf{p}_{g^*}$
end while
return p
```

To find this set of constraints, we make use of the fact that the metric $\mathcal{M}$ defining $\mathcal{U}$ is additive, specifically the fact that $\mathcal{U}$ can be expressed in terms of scalars $c_1^{(g)}, c_0^{(g)} \in [0,1]$ which give the respective cost of positively or negatively classifying an agent from group G_g. That is, given a perturbation $\mathbf{p} \in [-1,1]^n$ and fair model f_F, unfairness can be written as,

$$\begin{aligned}
&\mathcal{U}\big(h_F(\mathcal{X}) + \mathbf{p}, G\big) \\
=&\big|\mathcal{M}\big(h_F(\mathcal{X}) + \mathbf{p} : g = 1\big) - \mathcal{M}\big(h_F(\mathcal{X}) + \mathbf{p} : g = 0\big)\big| \\
=&\Bigg| \sum_{i \in G_1} c_1^{(1)}\big(h_F(\mathbf{x}_i) + p_i\big) + c_0^{(1)}\big(1 - \big(h_F(\mathbf{x}_i) + p_i\big)\big) \\
&\qquad - \sum_{j \in G_0} c_1^{(0)}\big(h_F(\mathbf{x}_j) + p_j\big) + c_0^{(0)}\big(1 - \big(h_F(\mathbf{x}_j) + p_j\big)\big)\Bigg| \\
=&\Bigg|\bigg(\sum_{i \in G_1}(c_1^{(1)} - c_0^{(1)})p_i - \sum_{j \in G_0}(c_1^{(0)} - c_0^{(0)})p_j\bigg) \\
&\quad + \bigg(\sum_{i \in G_1} c_1^{(1)} h_F(\mathbf{x}_i) + c_0^{(1)}\big(1 - h_F(\mathbf{x}_i)\big) - \sum_{j \in G_0} c_1^{(0)} h_F(\mathbf{x}_j) - c_0^{(0)}\big(1 - h_F(\mathbf{x}_j)\big)\bigg)\Bigg|
\end{aligned}$$

Since $c_0^{(g)}$, $c_1^{(g)}$, and $h_F(\mathcal{X})$ are constant

$$u := \sum_{i \in G_1} c_1^{(1)} h_F(\mathbf{x}_i) + c_0^{(1)}\big(1 - h_F(\mathbf{x}_i)\big) - \sum_{j \in G_0} c_1^{(0)} h_F(\mathbf{x}_j) - c_0^{(0)}\big(1 - h_F(\mathbf{x}_j)\big)$$

is also constant. Thus the fairness constraint can be expressed as

$$\begin{aligned}
&\mathcal{U}\big(h_F(\mathcal{X}) + \mathbf{p}, G\big) \leq \beta \\
\iff& -\beta - u \leq \sum_{i \in G_1}(c_1^{(1)} - c_0^{(1)})p_i - \sum_{j \in G_0}(c_1^{(0)} - c_0^{(0)})p_j \leq \beta - u.
\end{aligned} \tag{8.18}$$

This formulation of unfairness is similar to the case of deterministic DOS with the key difference being the domain of the optimization variables $\mathbf{p}$. From this inequality, we see that for any two agents i_1, i_2 from the same group, increasing or decreasing the score of i_1 has the same effect on unfairness as equivalently increasing or decreasing the score of i_2. More specifically, let $i_1, i_2 \in G_g$, then for any potential solution $\mathbf{p}$, let $\mathbf{p}'$ be any other potential solution with $p'_j = p'_j$ if $j \neq i_1, i_2$, and $p_{i_1} + p_{i_2} = p'_{i_1} + p'_{i_2}$. Both $\mathbf{p}$ and $\mathbf{p}'$ have equivalent fairness. This observation can be used to order both groups in terms of increase in p_i required for agent i to prefer f_P over f_C.

To induce this ordering, consider any two agents $i_1, i_2 \in G_g$ with

$$h_C(\mathbf{x}_{i_1}) - h_F(\mathbf{x}_{i_1}) \leq h_C(\mathbf{x}_{i_2}) - h_F(\mathbf{x}_{i_2}),$$

that is, i_1 requires at least as large a score shift as i_2 to prefer f_F over f_C. Let S_1 be any set of m popularity constraints which include $h_C(\mathbf{x}_{i_1}) \leq h_F(\mathbf{x}_{i_1}) + p_{i_1}$, but not $h_C(\mathbf{x}_{i_2}) \leq h_F(\mathbf{x}_{i_2}) + p_{i_2}$, and let

$$S_2 = \left(S_1 \setminus \{h_C(\mathbf{x}_{i_1}) \leq h_F(\mathbf{x}_{i_1}) + p_{i_1}\}\right) \cup \{h_C(\mathbf{x}_{i_2}) \leq h_F(\mathbf{x}_{i_2}) + p_{i_2}\}.$$

Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be the solutions corresponding to Program 8.15 with constraint set S_1 and S_2 respectively. Then $\|\mathbf{p}_2\|_q \leq \|\mathbf{p}_1\|_q$. That is, choosing to enforce that i_2 prefers f_F over f_C is at least as good as choosing to enforce the preference of i_1 for f_F over f_C. To see this, consider the the scores of agents i_1 and i_2 under solution $\mathbf{p}_1$, i.e. $f_F(\mathbf{x}_{i_1}) + p_{1,i_1}$ and $f_F(\mathbf{x}_{i_2}) + p_{1,i_2}$. Suppose that scores p_{1,i_1} and p_{1,i_2} are permuted creating $\mathbf{p}'_1$, i.e. $p'_{1,i_1} = p_{1,i_2}$ and $p'_{1,i_2} = p_{1,i_1}$. Then $\mathbf{p}_1$ and $\mathbf{p}'_1$, have equal unfairness. Moreover, by the construction of S_1 and S_2 it must be the case that

$$p_{1,i_1} \geq f_C(\mathbf{x}_{i_1}) - f_F(\mathbf{x}_{i_1}) \geq f_C(\mathbf{x}_{i_2}) - f_F(\mathbf{x}_{i_2}),$$

implying that $\mathbf{p}'_1$ constitutes to a feasible solution to the program corresponding to popularity constraints S_2, i.e. $\|\mathbf{p}'_1\|_q \geq \|\mathbf{p}_2\|_q$. Since permuting elements of $\mathbf{p}_1$ does not affect the value of any ℓ_q-norm, it must be the case that $\|\mathbf{p}_1\|_q = \|\mathbf{p}'_1\|_q \geq \|\mathbf{p}_2\|_q$. Thus if groups are ordered such that for $i \in G_g$ we have

$$h_C(\mathbf{x}_i) - h_F(\mathbf{x}_i) < h_C(\mathbf{x}_{i+1}) - h_F(\mathbf{x}_{i+1})$$

, and then one need only consider adding the constraint $h_C(\mathbf{x}_{i+1}) \leq h_F(\mathbf{x}_{i+1}) + p_{i+1}$ if the constraint $h_C(\mathbf{x}_i) \leq h_F(\mathbf{x}_i) + p_i$ has already been selected.

Suppose G_1 and G_0 are ordered in such a manner. Then, to decide which constraints to include, it suffices to determine the intergroup decisions since the intragroup decisions are then determined by the agent order. Since there are at most $m = \gamma n$ unique sets of constraints which preserve orderings within groups, and each set of constraints corresponds to a polynomial time solvable program, Program 8.15 is solvable in time $\Theta(\gamma n T)$ where $\Theta(T)$

is the time required to solve a single program (either a linear program or a semidefinite program). Moreover, each corresponding program (namely Program 8.15 with Constraint 8.17 replaced by S) can be solved by Algorithm 4. At high level this algorithm takes the agents in S (i.e., the set of agents which should prefer f_P) and sets each p_i to the minimum value, in terms of magnitude, such that $i \in S$ prefers f_P. If fairness is not violated by this change to p_i is optimal. In the case when fairness is violated, the algorithm iteratively increases (or decreases) elements of $\mathbf{p}$ such that unfairness is strictly decreasing while minimally increasing $\|\mathbf{p}\|_q$.

To see the optimality and running time of Algorithm 4, consider the first step, namely $p_i = \min\big(0, h_C(\mathbf{x}_i) - h_F(\mathbf{x}_i)\big)$ for all $i \in S$. This setting of $\mathbf{p}$ causes all agents in S to prefer f_P and is clearly the minimum perturbation to do so. Therefore, if fairness is not violated, then $\mathbf{p}$ is optimal, and the running time is $\Theta(\gamma n)$.

In the case that fairness is violated, the scores on groups G_0 and G_1 need to be further altered. In particular, let

$$U(f_p, \mathcal{X}, G) = \mathcal{M}(f_P(\mathcal{X}); g = 1) - \mathcal{M}(f_P(\mathcal{X}); g = 0)$$

i.e., U is $\mathcal{U}$ without absolute value. The rate of change in U with respect to increasing p_i is given by $c_1^{(g_i)} - c_0^{(g_i)}$ for each $i \in [n]$. Therefore, if $U(f_p, \mathcal{X}, G) < -\beta$, unfairness can only be fixed by increasing scores on each group G_g with $c_1^{(g)} - c_0^{(g)} > 0$ and decreasing scores on each group G_G with $c_1^{(g)} - c_0^{(g)} < 0$. On the other hand, if $U(f_p, \mathcal{X}, G) > \beta$, then unfairness can only be fixed by increasing scores on each group G_g with $c_1^{(g)} - c_0^{(g)} < 0$ and decreasing scores on each group G_G with $c_1^{(g)} - c_0^{(g)} > 0$. When increasing scores on G_g the only constraint is that $h_F(\mathbf{x}_i) + p_i \leq 1$ for all $i \in [n]$, but when decreasing scores the constraint $0 \leq h_F(\mathbf{x}_i) + p_i$ for all $i \in [n]$ needs to be considered as well as $h_C(\mathbf{x}_j) \leq h_F(\mathbf{x}_j) + p_j$ for all $j \in S$.

With respect to fairness agents from the same group are exchangeable in the sense that increasing (or decreasing) the score of any agent in G_g has the same effect on unfairness as increasing (or decreasing) the score of any other agent in G_g. Specifically, for $i, j \in G_g$ unfairness is invariant to any change in p_i, p_j which preservers $p_i + p_j$. Therefore, ignoring popularity, no optimal solution will set $p_i < 0$ and $p_j > 0$. Moreover, when $p_i + p_j$ must be preserved, the terms $|p_i| + |p_j|$, $p_i^2 + p_j^2$ and $\max\big(|p_i|, |p_j|\big)$ are all minimized when $p_i = p_j = \frac{p_i + p_j}{2}$. Therefore for $q \in \{1, 2, \infty\}$, if one where to increase $|p_i|$, say by value ε, then

$p_i + \text{sign}(p_i)\varepsilon$ is no better than $p_j + \frac{\text{sign}(p_i)\varepsilon}{|G_{g_i}|}$ for each $j \in G_{g_i}$. That is, ignoring popularity, it is optimal to uniformly distribute the weight of $\mathbf{p}$ over each group.

When considering both popularity constraints, as well as the need for perturbations to constitute valid probabilities, it then optimal to uniformly increase the weight on all agents $i \in G_g$ such that neither of these constraints is violated. This value is given as ε_g at each iteration. Let $\mathbf{p}$ be the solution produced by Algorithm 4 and let $\mathbf{p}^*$ be any optimal solution. Since both are solutions $h_F(\mathcal{X}) + \mathbf{p}$ and $h_F(\mathcal{X}) + \mathbf{p}^*$ are both β-fair, and for each $j \in S$ $h_F(\mathbf{x}_j) + p_j \geq h_C(\mathbf{x}_j)$ and $h_F(\mathbf{x}_j) + p_j^* \geq h_C(\mathbf{x}_j)$. If $\|\mathbf{p}\|_q \leq \|\mathbf{p}^*\|_q$, then $\mathbf{p}$ is also an optimal solution. Assume, by way of contradiction, that $\|\mathbf{p}\|_q > \|\mathbf{p}^*\|_q$, and consider two cases:

$$1.)\ \mathcal{U}(h_F(\mathcal{X}) + \mathbf{p}^*, G) < \beta \ \text{ and } 2.)\ \ \mathcal{U}(h_F(\mathcal{X}) + \mathbf{p}^*, G) = \beta.$$

In case (1), if $q = 1, 2$ then $i \notin S$ implies $p_i^* = 0$, and if $q = \infty$ then $i \notin S$ implies $|p_i^*| \leq \max_{j \in S}\left(|p_j^*|\right)$. To see this, let $q = 1, 2$ and $j \notin S$. Suppose that $|p_i^*| > 0$ and $u = \beta - \mathcal{U}(h_F(\mathcal{X}) + \mathbf{p}^*, G)$. Then $|p_i^*|$ can be decreased by at least $\frac{u}{|c_1^{(g_i)} - c_0^{(g_i)}|}$ without violating fairness. Doing so results in a strict decrease to $\|\mathbf{p}^*\|_q$. When $q = \infty$ and identical argument holds for $|p_i^*| > \max_{j \in S}\left(|p_j^*|\right)$.

In case (2), we order each group according to the maximum feasible perturbation to each agent, w.l.o.g. we show this for G_0 when $c_1^{(0)} - c_0^{(0)} > 0$ (a symmetric argument holds in other cases). For $i \in G_0$ let $\delta_i = -h_F(\mathbf{x}_i)$ if $i \notin S$ and $\delta_j = h_C(\mathbf{x}_i) - h_F(\mathbf{x}_i)$ if $i \in S$. Order G_0 such that for $i, j \in G_0$, $i < j$ implies $\delta_i \geq \delta_j$. Suppose that $\delta_j \leq p_i^* \leq p_j^* \leq 0$. Then any solution which has $p_i = \frac{p_i^* + p_j^*}{2}$ (such as $\mathbf{p}'$) is both feasible and at least as optimal as $\mathbf{p}^*$.

At each iteration the sets G_g represent the set of agents whose scores can feasibly still be perturbed, i.e. further perturbing will not push the score below 0, above 1, or violate a constraint in S. The entries of $\mathbf{p}$ corresponding to either G_0 or G_1 are updated by ε_0 or ε_1 respectively. By the definition of ε_g, at least one agent is removed from either G_0 or G_1. There are at most n agents between the two sets, and thus at most n iterations are run. Each iteration takes at most time time $\Theta(n)$, since ε_g is computed as the minimum over at most n choices and at most n entries of $\mathbf{p}$ are updated. Therefore Algorithm 4 runs in time $\Theta(n^2)$.

Thus, since Algorithm 4 may be used to solve the instances of Program 8.15 arising in Algorithm 3 in time $\Theta(n^2)$, DOS can be solved by Algorithm 3 in time $\Theta(\gamma n^3)$. □

8.6.4 DOS for Randomized Resource Allocation

Next we turn our attention to resource allocation, in which $k < n$ equally desirable resources are allocated to a population of size n. Recall that the randomized allocation scheme given by $I(\mathcal{X}, G)$ assigns resources to agents where $\mathbb{E}\big[I_i(\mathcal{X}, G)\big] \in [0,1]$ gives the probability that agent i receives a resource with allocation performed over population $(\mathcal{X}, G)$. For notational convenience, we use $I(i) = \mathbb{E}\big[I_i(\mathcal{X}, G)\big]$ to represent the probability that agent i receives the resources and suppress the expectation and implicit dependence on the population $(\mathcal{X}, G)$.

Scarce resource allocation is particularly well suited for DOS as true labels (with respect to the allocation decision) are typically unknown. In this case, DOS postprocessing is given by,

$$\min_{\mathbf{p}\in[-1,1]^n} \; \|\mathbf{p}\|_q \tag{8.19}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} I_F(i) + p_i \leq k \tag{8.20}$$

$$\mathcal{U}\big(I_F + \mathbf{p},\; G\big) \leq \beta \tag{8.21}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big[I_C(i) \leq I_F(i) + p_i\big] \geq \gamma \tag{8.22}$$

We now show that DOS in resource allocation settings remains tractable.

Theorem 8.6.4. *Let I_C and I_F be a conventional and β-fair allocation scheme, respectively, and $\mathcal{U}$ be derived from an additive efficacy metric $\mathcal{M}$ which is independent of Y (e.g., PR). Then for $q \in \{1, 2, \infty\}$ Program 8.19 can be solved in time $\Theta(\gamma n T)$ by Algorithm 3 which returns a γ-popular, β-fair allocation if one exists.*

Proof Sketch. In the case of scarce resources, agents can again be ordered in an identical fashion to the classification setting (Theorem 8.6.3). Note that for any solution $\mathbf{p}$ and any $i, j \in G_g$, the resource constraint $\sum_{i=1}^{n} I_F(i) + p_i \leq k$ is invariant to any change in p_i, p_j, which preserves $p_i + p_j$. Thus a similar argument to Theorem 8.6.3, with a few caveats relating to infeasible solutions, holds. Specifically, this yields γn programs (either LPs or SDPs), each of which is solvable in time $\Theta(T)$. Thus DOS post processing for resource allocation can be computed in time $\Theta(\gamma n T)$. □

8.6.5 k-QLS for Randomized Classification

Finally, we explore k-QLS postprocessing for randomized classifiers. k-QLS creates k intervals by the quantiles of $h_F(\mathcal{X})$, where k is chosen by the model designer. Specifically, let ρ_ℓ be the maximum score associated with quantile ℓ of $h_F(\mathcal{X})$. Each interval is given as $I_\ell = [\rho_{\ell-1}, \rho_\ell]$, with the understanding that $\rho_0 = 0$ and $\rho_k = 1$. On each interval I_ℓ, and for each group g, a parameter $p_\ell^{(g)}$ is learned. At prediction time, $\mathbb{E}\big[f_P\big(\mathbf{x}_i\big)\big] = p_\ell^{(g_i)}$ for i s.t. $h_F(\mathbf{x}_i) \in I_\ell$, .

Finding the optimal lottery probabilities can formulated as the following optimization problem:

$$\min_{\mathbf{p}\in[0,1]^{2k}} \mathcal{L}\big(f_P, \mathcal{X}, Y\big) + \lambda \| f_F(\mathcal{X}) - f_P(\mathcal{X}) \|_q^q \tag{8.23}$$

$$\text{s.t.} \quad \mathcal{U}\big(f_P, D\big) \leq \beta \tag{8.24}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big[f_C(\mathbf{x}_i) \leq f_P(\mathbf{x}_i, g_i)\big] \geq \gamma, \tag{8.25}$$

where $\mathcal{L}$ is expected training error. As was the case for DOS postprocessing with randomized classifiers, the constraint that γ fraction of the population prefers f_P over f_C is discrete and non-convex. Indeed, unlike DOS, the k-QLS problem becomes strongly NP-hard.

Theorem 8.6.5. *Postprocessing to achieve γ-popularity and β-fairness with k-QLS (i.e., solving Program 8.23) is strongly NP-hard when models are randomized, and $\mathcal{U}$ is derived from an additive efficacy metric.*

Proof. We reduce from exact m knapsack (E-mKP) [64] (which is strongly NP-hard when coefficients are rational), in which a set of n items, each with weight and value $w_i, v_i \in \mathbb{Q}_{\geq 0}$ and a capacity W are given. The objective is to select exactly m items s.t. value is maximized and capacity is not violated. For each item, an interval can be created which contains exactly one agent i where $y_i = 0$, $h_C(\mathbf{x}_i)$ is determined by w_i and $h_F(\mathbf{x}_i)$ is determined by v_i. Ignoring popularity, and for FNR-fairness, it is optimal to negatively classify all such agents. However when m of these agents needs to prefer f_P, an optimal solution will set $h_P(\mathbf{x}_i) = h_C(\mathbf{x}_i)$ for exactly m agents. By the construction of $h_C(\mathbf{x}_i)$ and $h_F(\mathbf{x}_i)$, each agent with $h_p(\mathbf{x}_i) > 0$ increases FNR and loss by a term proportional to w_i and $-v_i$ respectively. Thus selecting m agents to prefer f_P s.t. loss is minimized while not violating fairness corresponds to selecting m items which maximize value while not violating the capacity constraint. □

Remark 8.6.6. *The intractability stems entirely from the model designer's ability to choose the number of quantiles k: if k is fixed, the problem can be solved in polynomial time as shown in the following theorem. In practice, we can fix k to be small, thus obtaining a tractable algorithm.*

Theorem 8.6.7. *Let f_C and f_F be a conventional and a β-fair randomized classifier respectively. Let U be derived from an additive efficacy metric $\mathcal{M}$. Then for a fixed number of quantiles k, Program 8.23 for $q = \{1, 2, \infty\}$ can be solved in polynomial time, thus obtaining γ-popular β-fair decisions.*

Proof. As was the case for DOS applied to randomized classifiers, k-QLS applied to randomized classifiers is tractable if a specific set of $m = \gamma n$ agents is required to prefer f_P, rather than any m agents. When the number of intervals is constant it is straightforward to induce an ordering on agents which explores only a polynomial number of constraint sets. Specifically, let

$$G_{(g,\ell)} = \{i \in [n] : g_i = g \ \text{ and } \ h_F(\mathbf{x}_i) \in I_\ell\},$$

then agents in each G_g can be ordered by the magnitude of $p_\ell^{(g)}$ required such that they prefer f_P to f_C. Order G_g such that for $i, j \in G_g$ if $i < j$ then $h_C(\mathbf{x}_j) \leq h_C(\mathbf{x}_i)$, then if agent $i \in G_g$ prefers f_P to f_C, so does every $j \leq i$. There are $2k$ such sets, each containing at most n/k agents. Since the popularity over each G_g can be parameterized by the identity of the agent with the largest value of $h_C(\mathbf{x})$ who prefers f_P, there are no more than $(\gamma n)^k$ unique values under this parameterization, and thus no more than $(\gamma n)^k$ sets of constraints need be examined; each examination requires only polynomial time. □

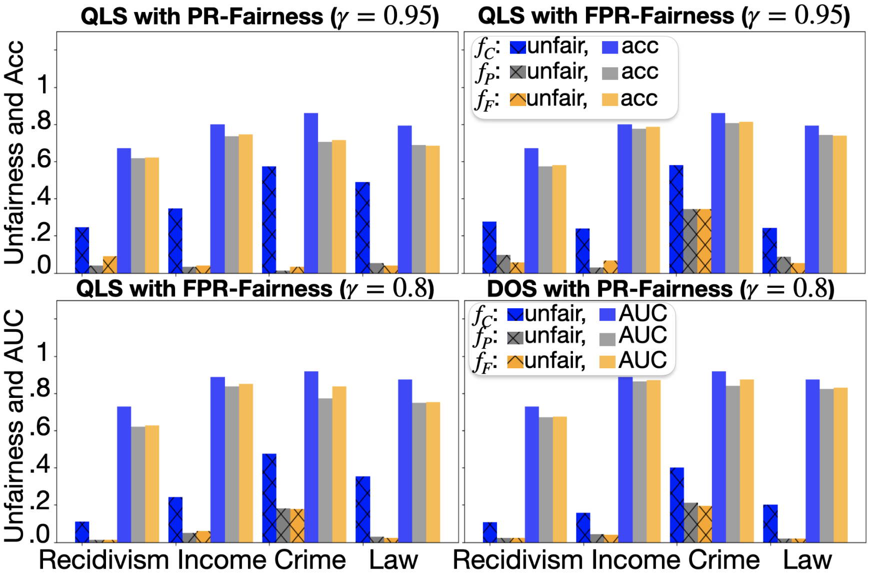

Figure 8.1: Model performance and unfairness on test data (3-fold average) for deterministic models with $\gamma = 0.95$ (top) and randomized models with $\gamma = 0.8$ (bottom). The conventional classifier f_C, fair classifier f_F (learned via reductions), and the fair popular classifier f_P (learned via our postprocessing technique), each using Logistic Regression.

8.7 Experiments

We next empirically investigate the relationship between popularity and fairness, and evaluate the efficacy of the proposed postprocessing algorithms. The experimental setup is identical to that of previous chapters and follows from Chapter 4.

We begin by examining the efficacy of our proposed postprocessing techniques DOS and k-QLS (k=10). When classifiers are deterministic, performance is measured using balanced accuracy (balanced w.r.t. Y). When classifiers are randomized, performance is measured using ROC-AUC, calculated over model scores (i.e., expected outcomes).

Remark 8.7.1. *Both k-QLS and DOS may require solving a large number of LPs or SDPs, which may be expensive. However, both methods can be efficiently implemented in practice by either solving the programs in parallel, trimming down the number of programs with heuristics, or replacing all programs with a single integer program. The latter being the most efficient, typically finishing in under 60 seconds. Further details on these methods, and exact running times, are provided in*

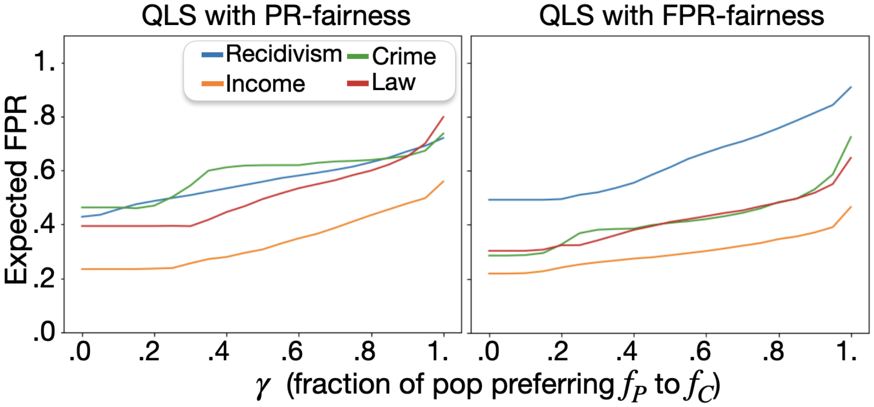

Figure 8.2: Expected False Positive Rate (FPR) of k-QLS, on randomized classifiers, as a function of γ.

	Recidivism	Income	Crime	Law School
Deterministic DOS	0.001	0.001	0.001	0.001
Deterministic k-QLS	0.021	0.092	0.033	0.026
Randomized DOS	0.351	1.645	0.931	0.619
Randomized k-QLS	44.121	67.121	54.379	52.947

Table 8.2: Running time in seconds (rounded to three digits), of DOS and k-QLS for randomized models with $\gamma = 0.8$ and deterministic models with $\gamma = 0.95$. For deterministic models, the polynomial time algorithms are run, for randomized model a single integer program is run. Reported times are averaged across PR, TPR, and FPR fairness as well as all base model types (Logistic Regression, Gradient Boosted Trees, Support Vector Machine, and Neural Networks). Since DOS and k-QLS are postprocessing methods, reported running times do not include running time of the base models.

By casting popularity as an integer-valued constraint the corresponding integer program can be solved directly by modern solvers. In our experiments we solve those programs with CPLEX, and present the average running time for each approach in table 8.2.

Figure 8.1 shows that both k-QLS and DOS are able to achieve high levels of γ-popularity and β-fairness with little degradation in performance. In particular, deterministic classifiers (due to their higher natural popularity) are able to achieve greater levels of popularity compared to randomized models, with similar levels of degradation to performance. We observe similar results for other combinations of dataset, efficacy metric, and classifier type (Section B of the Appendix).

Finally, we consider the extent to which popularity may skew model efficacy. In particular, as the popularity coefficient γ increases, a larger fraction of the population is guaranteed to have scores from f_P, which are at least as large as those from f_C. Since popularity constraints ensure that agents scores do not decrease, achieving higher levels of popularity (i.e., higher γ) also incentivize the resulting f_P to maintain false positive errors made by f_C. Thus one would expect FPR to increase with γ.

This phenomenon is shown in Figure 8.2, which demonstrates that as γ increases, so does expected FPR. Although the expected FPRs vary between datasets and fairness definitions, the rate of increase is relatively similar across instances.

8.8 Discussion

As shown in Chapter 6 the deployment of group-fair classifiers, in place of conventional classifiers, may result large fractions of a population perceiving that they are made worse off by the change. To capture these effects we introduce the notion of popularity, which measures the fraction of agents preferring one classifier over another, and propose two postprocessing techniques (DOS and k-QLS) for achieving popularity while retaining good fairness properties. Both techniques provide efficient solutions for deterministic and randomized classifiers, as well as scarce resource allocation. We note that while in practice postprocessing can achieve popularity and fairness with minimal degradation to model performance, requiring higher levels of popularity can actually entrench any false positive errors made by the conventional model. Consequently, application of the proposed techniques need to carefully analyze the trade-offs not merely between popularity, group fairness, and overall accuracy, but also with specific measures of error, particularly the false positive rate.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification, 2018.

[3] A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

[4] E. Akyol, C. Langbort, and T. Basar. Price of transparency in strategic machine learning, 2016.

[5] A. A. Alsheikh-Ali, W. Qureshi, M. H. Al-Mallah, and J. P. Ioannidis. Public availability of published research data in high-impact journals. *PloS one*, 6(9):e24357, 2011.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[7] P. Awasthi, M. Kleindessner, and J. Morgenstern. Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics*, pages 1770–1780. PMLR, 2020.

[8] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[9] O. Ben-Porat, F. Sandomirskiy, and M. Tennenholtz. Protecting the protected group: Circumventing harmful fairness. *arXiv preprint arXiv:1905.10546*, 2019.

[10] E. Ben-Porath, E. Dekel, and B. L. Lipman. Optimal allocation with costly verification. *American Economic Review*, 104(12):3779–3813, 2014.

[11] J. Berger, M. Osterloh, K. Rost, and T. Ehrmann. How to prevent leadership hubris? comparing competitive selections, lotteries, and their combination. *The Leadership Quarterly*, 31(5):101388, 2020.

[12] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[13] J. Blocki, N. Christin, A. Datta, A. Procaccia, and A. Sinha. Audit games with multiple defender resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[14] J. Blocki, N. Christin, A. Datta, A. D. Procaccia, and A. Sinha. Audit games. *arXiv preprint arXiv:1303.0356*, 2013.

[15] T. Börgers and D. Krahmer. *An introduction to the theory of mechanism design.* Oxford University Press, USA, 2015.

[16] G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pages 6045–6061. PMLR, 2022.

[17] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

[18] J. Callan and A. Moffat. Panel on use of proprietary data. In *ACM SIGIR Forum*, volume 46, pages 10–18. ACM New York, NY, USA, 2012.

[19] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] A. Camacho and E. Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3:41–65, 05 2011.

[21] R. Canetti, A. Cohen, N. Dikkala, G. Ramnarayan, S. Scheffler, and A. Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, pages 309–318, 2019.

[22] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees, 2020.

[23] L. E. Celis, A. Mehrotra, and N. K. Vishnoi. Fair classification with adversarial perturbations, 2021.

[24] Y. Chen, J. Wang, and Y. Liu. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 236, 2020.

[25] J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099, 2020.

[26] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[27] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[28] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. *Information Systems/Algoritmi*, 2008.

[29] C. Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[30] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

[31] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 55–70, New York, NY, USA, 2018. Association for Computing Machinery.

[32] D. Dua and C. Graff. UCI machine learning repository, 2017.

[33] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness, 2011.

[34] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[35] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[36] A. Erlanson and A. Kleiner. Costly verification in collective decisions. *Theoretical Economics*, 15(3):923–954, 2020.

[37] A. Estornell, S. Das, and Y. Vorobeychik. Incentivizing truthfulness through audits in strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5347–5354, 2021.

[38] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact, 2015.

[39] V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[40] M. Hardt, M. Jagadeesan, and C. Mendler-Dünner. Performative power. *arXiv preprint arXiv:2203.17232*, 2022.

[41] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA, 2016. Association for Computing Machinery.

[42] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[43] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning, 2016.

[44] L. Hong and S. E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[45] L. Hu and Y. Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[46] L. Hu, N. Immorlica, and J. W. Vaughan. The disparate effects of strategic manipulation. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019.

[47] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[48] X. Huang, Z. Li, Y. Jin, and W. Zhang. Fair-adaboost: Extending adaboost method to achieve fair classification. *Expert Systems with Applications*, 202:117240, 2022.

[49] T. Jang, P. Shi, and X. Wang. Group-aware threshold adaptation for fair classification. *arXiv preprint arXiv:2111.04271*, 2021.

[50] T. Jang, P. Shi, and X. Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995, 2022.

[51] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[52] S. Jung, S. Chun, and T. Moon. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10348–10357, 2022.

[53] F. Kamiran and T. Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.

[54] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

[55] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects, 2020.

[56] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

[57] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

[58] A.-H. Karimi, J. Von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[59] D. Karlan and J. Zinman. Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1):433–464, 2010.

[60] D. Karlan and J. Zinman. Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1):433–464, 2010.

[61] M. Kasy and R. Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

[62] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, 2018.

[63] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[64] H. Kellerer, U. Pferschy, and D. Pisinger. Knapsack problems. In *Knapsack problems*. Springer, 2004.

[65] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[66] M. Kleindessner, S. Samadi, M. B. Zafar, K. Kenthapadi, and C. Russell. Pairwise fairness for ordinal regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3381–3417. PMLR, 2022.

[67] R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[68] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[69] P. Lahoti, K. P. Gummadi, and G. Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

[70] S. Levanon and N. Rosenfeld. Strategic classification made practical, 2021.

[71] S. Levanon and N. Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pages 12593–12618. PMLR, 2022.

[72] B. Li and Y. Vorobeychik. Evasion-robust classification on binary domains. *ACM Transactions on Knowledge Discovery from Data*, 12(4):1–32, 2018.

[73] B. Li, Y. Vorobeychik, and X. Chen. A general retraining framework for scalable adversarial classification, 2016.

[74] X. Li, P. Wu, and J. Su. Accurate fairness: Improving individual fairness without trading accuracy. *arXiv preprint arXiv:2205.08704*, 2022.

[75] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097, 2022.

[76] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[77] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[78] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[79] T. Lundy, A. Wei, H. Fu, S. D. Kominers, and K. Leyton-Brown. Allocation for social good: auditing mechanisms for utility maximization. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 785–803, 2019.

[80] T. Lundy, A. Wei, H. Fu, S. D. Kominers, and K. Leyton-Brown. Allocation for social good: Auditing mechanisms for utility maximization. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, page 785–803, 2019.

[81] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[82] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[83] D. McNamara, C. S. Ong, and R. C. Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 263–270, 2019.

[84] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[85] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[86] C. Mendler-Dünner, F. Ding, and Y. Wang. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems*, 35:31171–31185, 2022.

[87] S. Milli, J. Miller, A. D. Dragan, and M. Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 230–239, New York, NY, USA, 2019. Association for Computing Machinery.

[88] S. Milli, J. Miller, A. D. Dragan, and M. Hardt. The social cost of strategic classification. In *Conference on Fairness, Accountability, and Transparency*, page 230–239, 2019.

[89] D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR, 2020.

[90] R. B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979.

[91] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 466–477, New York, NY, USA, 2021. Association for Computing Machinery.

[92] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[93] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

[94] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif. Theoretical evidence for adversarial robustness through randomization. In *Neural Information Processing Systems*, 2019.

[95] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration, 2017.

[96] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[97] J. Pombal, P. Saleiro, M. A. Figueiredo, and P. Bizarro. Prisoners of their own devices: How models induce data bias in performative prediction. *arXiv preprint arXiv:2206.13183*, 2022.

[98] P. Putzel and S. Lee. Blackbox post-processing for multiclass fairness. *arXiv preprint arXiv:2201.04461*, 2022.

[99] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[100] A. Ross, H. Lakkaraju, and O. Bastani. Learning models for actionable recourse. *Advances in Neural Information Processing Systems*, 34:18734–18746, 2021.

[101] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Neural Information Processing Systems*, volume 32, 2019.

[102] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[103] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32, 2019.

[104] H. Shimao, W. Khern-am nuai, K. Kannan, and M. C. Cohen. Strategic best response fairness in fair machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 664–664, 2022.

[105] T. Tassier and F. Menczer. Social network structure, segregation, and equality in a labor market with referral hiring. *Journal of Economic Behavior & Organization*, 66(3-4):514–528, 2008.

[106] L. Tong, B. Li, C. Hajaj, C. Xiao, N. Zhang, and Y. Vorobeychik. Improving robustness of ML classifiers against realizable evasion attacks using conserved features. In *USENIX Security Symposium*, pages 285–302, 2019.

[107] S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16926–16937. Curran Associates, Inc., 2021.

[108] B. Ustun, Y. Liu, and D. Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.

[109] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[110] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[111] S. Venkatasubramanian and M. Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 284–293, 2020.

[112] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups, 2020.

[113] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021.

[114] L. Wightman and L. S. A. Council. *LSAC National Longitudinal Bar Passage Study.* LSAC research report series. Law School Admission Council, 1998.

[115] V. Xinying Chen and J. Hooker. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, pages 1–39, 2023.

[116] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021.

[117] K. Yang, B. Huang, J. Stoyanovich, and S. Schelter. Fairness-aware instrumentation of preprocessing˜ pipelines for machine learning. In *Workshop on Human-In-the-Loop Data Analytics (HILDA'20)*, 2020.

[118] S. Yeom and M. Fredrikson. Individual fairness revisited: Transferring techniques from adversarial robustness, 2020.

[119] T. Young and S. Hopewell. Methods for obtaining unpublished data. *Cochrane Database of Systematic Reviews*, 11, 2011.

[120] M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ml models with sensitive subspace robustness, 2020.

[121] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[122] C. Zhang, H. Zhang, and C.-J. Hsieh. An efficient adversarial attack for tree ensembles. *Advances in Neural Information Processing Systems*, 33:16165–16176, 2020.

[123] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, and C. Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

[124] Z. Zheng and B. Padmanabhan. Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution. *Management Science*, 52(5):697–712, 2006.

Appendix A

A.1 Fairness Reversals

Equivalence of α- and β-fairness

Lemma 5.2.1: *Suppose $\mathcal{M}$ is defined by PR, FPR, or TPR, and the hypothesis class is expressive enough to produce either the constant function $f(\mathbf{x}) = 1$ or the constant function $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$. Then for any $\alpha \in [0, 1]$ there exists a $\beta \geq 0$ such that the resulting optimal β-fair classifier f_β is also an optimal α-fair classifiers. Conversely for any $\beta \geq 0$ there exists an $\alpha \in [0, 1]$ such that the resulting optimal α-fair classifier f_α is also an optimal β-fair classifier.*

Proof. To prove this claim we first outline the general structure of the relationship between α and β fairness. For any $\alpha \in [0, 1]$, setting $\beta = U_\alpha$ where U_α is the unfairness of the optimal α-fair model will cause the optimal β fair model to also be an optimal α-fair model. For any $\beta \geq 0$, the constant function $f(\mathbf{x}) = 1$ or $f(\mathbf{x}) = 0$ is always feasible for PR, TPR, or FPR fairness. For any Pareto front of model (induced by loss and unfairness) there will always be an optimal β-fair model in that front, and because the objective of α-fairness is linear in α, all classifiers in the front can be achieved by a some value of α. First note that solutions to Programs 5.2, 5.3 are determined uniquely by the outcomes of f, namely $f(\mathcal{X})$. Thus $\mathcal{H}$ can be considered as consisting only of models which produce distinct values of $f(\mathcal{X})$.

To explicitly show this, we first prove the forward direction of our claim. Let $\alpha \in [0, 1]$, f_α be an optimal α-fair classifier and $\beta = U(f_\alpha, \mathcal{X}, Y, G)$. For this setting of β, the model f_α is a feasible solution to Program 5.2, so

$$\mathcal{L}(f_\beta, \mathcal{X}, Y) \leq \mathcal{L}(f_\alpha, \mathcal{X}, Y) \quad \text{and} \quad U(f_\beta, \mathcal{X}, Y, G) \leq U(f_\alpha, \mathcal{X}, Y, G),$$

implying that,

$$(1-\alpha)\mathcal{L}(f_\beta, \mathcal{X}, Y) + \alpha U(f_\beta, \mathcal{X}, Y, G) \leq (1-\alpha)\mathcal{L}(f_\alpha, \mathcal{X}, Y) + \alpha U(f_\alpha, \mathcal{X}, Y, G)$$

Thus for any $\alpha \in [0,1]$, setting $\beta = U(f_\alpha, \mathcal{X}, Y, G)$ will yield a model f_β which is also an optimal α-fair classifier.

To prove the converse direction we will show the following three facts, 1.) only models in a Pareto front(with respect to loss and unfairness) need be considered, 2.) an optimal β-fair model always exists in this Pareto front, and 3.) any model within this Pareto front can be found by some choice of α. Before showing these facts, we need two additional pieces of setup. First, when $\mathcal{M}$ is defined by PP, TPR, or FPR, either constant function $f(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$ or $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ will result in $U(f, \mathcal{X}, Y, G) = 0$. Thus, for any choice of $\beta \geq 0$ there always exists a feasible solution to Program 5.2. Second, for any two models f and f', we say $f \succ f'$ if $\mathcal{L}(f, \mathcal{X}, Y) \leq \mathcal{L}(f', \mathcal{X}, Y)$ and $U(f, \mathcal{X}, Y, G) \leq U(f', \mathcal{X}, Y, G)$, and at least one inequality is strict. Define the Pareto front of hypothesis class $\mathcal{H}$ with respect to loss and unfairness as,

$$P = \{f \in \mathcal{H} : \nexists f' \in \mathcal{H} \;\; \text{s.t.} \;\; f' \succ f\}.$$

Now, let $\beta \geq 0$, then there exists some $f \in P$ which is an optimal β-fair classifier by identical reasoning as the forward direction. Thus, we need only show that for any $f \in P$ there exists a corresponding α such that solving Program 5.3 will yield a classifier f_α which is equivalent to f in the sense that both classifier have equal loss and unfairness. To do this, we first show that for any $\alpha \in [0,1]$ the classifier f_α is equivalent to some $f \in P$. Let f' be the optimal α-fair classifier the hypothesis class is restricted to P. Then by the definition of f_α,

$$\mathcal{L}(f_\alpha, \mathcal{X}, Y) \leq \mathcal{L}(f', \mathcal{X}, Y) \;\; \text{and} \;\; U(f_\alpha, \mathcal{X}, Y, G) \leq U(f', \mathcal{X}, Y, G),$$

and by definition of P, neither inequality can be strict. Hence both the loss and unfairness of f' and f_α are equal. This equivalence allows us to only focus on optimal α-fair classifiers from P, rather than from H.

To show that any f in P can be obtained by the choice of some α when the space of possible models is reduced to P, we induce an ordering on the classifiers in P via the value

$U(f, \mathcal{X}, Y, G)$. Note that

$$(1-\alpha)\mathcal{L}(f, \mathcal{X}, Y) + \alpha U(f, \mathcal{X}, Y, G)$$

is linear in α, and for $\alpha = 0$ or $\alpha = 1$ Program 5.3, restricted to P, will return respectively the model with the lowest loss and lowest unfairness. Thus both models at either extreme of our ordering on P can be obtained by some α, namely 0 or 1.

Next consider any model $f' \in P$ which is not at either extreme and let f_l and f_u be any two models in P such that

$$U(f_\ell, \mathcal{X}, Y, G) < U(f', \mathcal{X}, Y, G) < U(f_u, \mathcal{X}, Y, G)$$

and such that there exists values $\alpha_u \leq \alpha_l$ where f_u and f_l are the respective optimal α-fair classifiers in P (such classifiers must exist, namely those with the lowest loss and lowest unfairness). Moreover, not that if the above inequalities are not strict, then then the definition of P implies that the two corresponding classifiers are equivalent in terms of loss and unfairness.

Now let,

$$\alpha_{\text{sup}} = \sup\left\{\alpha : U(f', \mathcal{X}, Y, G) < U(f_\alpha, \mathcal{X}, Y, G)\right\}$$
$$\alpha_{\text{inf}} = \inf\left\{\alpha : U(f_\alpha, \mathcal{X}, Y, G) < U(f', \mathcal{X}, Y, G)\right\}.$$

Then $\alpha_{\text{inf}} \leq \alpha_{\text{sup}}$, and if the two are not equal, then any $\alpha \in (\alpha_{\text{inf}}, \alpha_{\text{sup}})$ with result in f' being the optimal α-fair classifier by the definition of P.

Now, suppose that $\alpha_{\text{inf}} = \alpha_{\text{sup}}$. Then since,

$$U(f_{\alpha_{\text{inf}}}, \mathcal{X}, Y, G) \leq U(f', \mathcal{X}, Y, G) \leq U(f_{\alpha_{\text{sup}}}, \mathcal{X}, Y, G)$$

it must be the case that

$$U(f_{\alpha_{\text{inf}}}, \mathcal{X}, Y, G) = U(f', \mathcal{X}, Y, G) \;\text{ and }\; \mathcal{L}(f_{\alpha_{\text{inf}}}, \mathcal{X}, Y, G) \leq \mathcal{L}(f', \mathcal{X}, Y, G)$$

But since $f' \in P$, it must also be the case that $\mathcal{L}(f_{\alpha_{\text{inf}}}, \mathcal{X}, Y, G) = \mathcal{L}(f', \mathcal{X}, Y, G)$. Thus $f_{\alpha_{\text{inf}}}$ and f' are equivalent, and for any $f' \in P$ there exists an α such that the optimal α-fair classifier is equivalent to f'. Combining this with the fact that for any β there exists an optimal β-fair classifier in P, implies that for any choice of β there exists an α such □

Lemma A.1.1. *Suppose that $\mathbb{P}\big(y=1|x\big)$ has a single crossing with $\mathbb{P}(y=1)$. Then error is negatively unimodal w.r.t. θ and the optimal base threshold is θ_C s.t. $\mathbb{P}\big(y=1|\theta_C\big)=\mathbb{P}\big(y=1\big)$.*

Proof: (Lemma A.1.1). The error of a classifier $f(x)=\mathbb{I}[x\geq\theta]$ is given by,

$$\begin{aligned}
& 1-\mathbb{P}\big(\mathbb{I}[x\geq\theta]=y\big)\\
= \;& 1-\mathbb{P}\big(x\geq\theta,y=1\big)-\mathbb{P}\big(x\leq\theta,y=0\big)\\
= \;& \mathbb{P}\big(y=0\big)+\mathbb{P}\big(x\leq\theta,y=1\big)-\mathbb{P}\big(x\leq\theta,y=0\big)
\end{aligned}$$

Since x is a continuous random variable and the terms involving θ are joint CDFs with well defined conditional PDFs, the derivative of the above expression w.r.t. to θ, exists and is equal to

$$\begin{aligned}
& p_{y,x}(y=1,x=\theta)-p_{y,x}(y=0,x=\theta)\\
= \;& p_x(x=\theta)\big(\mathbb{P}(y=1|x=\theta)-\mathbb{P}(y=0|x=\theta)\big)\\
= \;& p_x(x=\theta)\big(2\mathbb{P}(y=1|x=\theta)-1\big)
\end{aligned}$$

Since $\mathbb{P}\big(y=1|x=\theta\big)$ has a single crossing with $\mathbb{P}(y=1)$, the above derivative is *split* by the value 0, thus error is *negatively unimodal* with global minima at any θ_C s.t. $\mathbb{P}\big(y=1|x=\theta_C\big)=\mathbb{P}(y=1)$. □

Lemma A.1.2. *Suppose that fairness is defined in terms of Positive Rate (PR) and that $\mathbb{P}\big(g=1|x\big)$ has a single crossing with $\mathbb{P}\big(g=1\big)$, then*

1. *$PR_{\mathcal{D}}(\theta|g=1)\geq PR_{\mathcal{D}}(\theta|g=0)$ for any $\theta\in[0,1]$, (i.e. group 1 is advantaged under any threshold classifier), and*
2. *the unfairness term $\big|PR_{\mathcal{D}}(\theta|g=1)-PR_{\mathcal{D}}(\theta|g=0)\big|$ is* positively unimodal *w.r.t. θ and is maximized at any θ_U s.t. $\mathbb{P}\big(g=1|x=\theta_U\big)=\mathbb{P}\big(g=1\big)$.*

Proof: (Lemma A.1.2). For a classifier $f(x)=\mathbb{I}[x\geq\theta]$, we begin by demonstrating (1) the *unimodality* of $\mathbb{P}\big(x\geq\theta|g=1\big)-\mathbb{P}\big(x\geq\theta|g=0\big)$ and then use this propriety to show (2) the equivalence between $\mathbb{P}\big(x \geq \theta|g=1\big) - \mathbb{P}\big(x \geq \theta|g=0\big)$ and the unfairness term

$$\Big|\mathbb{P}\big(x \geq \theta|g=1\big) - \mathbb{P}\big(x \geq \theta|g=0\big)\Big|.$$

First, note that

$$\begin{aligned}
&= \mathbb{P}\big(x \geq \theta|g=1\big) - \mathbb{P}\big(x \geq \theta|g=1\big) \\
&= \frac{\mathbb{P}\big(g=1, x \geq \theta\big)}{\mathbb{P}\big(g=1\big)} - \frac{\mathbb{P}\big(g=0, x \geq \theta\big)}{\mathbb{P}\big(g=0\big)} \\
&= \frac{\mathbb{P}\big(g=1\big) - \mathbb{P}\big(g=1, x \leq \theta\big)}{\mathbb{P}\big(g=1\big)} - \frac{\mathbb{P}\big(g=0\big) - \mathbb{P}\big(g=0, x \leq \theta\big)}{\mathbb{P}\big(g=0\big)} \\
&= 1 - \frac{\mathbb{P}\big(g=1, x \leq \theta\big)}{\mathbb{P}\big(g=1\big)} - 1 + \frac{\mathbb{P}\big(g=0\big)\mathbb{P}\big(g=0, x \leq \theta\big)}{\mathbb{P}\big(g=0\big)} \\
&= -\frac{\mathbb{P}\big(g=1, \leq \theta\big)}{\mathbb{P}\big(g=1\big)} + \frac{\mathbb{P}\big(g=0, x \leq \theta\big)}{\mathbb{P}\big(g=0\big)}
\end{aligned}$$

Since each term involving θ is a joint CDF, the derivative of this term w.r.t to θ exists and is equal to

$$\begin{aligned}
& \frac{p_{g,x}(g=0, x=\theta)}{\mathbb{P}\big(g=0\big)} - \frac{p_{g,x}(g=1, x=\theta)}{\mathbb{P}\big(g=1\big)} \\
&= \frac{\mathbb{P}\big(g=0|x=\theta\big)p_x(x=\theta)}{\mathbb{P}\big(g=0\big)} - \frac{\mathbb{P}\big(g=1|x=\theta\big)p_x(x=\theta)}{\mathbb{P}\big(g=1\big)} \\
&= \frac{\big(1-\mathbb{P}\big(g=1|x=\theta\big)\big)h_x(x=\theta)}{\mathbb{P}\big(g=0\big)} - \frac{\mathbb{P}\big(g=1|x=\theta\big)h_x(x=\theta)}{\mathbb{P}\big(g=1\big)} \\
&= p_x(x=\theta)\frac{\mathbb{P}\big(g=1\big) - \mathbb{P}\big(g=1|x=\theta\big)}{\mathbb{P}\big(g=1\big)\mathbb{P}\big(g=0\big)}
\end{aligned}$$

Since $\mathbb{P}\big(g=1|x\big)$ is *split* by the value $\mathbb{P}\big(g=1\big)$ the above term is *split* by the value 0, thus by Lemma the term $\mathbb{P}\big(x \geq \theta|g=1\big) - \mathbb{P}\big(x \geq \theta|g=0\big)$ is *positively unimodal*, and is maximized at any θ_U s.t.

$$p_x(x=\theta_U)\frac{\mathbb{P}\big(g=1\big) - \mathbb{P}\big(g=1|x=\theta_U\big)}{\mathbb{P}\big(g=1\big)\mathbb{P}\big(g=0\big)} = 0$$

Since $p_x(x=\theta) > 0$ any such θ_U has the propriety that $\mathbb{P}\big(g=1|x=\theta_U\big) = \mathbb{P}\big(g=1\big)$. Thus concluding the proof of (2).

We now use (2) to show that (1) immediately follows. Note that for $\theta \in \{0, 1\}$ we have $\mathbb{P}\big(x \geq \theta | g = 1\big) = \mathbb{P}\big(x \geq \theta | g = 0\big)$. Since the function is *positively unimodal* and $\mathbb{P}\big(g = 1\big) > 0$ neither $\theta = 0$ nor $\theta = 1$ can be points corresponding to local maximums, hence for any θ we have

$$\begin{aligned}
&\mathbb{P}\big(x \geq \theta | g = 1\big) - \mathbb{P}\big(x \geq \theta | g = 0\big) \\
\geq\ &\mathbb{P}\big(x \geq 1 | g = 1\big) - \mathbb{P}\big(x \geq 1 | g = 0\big) \\
=\ &0
\end{aligned}$$

□

Lemma A.1.3. *Suppose that fairness is defined by either True Positive Rate or False Positive Rate and that g, y are conditionally independent given x. Suppose further that $\mathbb{P}\big(g = 1|x\big)$ has a single crossing with $\mathbb{P}\big(g = 1|y = 1\big)$ in the TPR case and by $\mathbb{P}\big(g = 1|y = 0\big)$ in the FPR case. Then when $\mathcal{M}$ is TPR or FPR,*

1. $\mathcal{M}_{\mathcal{D}}(\theta|g = 1) \geq \mathcal{M}_{\mathcal{D}}(\theta|g = 0)$ *for any* $\theta \in [0, 1]$, *(i.e. group 1 is advantaged under any threshold classifier), and*
2. *the unfairness term* $\big|\mathcal{M}_{\mathcal{D}}(\theta|g = 1) - \mathcal{M}_{\mathcal{D}}(\theta|g = 0)\big|$ *is* positively unimodal *w.r.t.* θ *and is maximized at any* θ_U *s.t.* $\mathbb{P}\big(g = 1|x = \theta_U\big) = \mathbb{P}\big(g = 1|y = 1\big)$ *in the TPR case and* $\mathbb{P}\big(g = 1|x = \theta_U\big) = \mathbb{P}\big(g = 1|y = 0\big)$ *in the FPR case.*

Proof: (Lemma A.1.3). This proof follows identically from Lemma A.1.2 when replacing terms related to PR with terms related to either FPR or TPR. □

Experiments

Figure A.1 show the single crossing conditions between $\mathbb{P}\big(y = 1|x\big)$, and $\mathbb{P}\big(g = 1|x\big)$, and their respective constant functions given in Lemmas A.1.1, A.1.2, A.1.3. We see that in each dataset, the single crossing conditions approximately holds in the sense that when the condition is violated, (i.e. crossing the respective horizontal line more than once) the violation is small in magnitude. Recall that the single crossing propriety implies the unimodality of the error and unfairness terms, which we see in Figure A.2 Small violations (both in magnitude and duration) of the single crossing condition amount to small changes in the derivative of

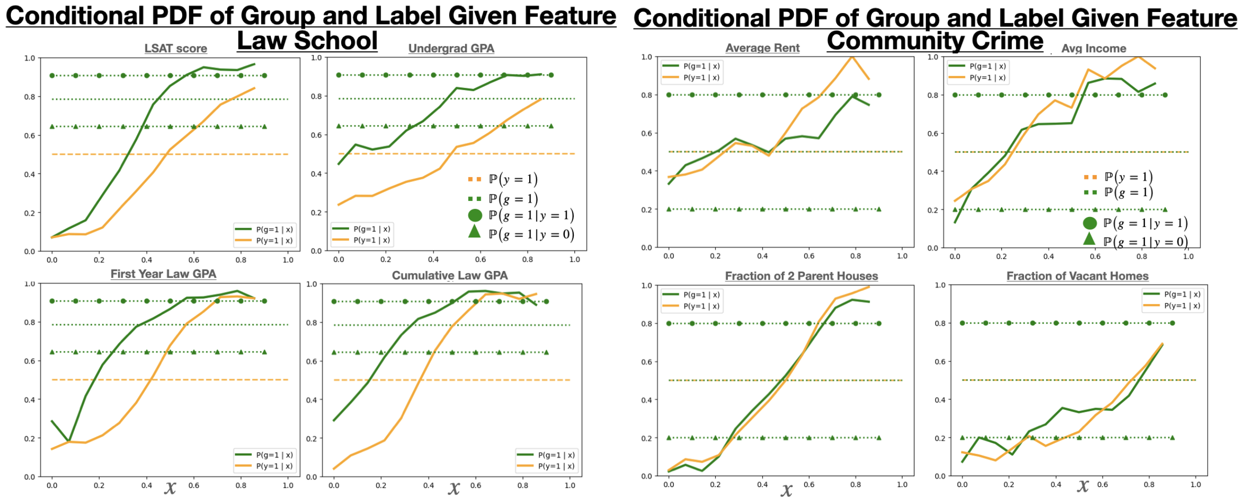

Figure A.1: Probabilities of group membership g (green) and true label y (orange). Probabilities conditioned on the feature x are given as solid lines, while those unconditioned are given as dotted, or dashed, lines. Recall that if the conditioned probabilities $\mathbb{P}(g = 1|x)$ and $\mathbb{P}(y = 1|x)$ having a single crossing with the respective unconditioned value (outlined in Lemmas A.1.1, A.1.2, A.1.3) then error and unfairness will be unimodal w.r.t. to the threshold θ. For example, in the case of PR fairness, if $\mathbb{P}(g = 1|x)$ has a single crossing with $\mathbb{P}(g = 1)$ and $\mathbb{P}(y = 1|x)$ has a single crossing with $\mathbb{P}(y = 1)$ then error and unfairness are unimodal w.r.t. to θ.

error or unfairness, which in term does not consequentially impact the unimodality of either term from an empirical perspective.

Figures A.4 and A.3 demonstrate additional instances of fairness reversals for varying levels of fairness given by α.

A.2 Individual Welfare

Figures A.5, A.6, A.7 show the popularity of the fair classifier f_C for the population (pop), advantaged group G_1, and disadvantaged group G_0. Unlike other fair learning schemes, we see that the EqOdds classifier, presented in Figure A.7, results one group always having popularity-1. This is due to the fact that EqOdds only perturbs the scores of a single, implying that the other group will always be at least as good under the fair model as they are under the conventional model.

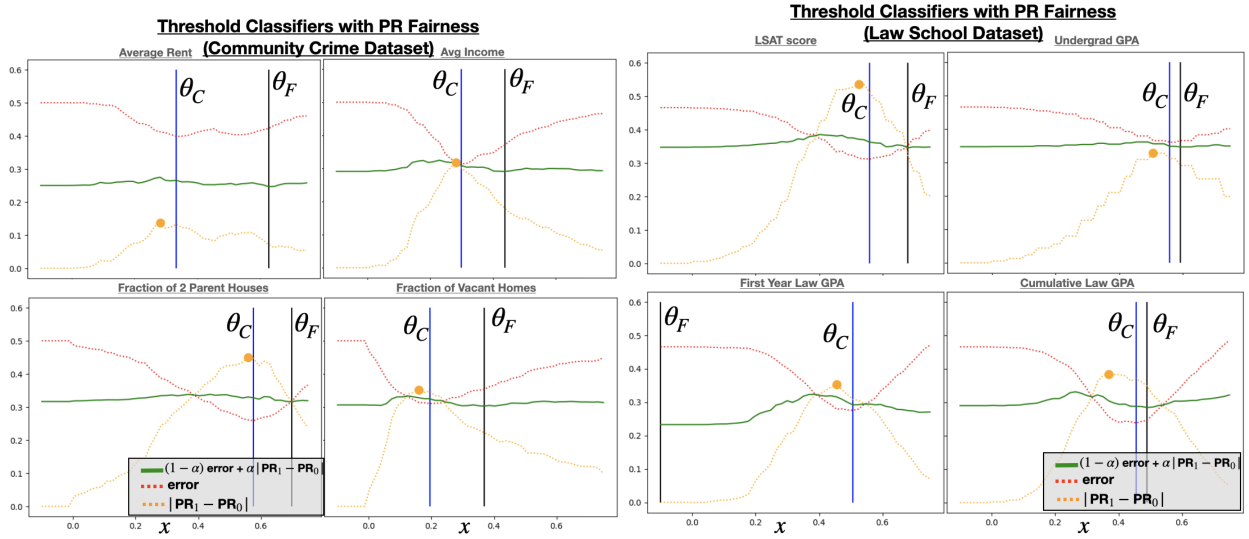

Figure A.2: Unfairness and error of threshold classifiers. Both error and unfairness are approximately unimodal w.r.t. threshold $\theta = x$. Thus error and unfairness are also unimodal w.r.t. the manipulation budget B for any manipulation cost function $c(x, x')$ which is monotone in $|x' - x|$. When this unimodality holds $\theta_C < \theta_F$ implies that strategic manipulation will lead to θ_C becoming more fair than θ_F. This fairness reversal is due to the fact that strategic manipulation amounts to lowering (shifting to the left) the threshold.

A.2.1 Individual Impact

Figures A.8-A.12 give a better sense of of how perceived impact overestimates realized impact in several specific cases. First we see that while the magnitudes of direct and realized impact can vary somewhat dramatically with the choice of dataset, baseline classifier, and fair learning scheme, there are is a nontrivial number of agents who are impacted across each dataset and pair of classifiers. Second, we generally see as θ increases and k decreases, realized

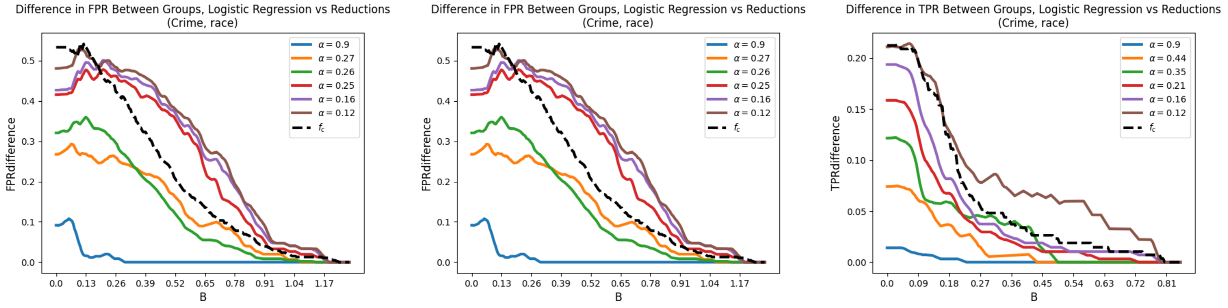

Figure A.3: Fairness reversals on the Community Crime dataset with Reductions Classifiers.

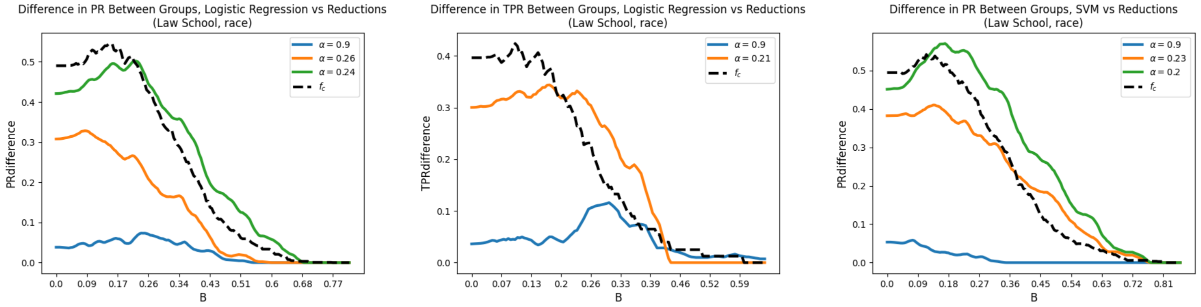

Figure A.4: Fairness reversals on the Law School dataset with Reductions Classifiers.

impact decrease, while perceived impact is, for the most part, independent of θ. Perceived impact can change with respect to θ in cases which agents tie for a resource. Third, for lower variance classifiers, such as Logistic Regression, we see the fraction of subsamples, an agent is impacted on, converge to 0 more rapidly. This is due to a randomness in resource allocation, due to model stability, meaning that the decisions of the baseline and fair classifiers are more stable across subsamples. Lastly, there are cases in which realized impact and perceived impact align, such as the the top two plots for the Law School dataset in Figure A.11, and cases in which perceived impact vastly overestimates realized impact, such as the plots for the Community Crime datasets in Figure A.9.

To look more closely at the ratio of impact between groups we present Figures A.13-A.16 which show the average fraction of each group which is impacted.

A general trend we see across most combinations of datasets, baseline classifiers, and fair learning schemes is that as resources become more scarce, the disadvantaged group is impacted at lower proportions, while for more abundant resources the disadvantaged group can suffer the majority of the impact across the population. Moreover, we see that there are people, from both of the sensitive groups, who are reliably impacted from the application of group fairness. More specifically, on every dataset, we see instances in which members of either group are impacted, even for large values of θ. This shows that there are identifiable individuals, regardless of group, who bare the burdens of group fairness at rates much higher than other members of the population. Moreover, we also see that the magnitude and distribution of impact varies significantly with respect to the choice of baseline classifier, fair learning scheme, and dataset.

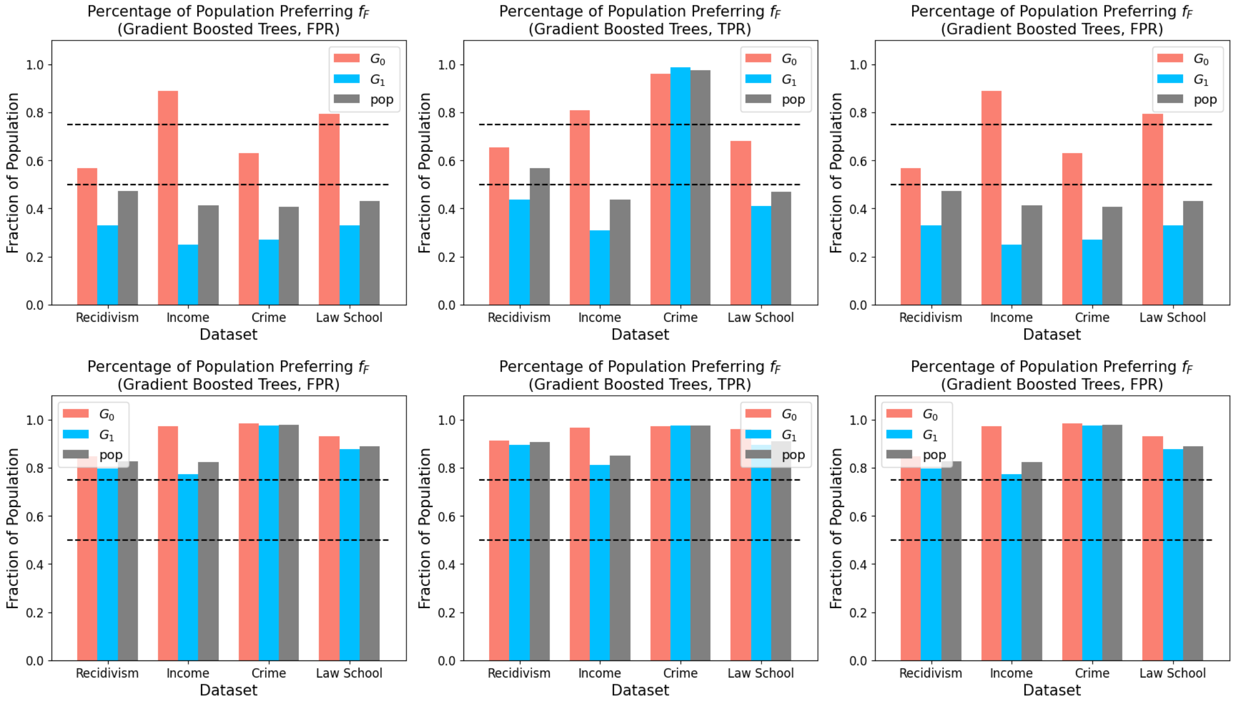

Figure A.5: Fraction of each population or group voting for f_F over f_C for randomized classifiers (top) and deterministic classifiers (bottom), when f_F is learned via the Reductions algorithm and each classifier uses Gradient Boosted Trees.

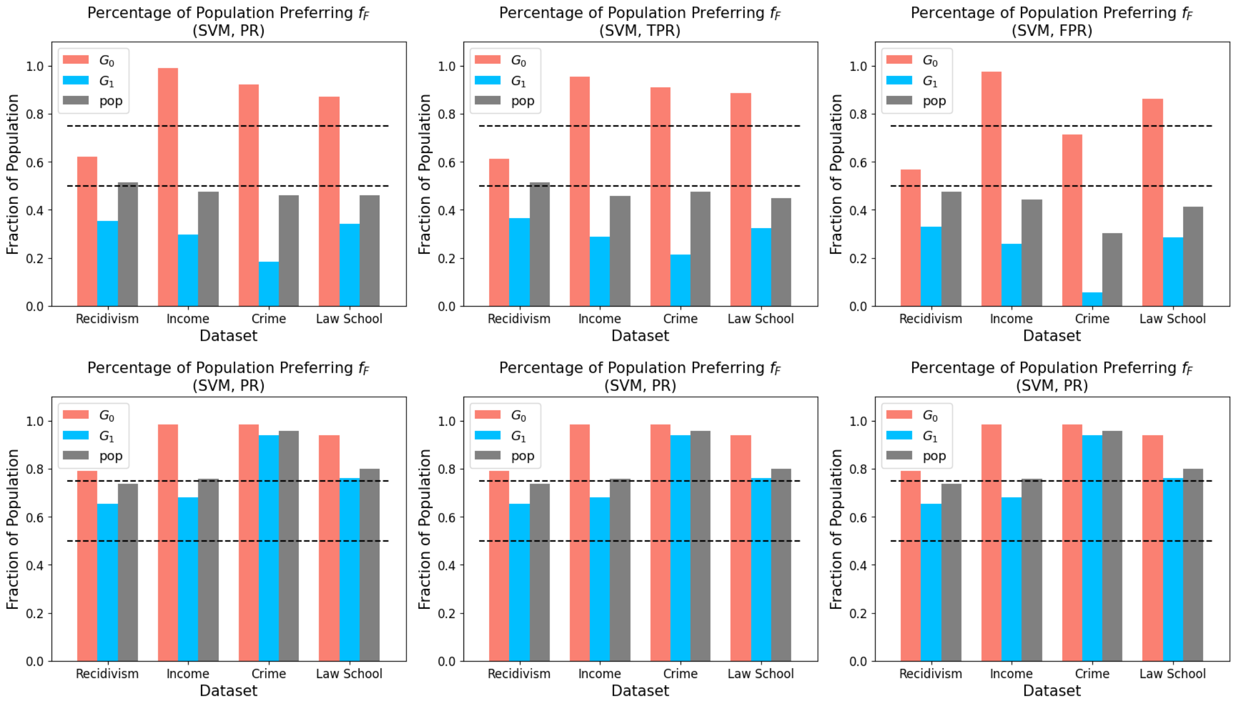

Figure A.6: Fraction of each population or group voting for f_F over f_C for randomized classifiers (top) and deterministic classifiers (bottom), when f_F is learned via the Reductions algorithm and each classifier uses SVMs.

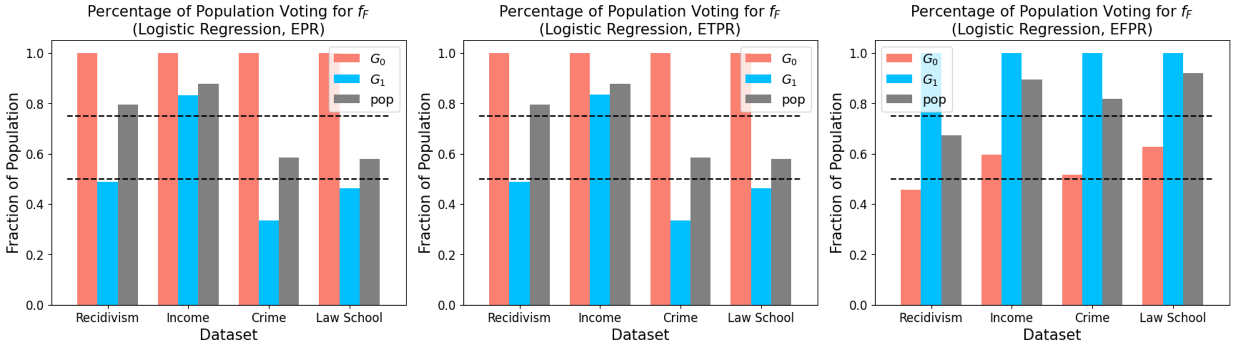

Figure A.7: Fraction of each population or group voting for f_F over f_C for randomized classifiers, when f_F is learned via the EqOdds algorithm and each classifier uses Logistic Regression. Due to the way in which EqOdds achieves fairness, the entirety of one group will always prefer f_F, since $f_F = f_C$ on that group.

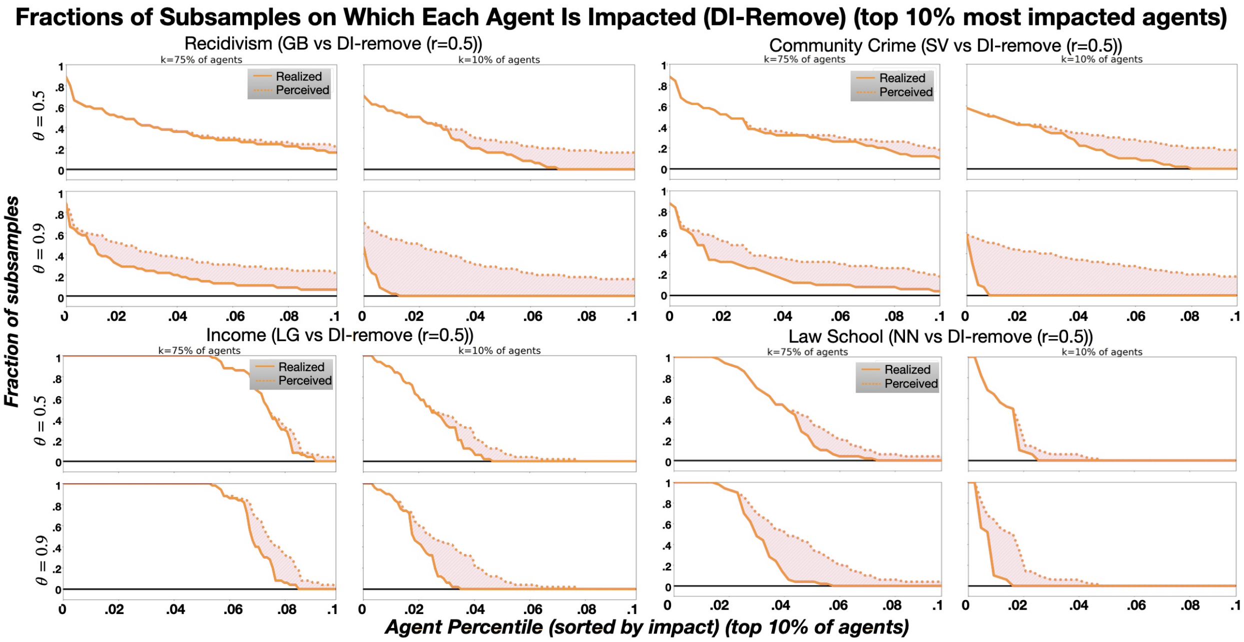

Figure A.8: Comparison of realized impact (solid) and perceived impact (dotted) when switching from a baseline classifier to DI-Remove. The shaded region gives a visual interpretation of how often, and to what degree, perceived impact overestimates realized impact. Only the top 10% of impacted agents are shown, realized and perceived impact are sorted independently.

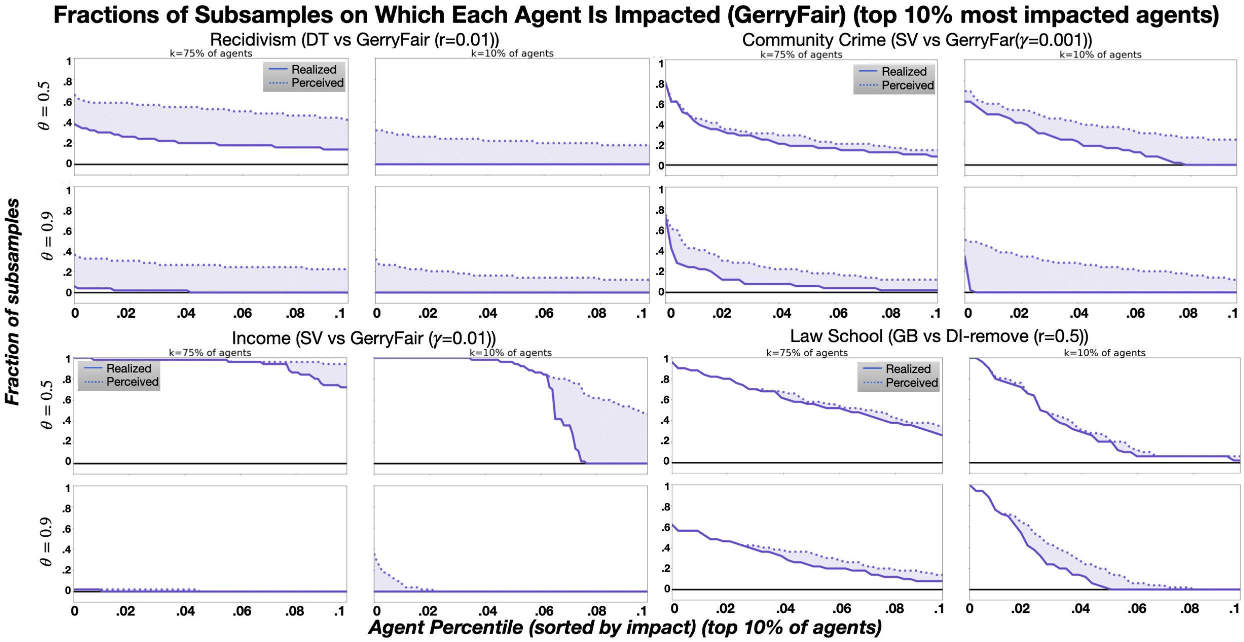

Figure A.9: Comparison of realized impact (solid) and perceived impact (dotted) when switching from a baseline classifier to GerryFair. The shaded region gives a visual interpretation of how often, and to what degree, perceived impact overestimates realized impact. Only the top 10% of impacted agents are shown, realized and perceived impact are sorted independently.

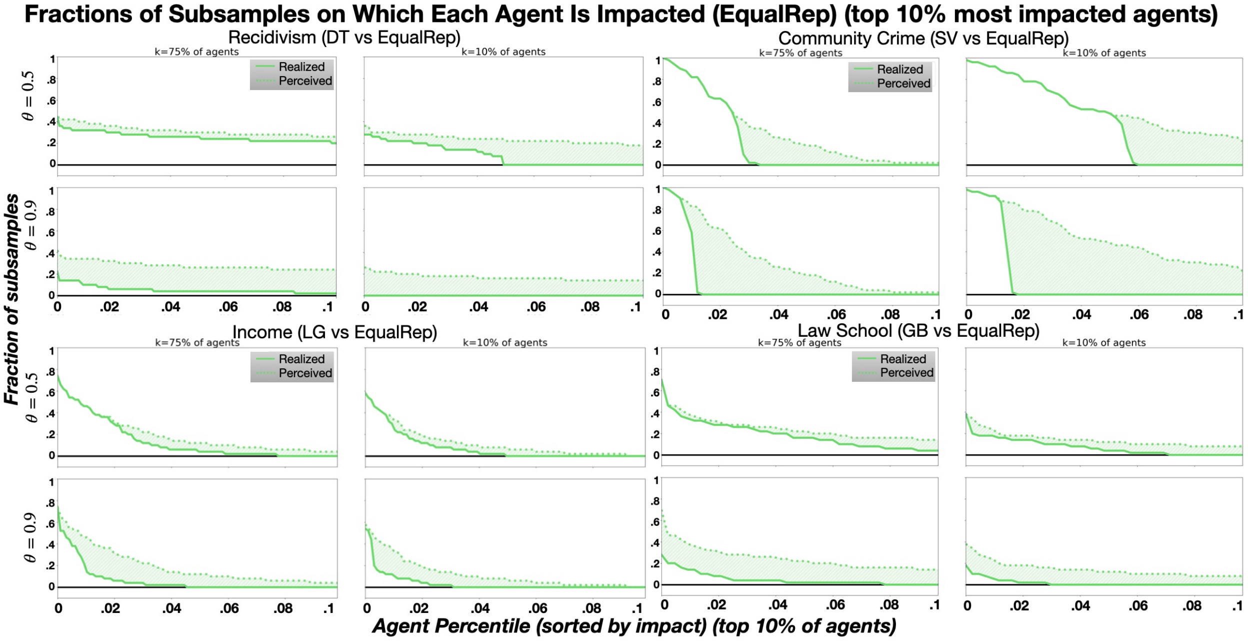

Figure A.10: Comparison of realized impact (solid) and perceived impact (dotted) when switching from a baseline classifier to EqualRep. The shaded region gives a visual interpretation of how often, and to what degree, perceived impact overestimates realized impact. Only the top 10% of impacted agents are shown,realized and perceived impact are sorted independently.

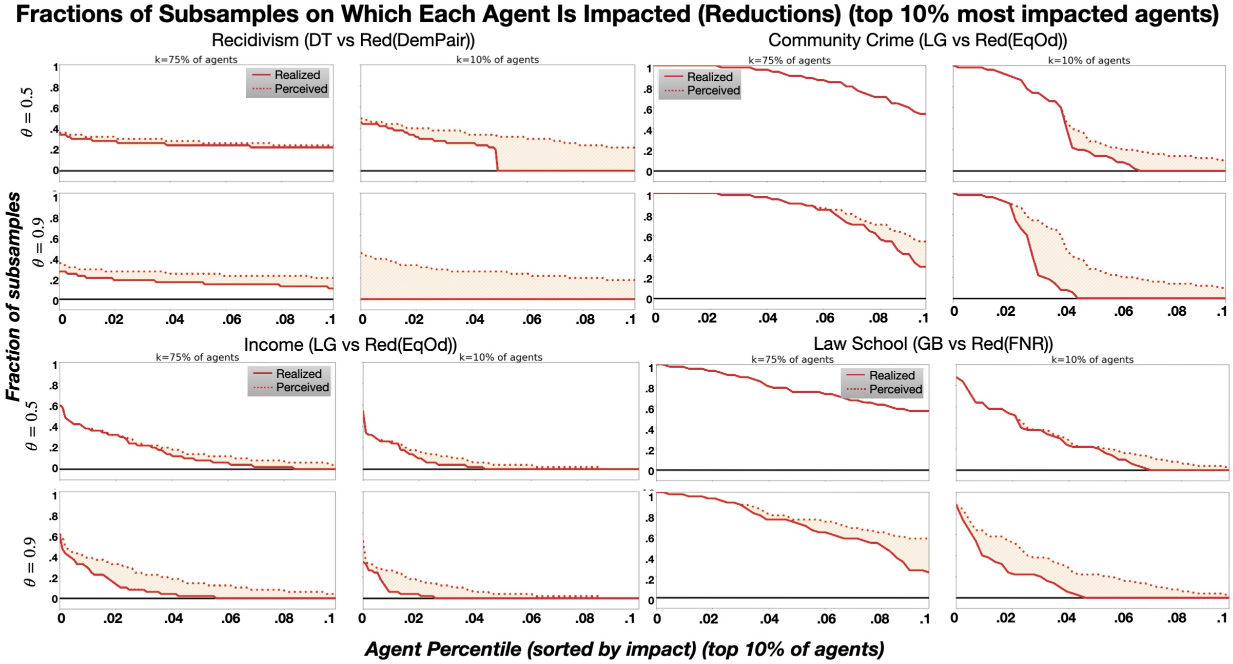

Figure A.11: Comparison of realized impact (solid) and perceived impact (dotted) when switching from SVM to Reduction wit FNR fairness. The shaded region gives a visual interpretation of how often, and to what degree, perceived impact overestimates realized impact. Only the top 10% of impacted agents are shown, realized and perceived impact are sorted independently.

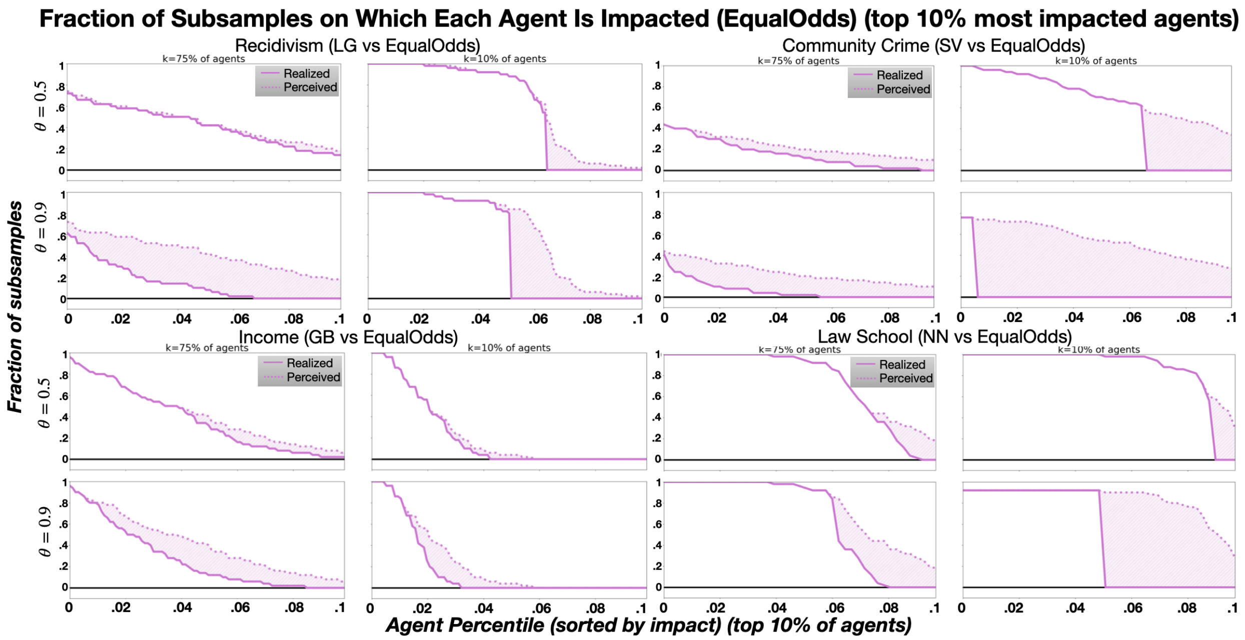

Figure A.12: Comparison of realized impact (solid) and perceived impact (dotted) when switching from a baseline classifier to EqualizeOdds with FNR relaxation. The shaded region gives a visual interpretation of how often, and to what degree, perceived impact overestimates realized impact. Only the top 10% of impacted agents are shown, realized and perceived impact are sorted independently..

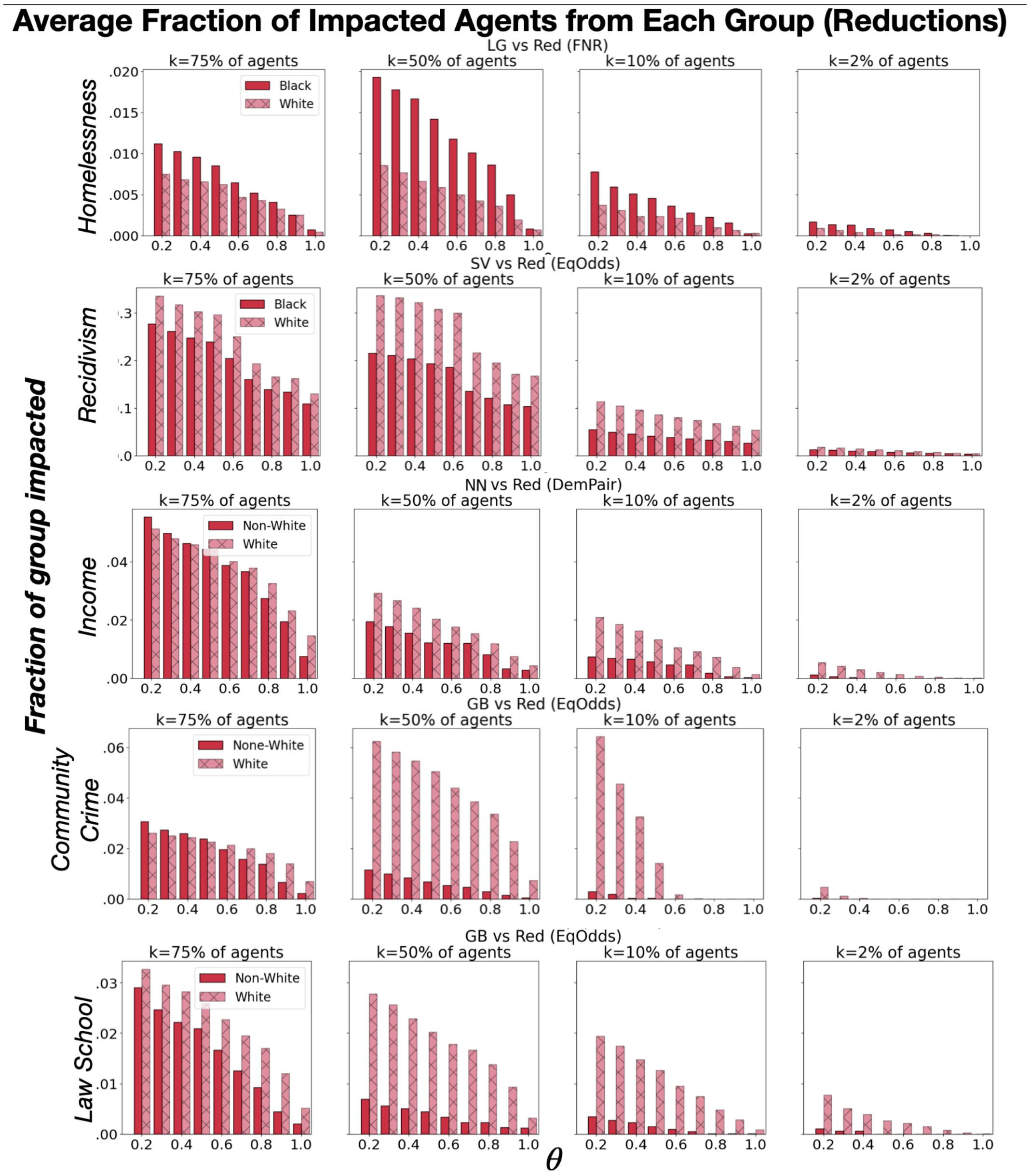

Figure A.13: Average fraction of each group impacted (realized), when switching from a baseline classifier to Reductions.

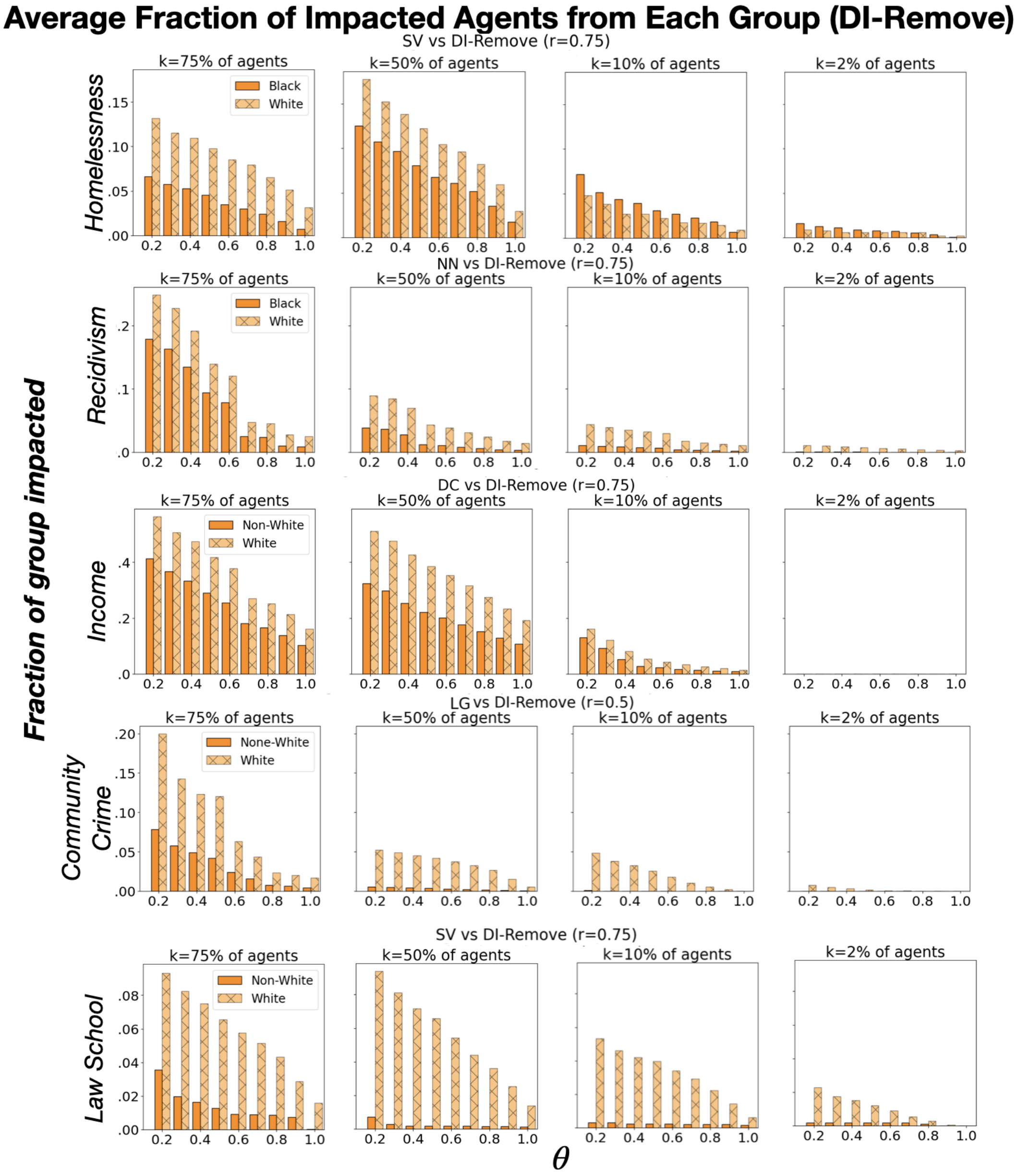

Figure A.14: Average fraction of each group impacted (realized), when switching from a baseline classifier to DI-Remove.

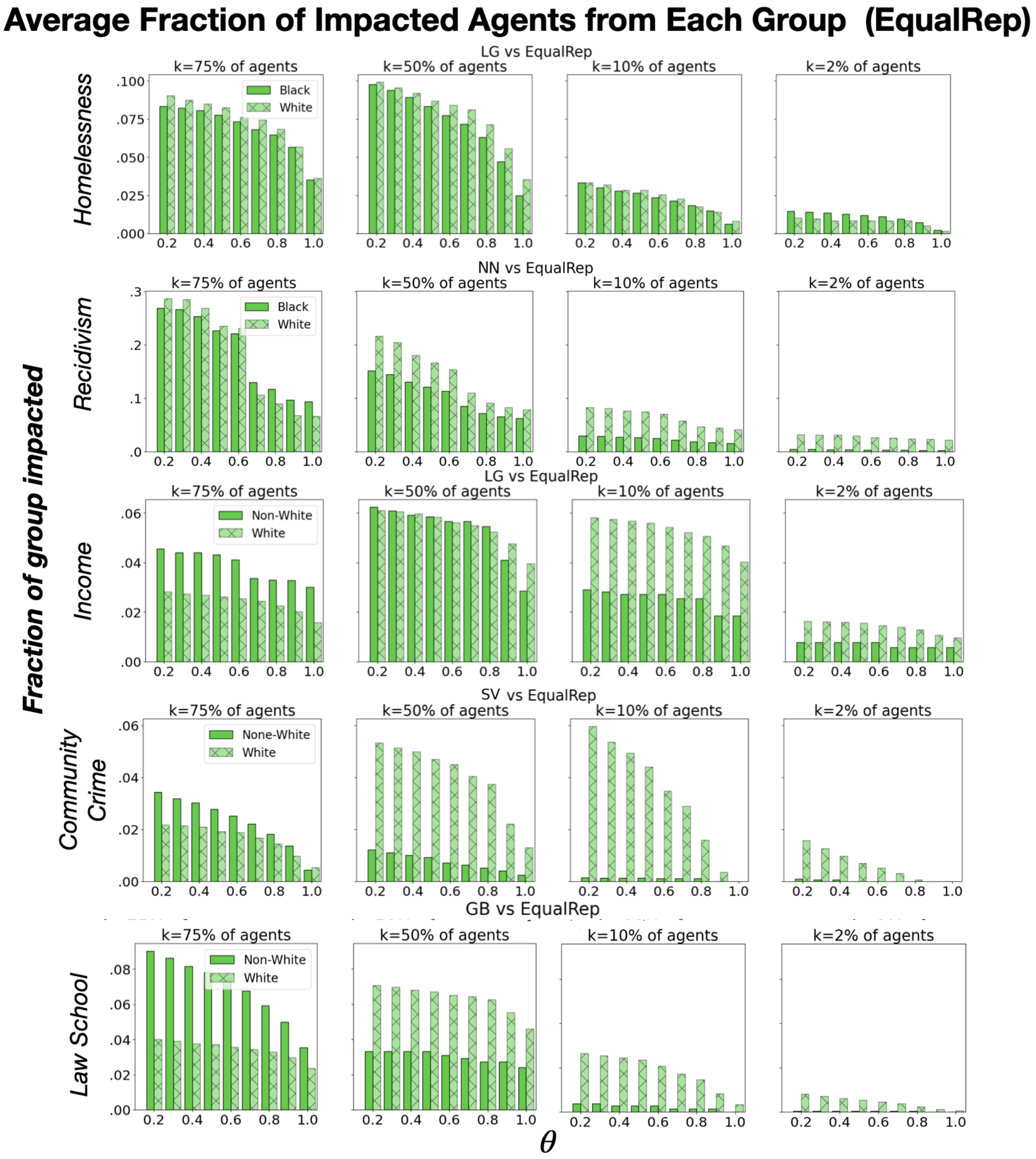

Figure A.15: Average fraction of each group impacted (realized), when switching from a baseline classifier to EqualRep.

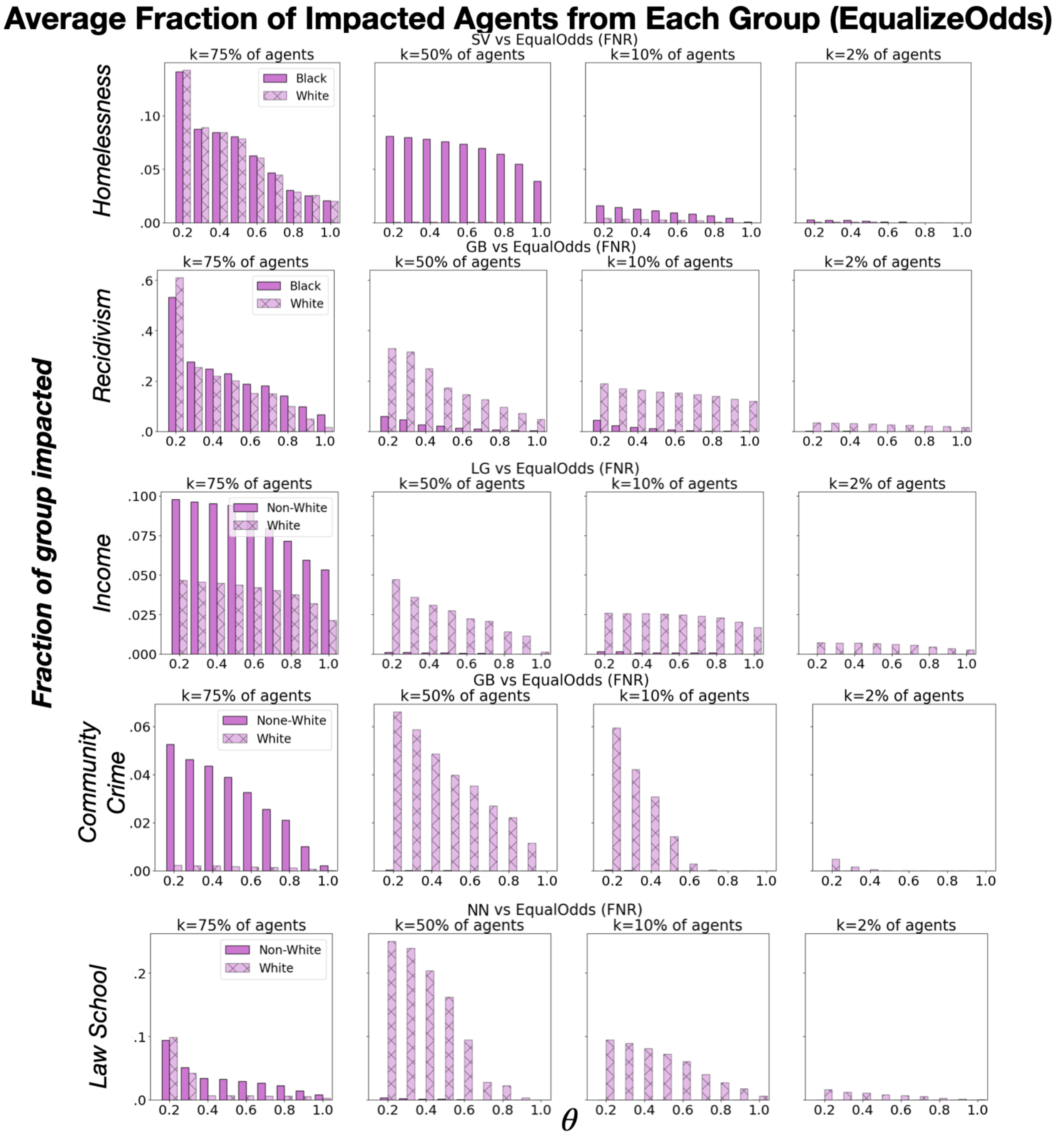

Figure A.16: Average fraction of each group impacted (realized), when switching from a baseline classifier to EqualizeOdds.

Appendix B

B.1 Aduiting and Recourse

Despite the ease of executing and defining the optimal audit policy, computing the expected number of agents which will prefer recourse is NP-hard.

Theorem B.1.1. *Let $U_{a,R}(\mathbf{x}) = u_a(\mathbf{x}_R) - c_R(\mathbf{x}, \mathbf{x}_R)$ and $U_{a,M}(\mathbf{x}) = u_a(\mathbf{x}) - c_A(\mathbf{x}, \mathbf{x}_M; g)$, then the objective value of a recourse maximizing principal,*

$$\begin{aligned} R^* = \max_{\alpha}\ & \mathbb{P}_{\mathbf{X}}\big(U_{a,R}(\mathbf{x}) \geq U_{a,M}(\mathbf{x})\big) \\ s.t.\ & \mathbb{E}_{\alpha}\Big[\sum_{\mathbf{z}' \in g(\mathbf{X})} \alpha(\mathbf{z}'; g(\mathbf{X}))|g(\mathbf{X})\Big] \leq B \quad \forall\ g(\mathbf{X}) \end{aligned}$$

is NP-hard to compute, even if the principal is given the optimal α, the classifier $\hat{y}$ is linear, the distribution $\mathcal{D}$ over agent features is uniform, and there are only $n = 2$ agents.

Proof. The proof is straightforward and follows directly from the hardness of computing the volume of a hypercube intersected by a hyperplane (HIH). An instance of HIH is given by a dimension-d hypercube $H = [0,1]^d$, and a hyperplane defined by the weight vector $\mathbf{w}$ and the bias θ. The objective is compute the volume of the region $S = \{\mathbf{x} \in H : \mathbf{w} \cdot \mathbf{x} \geq \theta\}$. To construct an instance of our problem, we define $\mathcal{X} = H$, $\hat{y}(\mathbf{x}) = \mathbb{I}[\mathbf{x} = \mathbf{1}]$, $B = 1$, $n = 2$, $C = 1$, and the distribution $\mathcal{D}$ over $\mathcal{X}$ to be uniform. Lastly, let $u_a(\mathbf{x}) = 1$ for all $\mathbf{x}$ and,

$$c_R(\mathbf{x}, \mathbf{1}) = \begin{cases} 0 & \text{if } \ \mathbf{w} \cdot \mathbf{x} \geq \theta \\ 2 & \text{if } \ \mathbf{w} \cdot \mathbf{x} < \theta \end{cases}$$

Under this construction $\mathbb{P}\big(\hat{y}(\mathbf{x}) = 1\big) = 0$. When only one agent reports $\mathbf{z}' = \mathbf{1}$ that agent is audited with probability 1, and when both agents report $\mathbf{z}' = \mathbf{1}$ the agents are audited with probability 0.5.

When an agent has $\mathbf{w} \cdot \mathbf{x} \geq \theta$ it is always optimal to choose recourse to $\mathbf{z} = \mathbf{1}$ since the cost of recourse is 0, while the cost of manipulation is at least 0.5. When an agent has $\mathbf{w} \cdot \mathbf{x} < \theta$ it is always optimal to manipulate since the cost of manipulation is upper-bounded by 1, while the cost of recourse is 2. Thus the probability that an agent chooses recourse is $\mathbb{P}\big(\mathbf{w} \cdot \mathbf{x} \geq \theta\big)$, i.e. the volume of S. □

B.2 Postprocessing for Popularity

Theorem (8.6.5)**.** *Postprocessing to achieve γ-popularity β-fairness with k-QLS (i.e., solving Program 8.23) is strongly NP-hard when models are randomized, $\mathcal{U}$ is derived from an additive fairness metric, and the number of quantiles k is determined by the input.*

Proof of Theorem 8.6.5. We reduce from the NP-hard problem exact m knapsack (EmKP), which is strongly NP-hard when coefficients are rational, which consists of n items, each with weight and value $w_i, v_i \in \mathbb{Q}_{\geq 0}$, a capacity $W \in \mathbb{Q}_{\geq 0}$, and a target m. The objective is to select exactly m items such that total value is maximized and the weight limit is not exceeded. To transform an instance of EmKP into an instance of k-QLS postprocessing, we map each item to an interval where the item weight corresponds to unfairness, item value corresponds to loss, and popularity is achieved when exactly m intervals have nonzero values of $p_\ell^{(g)}$. Specifically, for each item i create two agents i_0, i_1 such that for agent i_0, $g_{i_0} = y_{i_0} = 0$, and for i_1, $g_{i_1} = y_{i_1} = 1$. For the conventional and fair score function $h_C = \mathbb{E}[f_C], h_F = \mathbb{E}[f_F]$, let

$$h_C(\mathbf{x}_{i_0}) = \frac{w_i + \max_{j \in [n]}(w_j)}{2 \max_{j \in [n]}(w_j)} \quad \text{and} \quad h_F(\mathbf{x}_{i_0}) = \frac{v_i - 3W\big(1 + 2h_C(\mathbf{x}_{i_0})\big) \max_{j \in [n]}(v_j)}{4W h_C(\mathbf{x}_{i_0}) \max_{j \in [n]}(v_j)} \tag{B.1}$$

and,

$$h_C(\mathbf{x}_{i_1}) = h_F(\mathbf{x}_{i_1}) = 1.$$

In Equation B.1 note that $1/2 \leq h_C(\mathbf{x}_{i_0}) \leq 1$ and as such $0 \leq h_F(\mathbf{x}_{i_0}) \leq 1$. The particular values of both variables is selected to ensure that both h_C and h_F correspond to valid probabilities, and so that the loss and fairness constraint cancel out to yield a weight constraint over w_i and a maximize over v_i. Let the efficacy costs be defined as $c_{0,1}^{(0)} = c_{0,1}^{(1)} = 1$ and all others are 0, i.e. false positive fairness. Lastly let the popularity coefficient be $\gamma = \frac{n+m}{2n}$, maximum unfairness be $\beta = \frac{W}{2\max_{j\in[n]}(w_j)} + \frac{m}{2}$, the number of intervals be $k = 2n$, and the regularization coefficient be $\lambda = {}^1\!/_2$. Note that each of the k intervals then contains exactly one agent.

The key idea is that the construction of the groups, and choice of fairness definition, causes any optimal solution to positively classify all agents in G_1 since $g_j = y_j = 1$ for all $j \in G_1$. Doing so yields 0 loss on G_1 and makes no contribution to unfairness (since fairness is defined by FPR). Moreover, ignoring popularity, any optimal solution will negatively classify all agents in G_0 since $g_j = y_j = 0$ for all $j \in G_0$ and doing so yields 0 loss on G_0 and makes no false positive predictions. When adding the popularity constraint, i.e. $n + m$ of the $2n$ agents must have a an expected outcome under h_F which is at least as large as the expected outcome under h_C, the decisions on G_1 will remain invariant, but an optimal solution will select the lowest possible number of agents in G_0 (namely m) minimally increasing loss and not violating unfairness, and classify those agents positively with probability $h_C(\mathbf{x}_j)$ (i.e., their expected outcome under the conventional classifier). By the construction of the h_C and h_F in Equation B.1, these m agents will correspond to most profitable m items which do not exceed the weight limit.

To see why this is the case we first consider the loss term on each agent j in G_0 when that agent has expected outcome $p_j^{(0)}$,

$$\begin{aligned} & p_j^{(0)}(1 - y_j) + (1 - p_j^{(0)})y_j + \lambda(h_F(\mathbf{x}_j) - p_j^{(0)})^2 \\ =& p_j^{(0)}\big(1 + {}^1\!/_2 p_j^{(0)} - h_F(\mathbf{x}_j)\big) + {}^1\!/_2 h_F(\mathbf{x}_j)^2 \end{aligned}$$

since $0 \le h_F(\mathbf{x}_j) \le 1$, this term is monotonically increasing in $p_j^{(0)}$ and is minimized at $p_j^{(0)} = 0$. Thus without consideration of popularity or unfairness, the optimal solution is to set $p_j^{(0)} = 0$ for all $j \in G_0$. Moreover, by construction of the fairness cost coefficients $c_{0,1}^{(0)} = c_{0,1}^{(1)} = 1$, the fairness constraint can be written as

$$\begin{aligned} & \mathcal{U}(f_p, \mathcal{X}, Y, G) \le \beta \\ \iff & -\beta \le \sum_{i\in G_1} p_i^{(1)} c_{y_i,1}^{(1)} + (1 - p_i^{(1)}) c_{y_i,0}^{(1)} - \sum_{j\in G_0} p_j^{(0)} c_{y_j,1}^{(0)} + (1 - p_j^{(0)}) c_{y_j,0}^{(0)} \le \beta \\ \iff & \sum_{j\in G_0} p_j^{(0)} c_{0,1}^{(0)} \le \beta \end{aligned}$$

since $c_{0,1}^{(0)} = 1$, the left-hand side of the inequality is monotonically increasing in each $p_j^{(0)}$. Therefore, the fairness constraint adds no incentive to increase any $p_j^{(0)}$ on group 0, and thus only the popularity constraint will force $p_j^{(0)} > 0$ for some j.

Since $\gamma 2n = \frac{m+n}{2n} 2n = m + n$ number of agents need to prefer f_P to f_C (i.e., need $p_i^{(g)} \geq h_C(\mathbf{x}_i)$), and all n of the agents in G_1 trivially prefers f_P, the popularity constraint is satisfied only when m agents from G_0 prefer f_P.

Note that since both the unfairness term

$$\sum_{j \in G_0} p_j^{(0)} c_{0,1}^{(0)}$$

and the loss term

$$\sum_{j \in G_0} p_j^{(0)} \big(1 + {}^1\!/\!{}_2 p_j^{(0)} - h_F(\mathbf{x}_j)\big) + {}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2$$

corresponding to G_0 are both monotonically increasing in each $p_j^{(0)}$, the optimal solution is to set exactly m of the n variables $p_j^{(0)}$ to $h_C(\mathbf{x}_j)$ (i.e. the lowest possible value such that agent j prefers f_P to f_C). Let $\alpha_j \in \{0, 1\}$ correspond to an indicator that $p_j^{(0)} = h_C(\mathbf{x}_j)$, then k-QLS is equivalent to,

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \sum_{j \in G_0} \alpha_j \Big(& h_C(\mathbf{x}_j)\big(1 + {}^1\!/\!{}_2 h_C(\mathbf{x}_j) - h_F(\mathbf{x}_j)\big) + {}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2 \Big) && \text{(B.2)} \\ & + (1 - \alpha_j)({}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2) \\ \text{s.t.} \quad & \sum_{j \in G_0} \alpha_j h_C(\mathbf{x}_j) \leq \beta && \text{(B.3)} \\ & \sum_{j \in G_0} \alpha_j = m. && \text{(B.4)} \end{aligned}$$

Simplifying the objective and substituting the expressions for $h_C(\mathbf{x}_j)$ and $h_F(\mathbf{x}_j)$ yields

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \sum_{j\in G_0} \alpha_j \Big(h_C(\mathbf{x}_j)\big(1 + {}^1\!/\!{}_2 h_C(\mathbf{x}_j) - h_F(\mathbf{x}_j)) + {}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2 \Big) + (1-\alpha_j)({}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2) \\
\iff &\min_{\boldsymbol{\alpha}} \sum_{j\in G_0} \alpha_j \Big(h_C(\mathbf{x}_j)\big(1 + {}^1\!/\!{}_2 h_C(\mathbf{x}_j) - h_F(\mathbf{x}_j)) \Big) + {}^1\!/\!{}_2 h_F(\mathbf{x}_j)^2 \\
\iff &\min_{\boldsymbol{\alpha}} \sum_{j\in G_0} \alpha_j \Big({}^3\!/\!{}_4 - \frac{v_j}{4W \max_{i\in G_0}(v_i)}) \\
\iff &\max_{\boldsymbol{\alpha}} \sum_{j\in G_0} \alpha_j \frac{v_j}{4W \max_{i\in G_0}(v_i)} - \sum_{j\in G_0} \alpha_j {}^3\!/\!{}_4 \\
\iff &\max_{\boldsymbol{\alpha}} \sum_{j\in G_0} \alpha_j \frac{v_j}{4W \max_{i\in G_0}(v_i)}
\end{aligned}$$

where the final line stems from the fact that $\sum_{j\in G_0} = m$ and is thus a constant term, not affecting the optimization. Moreover, note that the denominator $4W \max_{i\in G_0}(v_i)$ is also constant, thus minimizing (B.2) is equivalent to maximizing the value of the knapsack.

Lastly, we need only show that the fairness term is equivalent to the original capacity constraint. The fairness constraint can be written then as

$$\begin{aligned}
&\sum_{j\in G_0} \alpha_j h_C(\mathbf{x}_j) \leq \beta \\
\iff &\sum_{j\in G_0} \alpha_j \frac{w_i + \max_{j\in[n]}(w_j)}{2\max_{j\in[n]}(w_j)} \leq \frac{W}{2\max_{j\in[n]}(w_j)} + \frac{m}{2} \\
\iff &\sum_{j\in G_0} \alpha_j \big(w_i + \max_{j\in[n]}(w_j)\big) \leq W + m \max_{j\in[n]}(w_j) \\
\iff &\sum_{j\in G_0} \alpha_j w_i \leq W
\end{aligned}$$

where the last line is again due to $\sum_{j\in G_0} \alpha_j = m$. Thus the fairness constraint is satisfied if and only if the original capacity constraint is satisfied. Thus, any solution to k-QLS which successfully minimizes loss such that unfairness is not violated and at least m agents from G_0 prefer f_P can be used as an optimal solution to the original EmKP problem by simply selecting all items j which correspond to nonzero values of $p_j^{(0)}$. Since EmKP is strongly NP-hard, so is k-QLS postprocessing on randomized classifiers. □

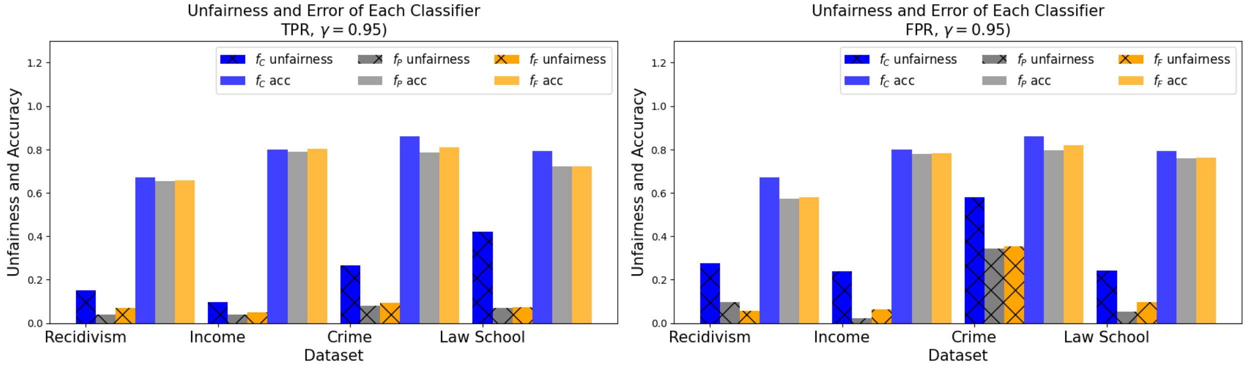

Figure B.1: Model performance and unfairness on test data (3-fold average) for deterministic models with $\gamma = 0.95$. The conventional classifier f_C, fair classifier f_F learned via the reductions algorithm, and the fair popular classifier f_P learned via our postprocessing techniques k-QLS, each using Gradient Boosted Trees.

Experiments

Figures B.1, B.2 show the accuracy and unfairness of the conventional model f_C, β-fair model f_F, and the γ-popular β-fair model f_P (learned via k-QLS) when model outcomes are deterministic. Figures B.3, B.3 show model AUC and unfairness of these model in the case of randomized classifiers. Figure B.5, B.6 show model AUC and unfairness when f_P is learned via the DOS algorithm in the case of randomized classifiers. Similar to the case of Logistic Regression, we see that f_P can achieve γ-popularity, and β-fairness, for relatively large levels of γ with minimal degradation to model performance.

Figure B.7 demonstrates the increased false positive rate errors made by γ-popular classifiers as γ increases.

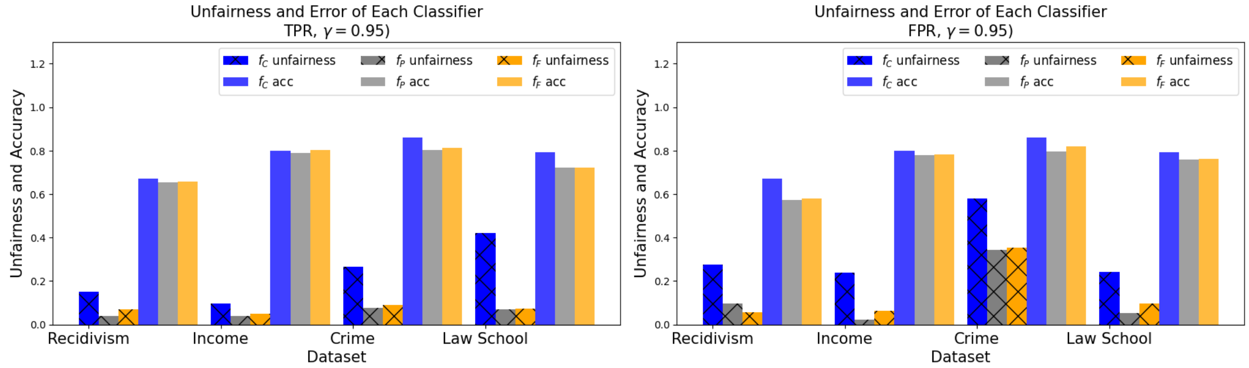

Figure B.2: Model performance and unfairness on test data (3-fold average) for deterministic models with $\gamma = 0.95$. The conventional classifier f_C, fair classifier f_F learned via the reductions algorithm, and the fair popular classifier f_P learned via our postprocessing techniques k-QLS, each using Support Vector Machines.

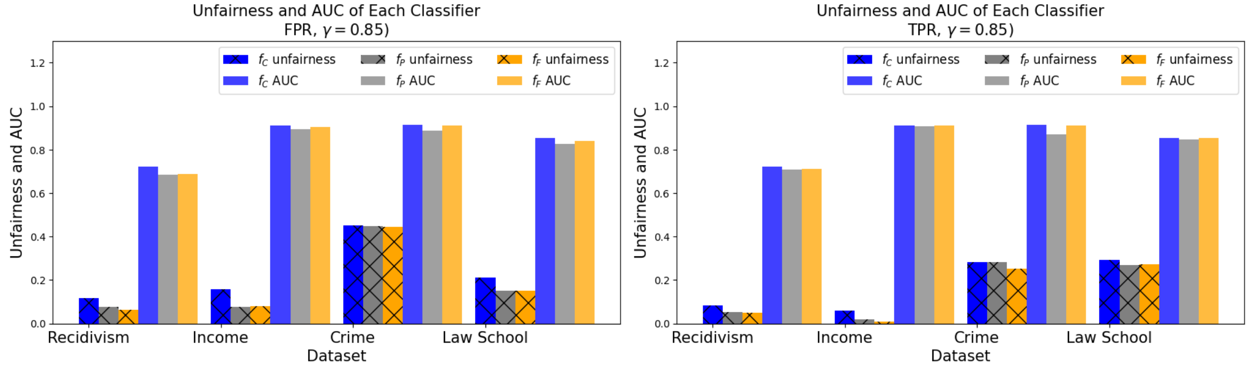

Figure B.3: Model performance and unfairness on test data (3-fold average) for randomized models with $\gamma = 0.85$. The conventional classifier f_C, fair classifier f_F learned via the reductions algorithm, and the fair popular classifier f_P learned via our postprocessing techniques k-QLS, each using Gradient Boosted Trees.

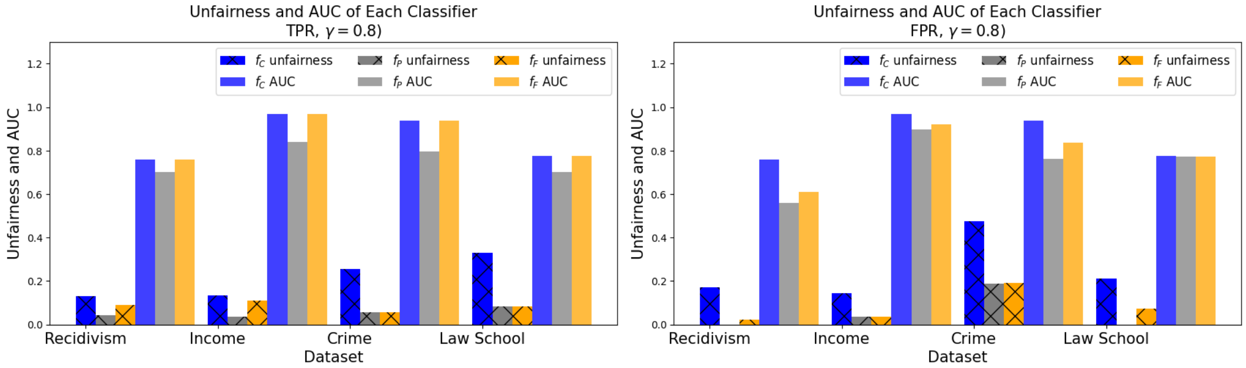

Figure B.4: Model performance and unfairness on test data (3-fold average) for randomized models with $\gamma = 0.8$. The conventional classifier f_C, fair classifier f_F learned via the KDE algorithm, and the fair popular classifier f_P learned via our postprocessing techniques k-QLS, each using Support Vector Machines.

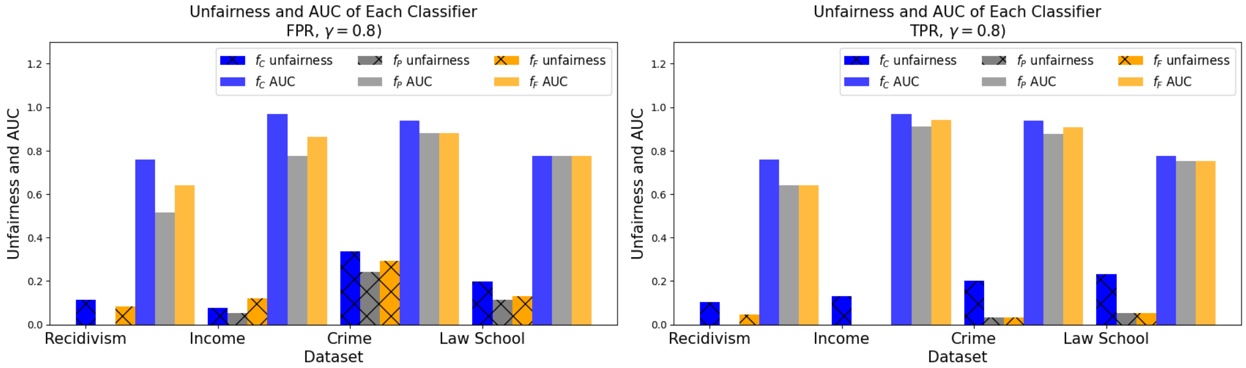

Figure B.5: Model performance and unfairness on test data (3-fold average) for randomized models with $\gamma = 0.8$. The conventional classifier f_C, fair classifier f_F learned via the reductions algorithm, and the fair popular classifier f_P learned via our postprocessing techniques DOS, each using Gradient Boosted Trees.

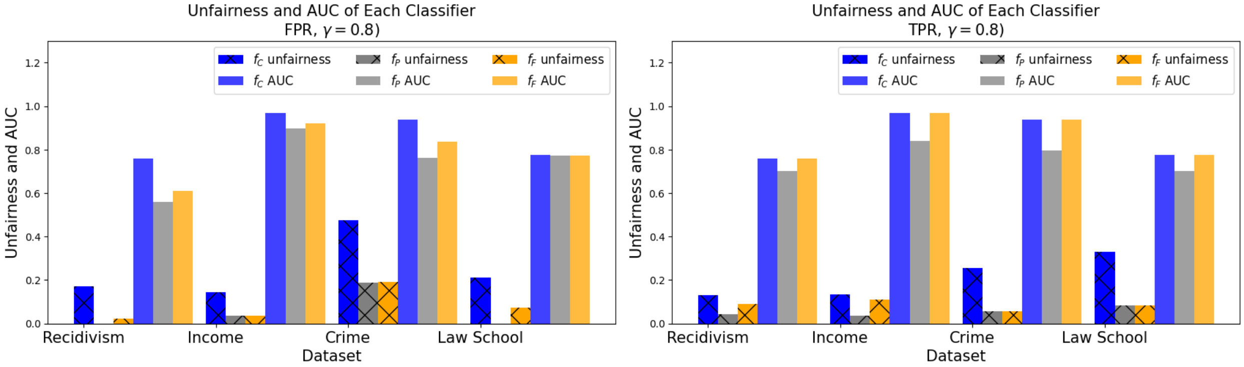

Figure B.6: Model performance and unfairness on test data (3-fold average) for randomized models with $\gamma = 0.8$. The conventional classifier f_C, fair classifier f_F learned via the KDE algorithm, and the fair popular classifier f_P learned via our postprocessing techniques DOS, each using Support Vector Machines.

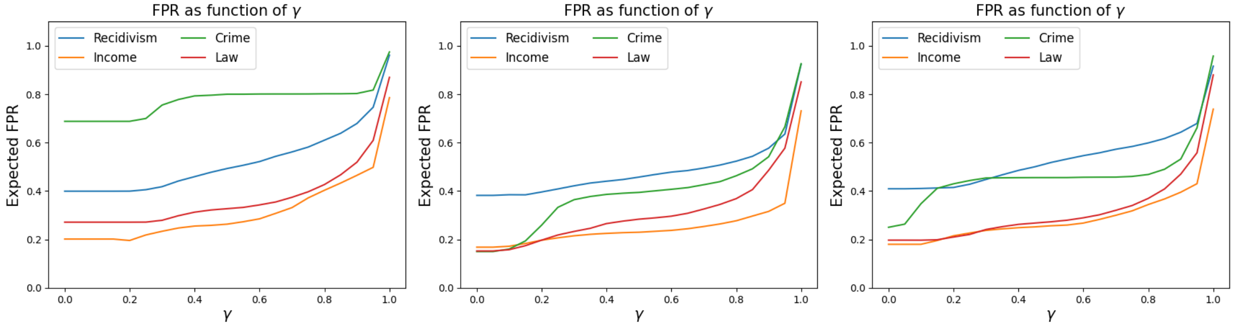

Figure B.7: Expected False Positive Rate (FPR) of k-QLS, on randomized classifiers for PR-fairness (left) TPR-fairness (center) and FPR-fairness (right),as a function of γ (Support Vector Machines).