Discrimination in Algorithmic Decision Making: From Principles to Measures and Mechanisms

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

by

Muhammad Bilal Zafar

Saarbrücken
February, 2019

Date of Colloquium:	February 04, 2019
Dean of Faculty:	Univ.-Prof. Dr. Sebastian Hack
Chair of the Committee:	Prof. Dr. Gerhard Weikum
Reporters	
First Reviewer:	Prof. Dr. Krishna P. Gummadi
Second Reviewer:	Dr. Manuel Gomez Rodriguez
Third Reviewer:	Prof. Dr. Sharad Goel, Ph.D.
Fourth Reviewer:	Prof. Dr. Paul Resnick, Ph.D.
Academic Assistant:	Dr. Rishiraj Saha Roy

Abstract

The rise of algorithmic decision making in a variety of applications has also raised concerns about its potential for discrimination against certain social groups. However, incorporating nondiscrimination goals into the design of algorithmic decision making systems (or, classifiers) has proven to be quite challenging. These challenges arise mainly due to the computational complexities involved in the process, and the inadequacy of existing measures to computationally capture discrimination in various situations. The goal of this thesis is to tackle these problems.

First, with the aim of incorporating existing measures of discrimination (namely, disparate treatment and disparate impact) into the design of well-known classifiers, we introduce a mechanism of decision boundary covariance, that can be included in the formulation of any convex boundary-based classifier in the form of convex constraints. Second, we propose alternative measures of discrimination. Our first proposed measure, disparate mistreatment, is useful in situations when unbiased ground truth training data is available. The other two measures, preferred treatment and preferred impact, are useful in situations when feature and class distributions of different social groups are significantly different, and can additionally help reduce the cost of nondiscrimination (as compared to the existing measures). We also design mechanisms to incorporate these new measures into the design of convex boundary-based classifiers.

Kurzdarstellung

Die Vielzahl der Anwendungen, die Algorithmen immer stärker an Entscheidungsprozessen beteiligen, wächst stetig. Dadurch werden Bedenken über die potenzielle Diskriminierung bestimmter gesellschaftlicher Gruppen aufgeworfen. Die Aufnahme von Nichtdiskriminierungszielsetzungen bei der Gestaltung algorithmischer Entscheidungs- bzw. Klassifizierungssysteme hat sich jedoch als grosse Herausforderung herausgestellt. Zum einen sind die nötigen Berechnungen komplex und zum anderen sind die existierenden Metriken unzureichend, um Diskriminierung in bestimmten Situationen rechnerisch zu erfassen. Das Ziel dieser Arbeit ist es, diese Problematik anzugehen.

Als erstes stellen wir einen Decision Boundary-basierten Kovarianzmechanismus vor, der genutzt werden kann, um existierende Diskriminierungsmetriken (also Disparate Treatment und Disparate Impact) beim Entwurf von gängigen Klassifizierungsalgorithmen einzusetzen. Der Ansatz kann für jeden konvexen Boundary-basierten Klassifizierungsalgorithmus in Form konvexer Constraints formuliert werden. Als nächstes definieren wir neue Diskriminierungsmetriken. Unsere erste Metrik namens Disparate Mistreatment kommt in Situationen zum Einsatz, in denen die Referenzdaten nicht zugunsten einer sozialen Gruppe verzerrt sind. Die übrigen beiden Metriken namens Preferred Treatment und Preferred Impact sind für Situationen konzipiert, in denen die Feature- und Klassenverteilungen unterschiedlicher sozialer Gruppen stark voneinander abweichen. Sie können dabei helfen, die Kosten von Nichtdiskriminierung im Vergleich zu bestehenden Metriken zu reduzieren. Wir zeigen ebenfalls, wie diese neuen Metriken in konvexen Boundary-basierten Klassifizierungsalgorithmen genutzt werden können.

Publications

Parts of this thesis have appeared in the following publications.

- “From Parity to Preference-based Notions of Fairness in Classification”. M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi and A. Weller. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, December 2017.

- “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment”. M. B. Zafar, I. Valera, M. Gomez-Rodriguez and K. P. Gummadi. In *Proceedings of the 26th International World Wide Web Conference (WWW)*, Perth, Australia, April 2017.

- “Fairness Constraints: Mechanisms for Fair Classification”. M. B. Zafar, I. Valera, M. Gomez-Rodriguez and K. P. Gummadi. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, April 2017.

Additional publications while at MPI-SWS.

- “A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices ”. T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar. In *Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining (KDD)*, London, UK, August 2018.

- “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning”. N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi and A. Weller. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, February 2018.

- "Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media". J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi and K. Karahalios. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, Portland, OR, February 2017.

- "Listening to Whispers of Ripple: Linking Wallets and Deanonymizing Transactions in the Ripple Network". P. Moreno-Sanchez, M. B. Zafar and A. Kate. In *Proceedings on Privacy Enhancing Technologies (PoPETS)*, 2016.

- "Message Impartiality in Social Media Discussions". M. B. Zafar, K. P. Gummadi and C. Danescu-Niculescu-Mizil. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*, Cologne, Germany, May 2016.

- "On the Wisdom of Experts vs. Crowds: Discovering Trustworthy Topical News in Microblogs". M. B. Zafar, P. Bhattacharya, N. Ganguly, S. Ghosh and K. P. Gummadi. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, Portland, OR, February 2016.

- "Strength in Numbers: Robust Tamper Detection in Crowd Computations". B. Viswanath, M. A. Bashir, M. B. Zafar, S. Bouget, S. Guha, K. P. Gummadi, A. Kate and A. Mislove. In *Proceedings of the 3rd ACM Conference on Online Social Networks (COSN)*, Palo Alto, CA, October 2015.

- "Sampling Content from Online Social Networks: Comparing Random vs. Expert Sampling of the Twitter Stream". M. B. Zafar, P. Bhattacharya, N. Ganguly, K. P. Gummadi and S. Ghosh. In *ACM Transactions on the Web (TWEB)*, 2015.

- "Characterizing Information Diets of Social Media Users". J. Kulshrestha, M. B. Zafar, L. E. Noboa, K. P. Gummadi and S. Ghosh. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, Oxford, UK, May 2015.

- "Inferring User Interests in the Twitter Social Network". P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh and K. P. Gummadi. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys)*, Silicon Valley, CA, October 2014. (**Short paper**)

- "Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale". P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly and K. P. Gummadi. In *Proceedings of the 17th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, Baltimore, MD, February 2014.

- “On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream”. S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly and K. P. Gummadi. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, Burlingame, CA, October 2013.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Algorithmic decision making in social domains

Data-driven algorithmic decision making has been used in applications involving human subjects for several decades. For instance, credit scoring algorithms were being deployed in practice in as early as the 1950s (FICO, 2018a; Furletti, 2002), and parole risk assessment algorithms have been in use since the 1970s (Hoffman and Beck, 1974; Kehl and Kessler, 2017). However, with the advent of complex learning methods, and convenient accessibility of "big data", algorithmic decision making is permeating into an ever-increasing number of human-centric applications, where algorithms are used to assist, or sometimes even replace human decision makers. Some examples include job screening (Posse, 2016), healthcare (Bhardwaj et al., 2017), community safety (Perry, 2013), product personalization (Covington et al., 2016), online ad delivery (Graepel et al., 2010) and social benefits assignments (Niklas et al., 2015).

Algorithmic decision making has shown great promise in increasing the accuracy and scalability of the applications under consideration. For example, a recent study by Liu et al. (2017) shows that machine learning models can achieve a performance comparable to that of humans when detecting cancer metastases. Goel et al. (2016) show that in applications such as stop-question-and-frisk (Meares, 2014)—where pedestrians are stopped by police officers on the suspicion of possessing illegal weapons—algorithmic decision making can recover the majority of illegal weapons, while making much fewer stops (6%) as compared to human decision makers (that is, the police officers). Similarly, Kleinberg et al. (2018) found that when making bail decisions, algorithms can significantly reduce the crime rate (by 25%) while maintaining the same incarceration rate. Several other studies have also shown evidence that algorithms can help increase the performance of the task at hand in domains ranging from hiring (Kuncel et al., 2013, 2014) to education (Dickson, 2017).

Algorithmic decision making also presents potential for several additional advantages, such as, reducing the arbitrariness and implicit human biases in decision making. For example, while different human judges are known to grant different decisions to similar defendants (Dobbie et al., 2018; Kleinberg et al., 2018), algorithms can be easily designed to overcome this issue. Similarly, whereas human judgments can be potentially swayed (unintentionally) by various factors ranging from unconscious human biases (Badger, 2016; Tatum, 2017) to the hunger level of human judges (Danziger et al., 2011), the design of algorithmic decision making systems suggests that they can trivially avoid these problems.

1.2 Discrimination in algorithmic decision making systems

Despite its apparent advantages, algorithmic decision making has also caused concerns about potential discrimination against people with certain social traits (*e.g.*, gender, race), also referred to as *sensitive features*.

For example, Sweeney (2013) found that Google's AdSense platform was disproportionately associating predominantly African-American names as having arrest records, as compared to the predominantly White names. A recent analysis by ProPublica claimed that COMPAS, a recidivism risk assessment tool used in courts across several locations in the United States (US), was biased against African-American defendants (Angwin et al., 2016). An analysis by Bolukbasi et al. (2016) revealed that the *word2vec* word embeddings (Mikolov et al., 2013) used in a number of downstream tasks such as translation, web search and sentiment analysis, were biased along gender stereotypes present in the society. Similarly, a number of other instance have been reported where algorithms (unintentionally) discriminated against certain social groups (Buolamwini and Gebru, 2018; Fussell, 2017; Pachal, 2015).

In this context, there have been calls from governments (Muñoz et al., 2016; Podesta et al., 2014), regulatory authorities (FTC, 2016; Goodman and Flaxman, 2016), civil rights unions (Eidelman, 2017) and researchers (Barocas and Selbst, 2016; OŃeil, 2016; Pasquale, 2015) to tackle the potential discriminatory effects of algorithmic decision making. For example, a recent report by the US Federal Trade Commission (FTC, 2016) points out that data-driven algorithmic decision making can "create or reinforce existing disparities" or "create new justification for exclusion", and urged that "companies should assess the factors that go into an analytics model and balance the predictive value of the model with fairness considerations". Similarly, Recital 71 of the European General Data Protection Regulation (GDPR) that came into effect in May 2018, requires organizations handling personal data of European Union (EU) users to "prevent, inter alia, discriminatory effects on natural persons on the basis of” certain social traits such as sexual orientation and ethnic origin (Goodman and Flaxman, 2016; Goodman, 2016).

1.3 Challenges in tackling discrimination

While avoiding discrimination based on certain socially salient traits (*e.g.*, gender, race) is a legal **principle** in many countries (Altman, 2016; Civil Rights Act, 1964), eliminating discrimination from algorithmic decision outcomes poses a tough challenge. Two of the major reasons for this difficulty are:

I. Algorithmic decision making systems are typically designed to optimize for prediction accuracy while enabling efficient training. Efficient training here refers to finding the optimal algorithm parameters rapidly, and is a crucial property while learning from large training datasets (Bishop, 2006). Incorporating nondiscrimination **mechanisms** into these systems—*i.e.*, optimizing for prediction accuracy *under* nondiscrimination constraints—while simultaneously preserving efficient training, is often quite difficult.

II. While the nondiscrimination principle “enjoys impressive global consensus” (Altman, 2016), operationalizing this principle to **measure** discrimination (to eventually eliminate it) is a non-trivial task. Here, operationalization refers to the process of formalizing or *interpreting* a fuzzy concept so as to make it *measurable* for empirical observations (Lukyanenko et al., 2014). For example, what constitutes a discriminatory practice in one case might not do so in another. In fact, one widely accepted measure of discrimination (namely, disparate impact), is known to lead to “reverse discrimination” if applied out of context (Ricci, 2009).

1.4 Thesis contributions

This thesis tries to address the above challenges. Below, we discuss our research contributions towards this end.

I. Proposing mechanisms for existing nondiscrimination measures

Existing studies in discrimination-aware machine learning mostly quantify discrimination using two measures inspired by anti-discrimination legislation in various countries: ***disparate treatment*** and ***disparate impact*** (Barocas and Selbst, 2016). As we will discuss in detail in Section 2, while it is desirable to train decision making systems that are nondiscriminatory with respect to both the measures, doing so in practice is quite difficult due to computational complexities involved.

To overcome the computational issues in training nondiscriminatory classifiers, we propose a novel and intuitive mechanism of decision boundary covariance. This mechanism satisfies several desirable properties: (i) it can limit discrimination with respect to both disparate treatment and disparate impact; (ii) for a wide variety of convex boundary-based linear and non-linear classifiers (*e.g.*, logistic regression, SVM), it is convex and can be readily incorporated in their formulation without increasing their complexity, hence ensuring efficient learning; (iii) it allows for clear mechanisms to trade-off nondiscrimination and accuracy; and, (iv) it can be used to ensure nondiscrimination with respect to several sensitive features.

Experiments using both synthetic and real-world data show that our mechanism allows for a fine-grained control of the level of nondiscrimination, often at a small cost in terms of accuracy, and provides more flexibility than the state-of-the-art.

II. Proposing new measures of nondiscrimination (and designing mechanisms)

We also propose new measures of nondiscrimination that can avoid some shortcomings of the existing measures.

First, we argue that while the disparate impact measure of nondiscrimination might be quite intuitive in certain situations—*e.g.*, situations where the historical decisions in the training data are potentially biased (*i.e.*, groups of people with certain sensitive attributes may have historically received discriminatory treatment), its utility is somewhat limited in cases when the ground truth training labels are available. We then propose an alternative measure of nondiscrimination, ***disparate mistreatment***, which is useful in situations when the validity of historical decisions in the training data can be ascertained.

Next, we note that while existing measures of nondiscrimination in machine learning are based on parity (of treatment or impact), under some interpretations, a lack of parity might not necessarily constitute as discrimination. Specifically, drawing inspiration from the concepts of fair-divisions and envy-freeness in economics and game theory, we propose two additional measures of nondiscrimination: ***preferred treatment*** and ***preferred impact***. These measures are useful in situations when feature and class distributions of different groups subject to the decision making are significantly different. These measures are based on the idea that certain distributions of outcomes might be preferred by different groups even when the outcomes do not necessarily follow parity as specified by disparate treatment and disparate impact. We also show that these new measures can help reduce the cost of nondiscrimination.

We also extend our decision boundary covariance mechanism and incorporate the newly proposed nondiscrimination measures into the formulations of convex boundary-based classifiers, this time as convex-concave constraints. The resulting formulations can be solved efficiently using recent advances in convex-concave programming.

1.5 Thesis outline

The rest of this thesis is organized as follows:

- In Chapter 2, we provide background on discrimination in machine learning. Specifically, we discuss the concept of discrimination in the context of social sciences and law. We then describe how discrimination is measured in classification tasks.

- In Chapter 3, we design mechanisms to eliminate discrimination from classification outcomes, when it is measured using existing notions of disparate treatment and disparate impact.

- In Chapter 4, we propose a new measure of discrimination which we refer to as disparate mistreatment. We describe how disparate mistreatment can overcome some shortcomings of the existing measure of disparate impact. We also propose mechanisms to train classifiers without disparate mistreatment.

- In Chapter 5, we depart from the legal perspective of discrimination and introduce two new measures of discrimination: preferred treatment and preferred impact, which are inspired by ideas from economics and game theory. We then design mechanisms to train classifiers satisfying these two new (non)discrimination criteria.

- In Chapter 6, we review literature from various areas related to discrimination-aware algorithmic decision making.

- In Chapter 7, we add a discussion on the limitations of our work, and explore avenues of future work.

Chapter 2

Background

In this chapter, we provide background on important concepts used throughout this thesis. We start off by discussing the concept of discrimination. Next, considering that most existing notions of discrimination in machine learning literature are inspired by anti-discrimination laws, we describe different measures used to detect discrimination in legal domains in various countries. We then close the chapter by explaining how these measures are formalized in the area of machine learning.

2.1 What is discrimination?

After reviewing literature from various domains including law and philosophy, Altman (2016) defines discrimination as practices that:[1]

> "wrongfully impose a relative disadvantage on persons based on their membership in a salient social group"

While the definition is quite intuitive at the first glance, there are several important points to be considered:

Discrimination is a relative phenomenon. Altman notes that discrimination occurs when a person or a group is given disadvantageous treatment *relative to* some other group. He notes that this point is affirmed by the US Supreme Court case, *Brown v. Board of Education* (Brown, 1954) which ruled that racial segregation in public schools was discriminatory because it put African-Americans children at a *relative* disadvantage as compared to White children.

Moreover, Altman contrasts *differential treatment* with *relative disadvantage*, and mentions that not all groups that receive different treatment from each other are being

[1] Other sources like Lippert-Rasmussen (2006) and Cook (2015) provide similar definitions.

discriminated against. He argues that under the segregation practices in the American South, while the treatment of African-Americans and Whites was different from each other, and while this differential treatment might have held back the progress for everyone in the South, only African-Americans (and not Whites) were the victims of discrimination.

Not all groups are socially salient. While society can be divided into groups along different dimensions (*e.g.*, based on eye color, music preferences), not all ways of grouping people form salient social groups. According to Lippert-Rasmussen (2006), socially salient groups are the ones that are "important to the structure of social interactions across a wide range of social contexts".

On a more legal side, salient social groups (also called protected groups),[2] among other factors, are formed based on groupings that were the basis of consistent social injustices and oppression in the past (Altman, 2016; Barocas and Hardt, 2017). As a result, laws in different countries define socially salient groups accordingly. For example, with respect to employment, the protected features under the US anti-discrimination law are: race, color, gender, religion, national origin, citizenship, age, pregnancy, familial status, disability status, veteran status and genetic information (Barocas and Hardt, 2017). EU law has a very similar list of protected grounds. Interestingly, EU law also designates language as a protected ground (Fribergh and Kjaerum, 2010).

Finally, based on the contemporary discourse in a society, the definition of salient social groups is subject to change (Zarsky, 2014). For example, under US law, genetic information was only designated as a protected feature[3] in 2008 (Green et al., 2015).

Not all domains are regulated. Not all application domains in a society are regulated by anti-discrimination laws. For example, under the US law, the regulated domains are credit, education, employment, housing, public accommodation and marketing (Barocas and Hardt, 2017). Furthermore, the designation of protected groups may also vary across various domains. For example, under the US anti-discrimination law, health insurers

[2] While legal literature refers to salient social groups as "protected groups" (Barocas and Selbst, 2016), some studies in machine learning literature also refer to them as "sensitive feature groups" (Pedreschi et al., 2008). Thus, we will be using the terms *salient social group*, *protected group* and *sensitive feature group* interchangeably.

[3] We refer to the features or traits that form the basis of protected groups (*e.g.*, the feature race forms the groups: African-American, Hispanic, ...) as *socially salient group memberships*, *protected features* or *sensitive features*.

are prohibited from discriminating based on genetic information, but no such provision exists with respect to gender, race or religion (Avraham et al., 2014; GINA, 2008).

Discrimination involves groups. A point worth mentioning at this stage is that the phenomenon of discrimination by definition involves having discernible groups. For example, an employer putting applicants at relative disadvantage arbitrarily (without regard to their salient social group membership) might be unfair to the applicants in question, but (s)he will not be committing discrimination. Such scenarios involving *individual-level fairness* have previously been considered in moral philosophy (Rawls, 2009) as well as in machine learning (Dwork et al., 2012; Joseph et al., 2016; Speicher et al., 2018). On a high-level, these individual-level fairness notions require that all individuals at the same level of qualification (regardless of their group membership) should be treated similarly.

The wrongs of arbitrary rejections vs. the discriminatory rejections (based on salient social groups) are different. According to Arneson (2015): "Whereas being the object of discrimination because one belongs to a group that has been targeted for oppressive treatment in the past is likely to be a wound to one's sense of dignity and self-respect, being the victim of whimsical or idiosyncratic hiring practices is less likely to inflict a significant psychic wound over and above the loss of the job itself. Also, since whimsical discrimination is idiosyncratic, it will not lead to cumulative harm by causing anyone to be the object of economic discrimination time after time (unless whimsical hiring were common and one were extremely unlucky)".

For further discussion into the concept of discrimination (and related ideas), we point the interested reader to Altman (2016) and Arneson (2015) and references therein.

2.2 Measures of discrimination in legal domains

Having analyzed the definition of discrimination in Section 2.1, the question that arises now is, how does one operationalize this definition? That is, how does one empirically *measure* if a (algorithmic) decision making system is discriminatory? Recall from Section 2.1 that in measuring discrimination, our aim is to see if a decision making system imposes ***wrongful relative disadvantage*** on certain socially salient groups.

Since much of the work in discrimination-aware machine learning until now has been inspired by anti-discrimination legislation, we now briefly survey how discrimination is measured in various legal systems. Specifically, our goal will be to understand how anti-discrimination laws interpret wrongful relative disadvantage in the definition of discrimination in Section 2.1.

For the sake of conciseness, we will mostly focus on anti-discrimination legislation from the US and the EU. Our terminology will be driven by the US anti-discrimination laws, and we will mention the terminology used in the EU law whenever significant differences arise. For a more detailed account into the discussion that follows, we point the reader to (Altman, 2016; Bagenstos, 2015; Barocas and Selbst, 2016; FDIC's Compliance Examination Manual, 2017; Fribergh and Kjaerum, 2010; Gano, 2017; Romei and Ruggieri, 2014; Siegel, 2014).

Anti-discrimination laws mostly differentiate between two distinct forms of discrimination: ***disparate treatment*** and ***disparate impact***.

2.2.1 Disparate treatment

This measure is referred to as "direct discrimination" under the EU law (Fribergh and Kjaerum, 2010).

What constitutes disparate treatment?

According Title VII of the US Civil Rights Act of 1964, a decision making process suffers from disparate treatment if it: (i) *explicitly* or *formally* considers the sensitive group membership of a person in question, or (ii) it bases the decisions on some other factors with the *intent to discriminate* against certain groups (Barocas and Selbst, 2016). EU law also defines disparate treatment in a similar way (Fribergh and Kjaerum, 2010).

The specification above raises the following interesting points.

Once a decision maker explicitly considers the protected ground (*e.g.*, gender) in making the decision, even if the protected group membership has minimal impact on the decisions—perhaps because other (non-protected) features carried higher weight—this would still count as disparate treatment (Barocas and Selbst, 2016).

Also, a decision maker could implicitly base the decisions on sensitive features. For example, under the redlining practice in the US, a lender would deny credit to residence of certain neighborhoods based on the racial makeup of that neighborhood (Barocas and Selbst, 2016; Gano, 2017). This case would also count as disparate treatment since the lender's decision to not issue credit is based on racial profiling of the neighborhood rather than considering the merits of individuals living in that neighborhood. According to Barocas and Selbst (2016): "Redlining is illegal because it can systematically discount entire areas composed primarily of members of a protected class, despite the presence of some qualified candidates."

Finally, under certain circumstances, it may be permissible to base decisions on the protected group membership information.

For example, under Title VII of the US Civil Rights Act of 1964, an employer can justify using the protected group membership information when it qualifies as a "Bona fide occupational qualification" (BFOQ) for the job under consideration (Berman, 2000). A sensitive feature can be considered a BFOQ when it is "reasonably necessary to the normal operation of that particular business". For example, due to safety reasons, mandatory retirement ages can be enforced on airline pilots or air traffic controllers since age is a BFOQ for these jobs (Altman, 2016).

Similarly, use of sensitive features in decision making could be permitted when the goal is to advance a compelling governmental interest (*e.g.*, affirmative action policies aimed at improving racial diversity in colleges). However, as MacCarthy (2017) notes, such scenarios (where sensitive features such as race are explicitly used in decision making) would likely be subject to strict judicial scrutiny by the courts, and would need to satisfy certain stringent criteria to pass the strict scrutiny test.

How is disparate treatment detected?

We briefly discuss how disparate treatment is detected in the legal domain, since this discussion would be useful in the later part of the thesis (Sections 2.4 and Chapter 7). In the discussion that follows, the *plaintiff* refers to the party that lodges a discrimination complaint before a court (*e.g.*, a potential employee who was rejected) and the *defendant* refers to the party against whom the case is lodged (*e.g.*, the employer).

A disparate treatment liability can be established in two different ways:

The first method is where the plaintiff can show **direct evidence** that the protected group membership was a motivating factor in the defendant's decision, *e.g.*, a bar advertising publicly that they do not serve certain minorities (Altman, 2016).

The plaintiff can show **indirect evidence** of discrimination. Under US legal system, this is done via *McDonnell-Douglas burden-shifting scheme* or *Price-Waterhouse mixed motive regime* (Barocas and Selbst, 2016; Gano, 2017), whereas under EU law, a *comparator* framework is used (Fribergh and Kjaerum, 2010). Roughly, this method requires the plaintiff to show that the action to reject the plaintiff could not have been taken had the defendant not taken the sensitive group membership into account, *i.e.*, the plaintiff would not have received the negative outcome had their sensitive group membership been different (*e.g.*, had she been White and not African-American).

Finally, under the US anti-discrimination doctrine, while many sources argue that disparate treatment always corresponds to intentional discrimination—*i.e.*, the decision maker *knowingly* basing decisions on the protected group membership of a person (either directly, or via a proxy) (Federal Reserve, 2016; Gano, 2017; Gold, 2004)—others argue that disparate treatment may very well stem unintentionally, *e.g.*, from unconscious biases (Barocas and Selbst, 2016; Krieger and Fiske, 2006). However, as Barocas and Selbst (2016) note, "the law does not adequately address unconscious disparate treatment", and it is not entirely clear how such cases would be addressed.[4] On the other hand, the EU law does not require the presence of intent in order to establish a disparate treatment liability (Fribergh and Kjaerum, 2010; Maliszewska-Nienartowicz, 2014).

2.2.2 Disparate impact

This measure is referred to as "indirect discrimination" under the EU law (Fribergh and Kjaerum, 2010).

What constitutes disparate impact?

Under both US and EU laws, disparate impact occurs when "facially neutral" decision making (*e.g.*, a hiring exam) results in disproportionately adverse impact on a certain protected group (Barocas and Selbst, 2016).

Adverse impact here is said to occur when the success rates for persons from different groups (*e.g.*, African-Americans vs. Whites) are substantially different. How different is "substantially different" is often determined on a case-by-case basis in the EU law (Fribergh and Kjaerum, 2010). The same holds true for the US justice system. However, as a rough guideline in the hiring domain, the US Equal Employment Opportunity Commission suggests having an impact ratio between the two groups to be no less than 80% (Biddle, 2005). As an example, a scenario where 50% of White applicants get hired, whereas only 10% of African-American applicants get accepted, the impact ratio is $\frac{10}{50} = 0.2$.

It is vital to note that disproportionally adverse impact **does not** automatically constitute a disparate impact liability. Both US and EU legislations accommodate a business necessity defense that can justify the adverse impact. For more details regarding this justification, we next describe how a disparate impact liability is established.

How is disparate impact detected?

Under the US judicial system, the process of establishing a disparate impact liability proceeds as follows (Barocas and Selbst, 2016): (i) The plaintiff shows that a facially neutral decision making process (*e.g.*, a hiring exam) led to disproportionate adverse

[4] As we discuss shortly in Section 2.2.2, some authors argue that the disparate impact doctrine might be better suited to handle unconscious biases (Siegel, 2014).

impact on the protected group. (ii) The defendant can then show that the decision making process is related to the job and is a "business necessity", *i.e.*, the adverse impact is unavoidable. (iii) The plaintiff can counter by demonstrating that the defendant could have used an alternative decision making regime that would achieve the same outcome utility for the defendant while having lesser adverse impact. EU courts allow a similar business necessity defense (Fribergh and Kjaerum, 2010).

For example, in the US Supreme Court case *Griggs vs. Duke Power Co.* (Griggs, 1971), the court was able to establish that the hiring criteria of Duke Power Co. was not job-related, hence the adverse impact on African-Americans constituted a case of disparate impact. On the other hand, in *Ricci vs. DeStefano* (Ricci, 2009), the court found *no evidence* that the promotion test used by the New Haven Fire Department was not related to the job and hence ruled that there would be no disparate impact liability.

The justification behind disparate impact as a discrimination measure

Disparate impact is known to be a highly controversial notion of discrimination with some arguing about its validity as a suitable discrimination measure (Altman, 2016; Barocas and Selbst, 2016).

However, Siegel (2014) notes that disparate impact can be useful as a discrimination measure when one aims to either root out *well-hidden disparate treatment* (*e.g.*, an employer using proxies to intentionally discriminate against protected groups) or to address *unconscious and structural discrimination* that can arise as a result of historical biases. Specifically, she gives the following reasons about the effectiveness of disparate impact as measure of discrimination.

> "Why impose disparate impact liability? Judges and commentators, both liberal and conservative, understand disparate impact liability to redress at least three kinds of discrimination that are common in societies that have recently repudiated centuries old traditions of discrimination.
>
> The first is covert *intentional discrimination* [emphasis added]. Once a society adopts laws prohibiting discrimination, discrimination may simply go underground. When discrimination is hidden, it is hard to prove. Disparate impact tests probe facially neutral practices to ensure their enforcement does not mask covert intentional discrimination.
>
> The second is *implicit or unconscious bias* [emphasis added]. Discrimination does not end suddenly; it fades slowly. Even after a society repudiates a system of formal hierarchy, social scientists have shown that traditional

> norms continue to shape judgments in ways that may not be perceptible even to the decision maker herself. Disparate impact tests probe facially neutral practices to ensure their enforcement does not reflect implicit bias or unconscious discrimination.
>
> The third form of bias is sometimes termed *structural discrimination* [emphasis added]. An employer acting without bias may adopt a standard that has a disparate impact on groups because the standard selects for traits whose allocation has been shaped by past discrimination, whether practiced by the employer or by others with whom the employer is in close dealings. Disparate impact tests probe facially neutral practices to ensure their enforcement does not unnecessarily perpetuate the effects of past intentional discrimination."

Regardless, disparate impact remains a contentious measure, and its applicability is assessed on a case-to-case basis—see for example *Griggs vs. Duke Power Co.* (Griggs, 1971), *Ricci vs. DeStefano* (Ricci, 2009), *Texas Department of Housing and Community Affairs vs. Inclusive Communities Project, Inc.* (Inclusive Communities, 2015) and *Fisher vs. University of Texas* (Fisher, 2016).

In this thesis, when discussing disparate impact, we will assume that the administrator of the decision making system aims at removing *substantial differences* between the beneficial outcome rates for different groups. That is, given a decision making system where the beneficial outcome rates are different for different groups, the administrator might be interested in accessing an array of decision making outcomes, with decreasing values of disparity in beneficial outcome rates (*e.g.*, where the disparity in beneficial outcome rates is 0.5, 0.4, . . ., 0.0). However, as described above, a disparity in decision outcomes does not always generate a disparate impact liability for the system administrator—in the case of a legitimate business necessity, the system administrator could still justify the disparity.

Finally, somewhat related with the disparate impact doctrine is the notion of affirmative action (Barocas and Selbst, 2016; MacCarthy, 2017; Siegel, 2014). The goal of affirmative action is often to correct for historical discrimination against certain groups. Affirmative action may involve (among other things) giving preferential treatment to these groups (*e.g.*, by setting up quotas, giving special treatment to these groups). However, affirmative action is allowed under very special circumstances and is known to be highly controversial (Fribergh and Kjaerum, 2010; Fullinwider, 2018).

2.2.3 How do disparate treatment and disparate impact capture wrongful relative disadvantage?

The reasons for interpreting disparate treatment and disparate impact to be causing wrongful relative disadvantage are plentiful. Here, we describe a few of these reasons. A detailed discussion on them can be found in (Altman, 2016).

A decision making process incurring disparate treatment (*i.e.*, intentionally basing decisions on sensitive feature information) can be interpreted as causing wrongful relative disadvantage since it judges people based on immutable traits that they do not have any control over (*e.g.*, race, national origin), and it may cause arbitrary and inaccurate stereotyping that is not relevant to the task at hand.

Similar arguments hold for disparate impact, with the addition that disparate impact also tries to capture implicit biases in the decision making process, as well as the structural discrimination where the biased historical treatment of certain groups results in these groups consistently getting disadvantageous outcomes in the present.

We now move on to the design of algorithmic decision making systems, and see how disparate treatment and disparate impact are measured in the context of algorithmic decision making.

2.3 Setup of a binary classification task

In this thesis, we focus on a specific (supervised) learning task: classification. Moreover, we only consider binary classification tasks. The reason is as follows: discrimination analysis often involves tasks where the outcomes are binary in nature, with a clear distinction between a desirable (*e.g.*, getting accepted for a job) and an undesirable (*e.g.*, getting rejected from a job) outcome. However, the techniques proposed in the later sections can be extended to m-ary classification tasks as well.

In a binary classification task, given a training set, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, consisting of N users, one aims at learning a mapping between user feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and the class labels $y \in \{-1, 1\}$. Here, one assumes that $(\boldsymbol{x}, y)$ are drawn from an unknown feature distribution $f(\boldsymbol{x}, y)$.

Learning this mapping can be done using various methods. In this thesis, we focus on a broad class of learning methods: convex decision boundary-based classifiers such as logistic regression, linear and non-linear support vector machines (SVMs), *etc.*

Under convex boundary-based classifiers, the learning reduces to finding a decision boundary defined by a set of parameters $\boldsymbol{\theta}$ in the feature space that separates the users in the training set according to their class labels. One typically looks for a decision boundary, denoted as $\boldsymbol{\theta}^*$, that minimizes a certain loss function $L(\boldsymbol{\theta})$ over the training set, *i.e.*, $\boldsymbol{\theta}^* = \mathrm{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. For convex boundary-based classifiers, L is a convex function of the decision boundary parameters $\boldsymbol{\theta}$, meaning that the globally optimal solution, $\boldsymbol{\theta}^*$, can be found *efficiently* even for large datasets.

Then, for a given *unseen* feature vector $\boldsymbol{x}$, one predicts the class label $\hat{y} = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$, and $\hat{y} = 1$ otherwise. Here, $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from $\boldsymbol{x}$ to the decision boundary, $\boldsymbol{\theta}^*$.

We now give examples of some well-known convex boundary-based classifiers:

Logistic regression. In logistic regression (and other linear convex boundary-based classifiers), the distance from decision boundary is denoted as $d_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x}$. In other words, the decision boundary is represented by the hyperplane $\boldsymbol{\theta}^T\boldsymbol{x} = 0$, since we predict $\hat{y} = 1$ if $d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0$ and $\hat{y} = 1$ if $d_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0$.

Next, in logistic regression, one maps the feature vectors $\boldsymbol{x}$ to the class labels y by means of a probability distribution:

$$p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-d_{\boldsymbol{\theta}}(\boldsymbol{x})}} = \frac{1}{1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}}, \tag{2.1}$$

It is easy to see that a point lying *at* the decision boundary, *i.e.*, with $d_{\boldsymbol{\theta}}(x) = 0$, has $p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}) = 0.5$, and this probability increases with an increase in the (signed) distance from the boundary.

One obtains the optimal value of $\boldsymbol{\theta}$ by solving the following maximum likelihood problem over the training set (Murphy, 2012):

$$\underset{\boldsymbol{\theta}}{\text{minimize}} - \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x}, \boldsymbol{\theta}). \tag{2.2}$$

Linear SVM. In the case of a linear SVM, the optimal decision boundary corresponds to the maximum margin decision hyperplane (Bishop, 2006). This boundary is found by solving the following optimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & ||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} & y_i\boldsymbol{\theta}^T\boldsymbol{x}_i \geq 1 - \xi_i, \forall i \in \{1, \ldots, N\} \\ & \xi_i \geq 0, \forall i \in \{1, \ldots, N\}, \end{array} \tag{2.3}$$

where $\boldsymbol{\theta}$ and ξ are the variables. Here, minimizing $||\boldsymbol{\theta}||^2$ corresponds to maximizing the margin between the *support vectors* assigned to the two classes, and $C\sum_{i=1}^{n}\xi_i$ penalizes the number of data points falling inside the margin.

Nonlinear SVM. In a nonlinear SVM, the decision boundary is represented by the hyperplane $\boldsymbol{\theta}^T\Phi(\boldsymbol{x}) = 0$, where $\Phi(\cdot)$ is a nonlinear transformation that maps every feature vector $\boldsymbol{x}$ into a higher dimensional transformed feature space. Similar to the case of a linear SVM, one may think of finding the parameter vector $\boldsymbol{\theta}$ by solving a constrained quadratic program. However, the dimensionality of the transformed feature space can be large, or even infinite, making the corresponding optimization problem difficult to solve. Fortunately, we can leverage the *kernel trick* (Schölkopf and Smola, 2002) and resort instead to the dual form of the problem, which can be solved efficiently.

In particular, the dual form is given by (for conciseness, we use the dual form notation of Gentle et al. (2012)):

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{minimize}} & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \\ \text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C, \\ & \boldsymbol{y}^T\boldsymbol{\alpha} = 0, \end{array} \tag{2.4}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$ are the dual variables, $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T$ are the class labels, $\boldsymbol{G}$ is the $N \times N$ Gram matrix with $\boldsymbol{G}_{i,j} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$ denotes the inner product between a pair of transformed feature vectors. The distance from decision boundary is computed as: $d_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{N}\alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i)$.

Finally, the optimization problems above can be altered easily to accommodate cases where one wants to assign different cost to different type of errors, *e.g.*, assigning different cost to false positives and false negatives (Bishop, 2006).

2.4 Disparate treatment and disparate impact in binary classification

Continuing from the setup of a binary classifier in Section 2.3, we also assume that each user feature vector $\boldsymbol{x}$ in the dataset $\mathcal{D}$ is accompanied by a sensitive feature $z \in \{0, 1\}$.[5] The sensitive feature is also drawn from an unknown distribution $f(z)$ and it may be

[5] Recall from Section 2.1 that we use sensitive feature, protected feature and socially salient group membership interchangeably.

dependent on the non-sensitive feature vectors $\boldsymbol{x}$ and class labels y, *i.e.*, $f(\boldsymbol{x}, y, z) = f(\boldsymbol{x}, y|z)f(z) \neq f(\boldsymbol{x}, y)f(z)$.

Notice that (i) we defined only one sensitive feature, and (ii) defined it to be binary. This is merely for the sake of exposition. In the later sections, we will provide examples of polyvalent and several sensitive features wherever necessary.

With this specification, we can formally describe the absence of disparate treatment and disparate treatment in the outcomes of a binary classification task.

No disparate impact. A binary classifier does not suffer from disparate impact if:

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1), \tag{2.5}$$

i.e., if the probability that a classifier assigns a user to the positive class $\hat{y} = 1$ is the same for both values of the sensitive feature z, then there is no disparate impact.

No disparate treatment. Assume that $\boldsymbol{x} \circ z$ represents the concatenation of the non-sensitive feature vector $\boldsymbol{x}$ and the sensitive feature z. Also, with slight abuse of notation, we assume that $\hat{y}(\boldsymbol{x} \circ z)$ represents the decision of a classifier for a user with the given non-sensitive and sensitive features.[6] Then, a binary classifier does not suffer from disparate treatment if:

$$\hat{y}(\boldsymbol{x}_i \circ 0) = \hat{y}(\boldsymbol{x}_i \circ 1) \quad \forall i \in \{1, \dots, N\} \tag{2.6}$$

i.e., if the decision of the classifier does not change with a change in the user's sensitive feature value, then there is no disparate treatment.

Relating our specification of disparate treatment in Eq. (2.6) to the definition of disparate treatment in Section 2.2.1, we notice that Eq. (2.6) only accounts for scenarios when the sensitive feature is *directly* used in the classification task. That is, Eq. (2.6) would not detect scenarios when a decision maker uses a proxy feature such as location *with the intent* of discriminating against a certain sensitive feature group.

The difficulty with detecting such implicit disparate treatment via proxy variables is that in any classification task, most non-sensitive features (*e.g.*, educational-level, location) will likely have non-zero correlation with the sensitive feature (*e.g.*, gender). For example, a 2007 analysis of credit-based insurance scores by US Federal Trade Commission (FTC, 2007) shows that a number of "informative" features are correlated

[6] For example, for convex boundary-based classifiers, $\hat{y}(\cdot)$ would be the sign of the distance from decision boundary. For a decision tree classifier, this would be the label of the corresponding leaf node.

with race. Under such situations, it is very difficult to determine whether or not the decision maker had an intent to discriminate while using certain non-sensitive features.

To counter such scenarios, the disparate impact test (Eq. 2.5) would be a more suitable tool to detect discrimination. In fact, as Siegel (2014) notes, one of the utilities of disparate impact tests is to detect "covert intentional discrimination" and "probe facially neutral practices to ensure their enforcement does not mask covert intentional discrimination".

Having formally described disparate treatment and disparate impact in the context of classification tasks, we now move on to design classifiers that can avoid these two forms of discrimination.

Chapter 3

Classification without disparate treatment and disparate impact

While it is desirable to design classifiers free of disparate treatment as well as disparate impact, controlling for both forms of discrimination simultaneously is challenging. One could avoid disparate treatment by ensuring that the decision making process does not have access to sensitive feature information (and hence cannot make use of it). However, ignoring the sensitive feature information may still lead to disparate impact in outcomes: since automated decision-making systems are often trained on historical data, if a group with a certain sensitive feature value was discriminated against in the past, this unfairness may persist in future predictions, leading to disparate impact (Barocas and Selbst, 2016; Dwork et al., 2012). Similarly, avoiding disparate impact in outcomes by using sensitive feature information while making decisions would constitute disparate treatment, and may also lead to reverse discrimination (Ricci, 2009).

In this chapter, our goal is to design classifiers—specifically, convex margin-based classifiers like logistic regression and support vector machines (SVMs)—that avoid *both* disparate treatment and disparate impact, and can additionally accommodate the "business necessity" clause of disparate impact doctrine (Section 2.2.2). According to the business necessity clause, an employer can justify a certain degree of disparate impact in order to meet certain performance-related constraints (Barocas and Selbst, 2016). However, the employer needs to ensure that the current decision making incurs the *least possible* disparate impact under the given constraints.

Since it is very challenging to directly incorporate the disparate impact requirement into the design of many well-known classifiers like logistic regression or SVM, we introduce a novel and intuitive mechanism of decision boundary covariance: the covariance between the sensitive features and the signed distance between the users' feature vectors and the decision boundary of the classifier. The decision boundary covariance serves as a tractable proxy for measuring and limiting the disparate impact of a classifier.

Our covariance mechanism allows us to derive two complementary formulations for training nondiscriminatory classifiers: one that maximizes accuracy subject to nondiscrimination constraints, and enables compliance with disparate impact doctrine in its basic form (*i.e.*, ensuring parity in beneficial outcomes for different sensitive feature groups); and another that minimizes discrimination subject to accuracy constraints, and can help fulfill the business necessity clause of disparate impact doctrine. Remarkably, both formulations can also avoid disparate treatment, since they do not use sensitive feature information while making decisions, *i.e.*, their decisions satisfy Eq. (2.6).[7] Our mechanism additionally satisfies several desirable properties: (i) for a wide variety of convex boundary-based linear and non-linear classifiers (*e.g.*, logistic regression, SVM), it is convex and can be readily incorporated in their formulation without increasing their complexity, hence ensuring efficient learning; (ii) it allows for clear mechanisms to trade-off nondiscrimination and accuracy; and, (iii) it can be used to ensure nondiscrimination with respect to several sensitive features. Experiments using both synthetic and real-world data show that our mechanism allows for a fine-grained control of the level of nondiscrimination, often at a small cost in terms of accuracy, and provides more flexibility than the state-of-the-art.

Relevant publication

Results presented in this chapter are published in (Zafar et al., 2017b).

3.1 Methodology

First, to comply with the disparate treatment criterion in Eq. (2.6), we specify that the sensitive feature should not be a part of the decision making process *i.e.*, $\boldsymbol{x}$ and z consist of disjoint feature sets.

Next, for training a classifier adhering to the disparate impact criterion in Eq. (2.5), one can add this criterion into the classifier formulation as follows:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y}=1|z=0) - P(\hat{y}=1|z=1) \leq \epsilon, \\ & P(\hat{y}=1|z=0) - P(\hat{y}=1|z=1) \geq -\epsilon, \end{array} \tag{3.1}$$

where a smaller value of $\epsilon \in \mathbb{R}^+$ would result in a classifier more adherent to Eq. (2.5).

[7] As we explain shortly in Section 3.1, the sensitive feature information is needed only during the training phase to learn nondiscriminatory classifier parameters.

Unfortunately, it is very challenging to solve the above optimization problem for convex boundary-based classifiers, since for many such classifiers (*e.g.*, SVM) the probabilities are a non-convex function of the classifier parameters $\boldsymbol{\theta}$ and, therefore, would lead to non-convex formulations, which are difficult to solve efficiently. Secondly, as long as the user feature vectors lie on the same side of the decision boundary, the probabilities are invariant to changes in the decision boundary. In other words, the probabilities are functions having saddle points. The presence of saddle points furthers complicate the procedure for solving non-convex optimization problems (Dauphin et al., 2014).

To overcome these challenges, we next introduce a novel measure of decision boundary covariance which can be used as a proxy to efficiently design classifiers satisfying Eq. (2.5).

Our measure of decision boundary covariance stems from the intuition that if two groups have high disparity in their probabilities of being assigned to the positive class, *i.e.*, if Eq. (2.5) is far from being satisfied, then the average signed distances from decision boundary for the two groups are also likely to be quite different from each other. Hence, by controlling the relationship between the sensitive feature and the signed distance from decision boundary, one could hope to limit disparate impact in the predicted labels. We now formalize this intuition below.

3.1.1 Decision boundary covariance

Our measure of decision boundary covariance is defined as the covariance between the users' sensitive feature, z, and the signed distance from the users' feature vectors to the decision boundary, $d_{\boldsymbol{\theta}}(\boldsymbol{x})$, *i.e.*:

$$
\begin{aligned}
\text{Cov}(z, d_{\boldsymbol{\theta}}(\boldsymbol{x})) &= \mathbb{E}[(z-\bar{z})(d_{\boldsymbol{\theta}}(\boldsymbol{x}) - \bar{d}_{\boldsymbol{\theta}}(\boldsymbol{x}))] \\
&= \mathbb{E}[(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}) - (z-\bar{z})\bar{d}_{\boldsymbol{\theta}}(\boldsymbol{x})] \\
&= \mathbb{E}[(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}[(z-\bar{z})]\bar{d}_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&\approx \frac{1}{N} \sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}),
\end{aligned}
\tag{3.2}
$$

where $\mathbb{E}[(z-\bar{z})]\bar{d}_{\boldsymbol{\theta}}(\boldsymbol{x})$ cancels out since $\mathbb{E}[(z-\bar{z})] = 0$. Since in linear models for classification, such as logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\boldsymbol{\theta}^T\boldsymbol{x} = 0$, Eq. (3.2) reduces to $\frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x}$.

In contrast to the probabilities in Eq. (3.1), the decision boundary covariance (Eq. (3.2)) is a convex function with respect to the decision boundary parameters $\boldsymbol{\theta}$, since $d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is convex with respect to $\boldsymbol{\theta}$ for all linear, convex margin-based classifiers.[8] Hence, it can be easily included in the formulation of these classifiers while retaining efficient training.

Moreover, note that, if a decision boundary satisfies Eq. (2.5), then the (empirical) covariance will be approximately zero for a sufficiently large training set.

3.1.2 Maximizing accuracy under nondiscrimination constraints

In this section, we design classifiers that maximize accuracy subject to nondiscrimination constraints (*i.e.*, satisfying Eq. (2.5)), and thus may be used to ensure compliance with the disparate impact doctrine in its basic form.

To this end, we replace the probabilities in Eq. (3.1) with decision boundary covariance and find the decision boundary parameters $\boldsymbol{\theta}$ by minimizing the corresponding loss function over the training set under nondiscrimination constraints, *i.e.*:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}) \leq c, \\ & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq -c, \end{array} \tag{3.3}$$

where $c \in \mathbb{R}^+$ is the covariance threshold, which specifies an upper bound on the covariance between each sensitive feature and the signed distance from the feature vectors to the decision boundary. In this formulation, c trades off nondiscrimination and accuracy, such that as we decrease c towards zero, the resulting classifier will be more compliant with Eq. (2.5) but will potentially suffer from a larger loss in accuracy. Note that since the above optimization problem is convex, our scheme ensures that the trade-off between the classifier loss function and decision boundary covariance is Pareto-optimal.

Finally, for *multiple sensitive features* (*e.g.*, gender, race), one can include constraints for each sensitive feature separately. For *polyvalent sensitive features* having $k \geq 2$ values, one can first convert the sensitive feature into k binary sensitive features using a one hot encoding scheme, and then add constraints for each of the k sensitive features.

Remarks. It is important to note that the distance to the margin, $d_{\boldsymbol{\theta}}(\boldsymbol{x})$, only depends on the non-sensitive features $\boldsymbol{x}$ and, therefore, the sensitive feature z is not needed while making decisions. In other words, we account for *disparate treatment*, by removing the sensitive features from the decision making process and, for *disparate impact*, by adding nondiscrimination constraints during (only) the training process of the classifier.

[8] For non-linear convex margin-based classifiers like non-linear SVM, equivalent of $d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is still convex in the transformed kernel space.

Next, we specialize problem (3.3) for a logistic regression and a non-linear SVM classifier.

Logistic Regression without disparate impact. Continuing from the setup of a logistic regression classifier in Section 2.3, one can modify it to add disparate impact constraints as follows:

$$\begin{array}{lll} \underset{\boldsymbol{\theta}}{\text{minimize}} & -\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) & \big\}\ \text{Logistic regression formulation} \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x} \leq c, & \multirow{2}{*}{$\Big\}$ Disparate impact constraints} \\ & \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x} \geq -c. & \end{array} \tag{3.4}$$

Linear SVM without disparate impact. The formulation of the linear SVM classifier in Section 2.3 can be extended to include disparate impact constraints as follows:

$$\left.\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & \|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} & y_i\boldsymbol{\theta}^T\mathbf{x}_i \geq 1-\xi_i, \forall i \in \{1,\ldots,N\} \\ & \xi_i \geq 0, \forall i \in \{1,\ldots,N\}, \end{array}\right\}\ \text{SVM formulation}$$

$$\left.\begin{array}{l} \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z_i-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x}_i \leq c, \\ \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z_i-\bar{z})\,\boldsymbol{\theta}^T\boldsymbol{x}_i \geq -c, \end{array}\right\}\text{Disparate impact constraints} \tag{3.5}$$

Non-linear SVM without disparate impact. One can extend the formulation of non-linear SVM in Eq. 2.4 to include the disparate impact constraints. Specifically, one can use the kernel trick in the constraints as well:

$$\left.\begin{array}{ll} \underset{\alpha}{\text{minimize}} & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \\ \text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C, \\ & \boldsymbol{y}^T\boldsymbol{\alpha} = 0, \end{array}\right\}\ \text{SVM formulation}$$

$$\left.\begin{array}{l} \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\, d_{\boldsymbol{\alpha}}(\boldsymbol{x}) \leq c, \\ \frac{1}{N}\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z})\, d_{\boldsymbol{\alpha}}(\boldsymbol{x}) \geq -c, \end{array}\right\}\ \begin{array}{c}\text{Disparate impact}\\ \text{constraints}\end{array} \tag{3.6}$$

where $d_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{N}\alpha_i y_i k(\boldsymbol{x},\boldsymbol{x}_i)$ can still be interpreted as the signed distance from decision boundary (Schölkopf and Smola, 2002).

3.1.3 Minimizing disparate impact under accuracy constraints

In the previous section, we designed classifiers that maximize accuracy subject to nondiscrimination constraints. However, if the underlying correlation between the class labels and the sensitive features in the training set is very high, enforcing nondiscrimination constraints may result in underwhelming performance (accuracy) and thus be unacceptable in terms of business objectives. Disparate impact's "business necessity" clause accounts for such scenarios by allowing some degree of disparate impact in order to meet performance constraints. However, the employer needs to ensure that the decision making causes *least possible* disparate impact under the given performance (accuracy) constraints (Barocas and Selbst, 2016). To accommodate such scenarios, we now propose an alternative formulation that minimizes discrimination (disparate impact) subject to accuracy constraints.

To this aim, we find the decision boundary parameters $\boldsymbol{\theta}$ by minimizing the corresponding (absolute) decision boundary covariance over the training set under constraints on the classifier loss function, *i.e.*:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & \left| \frac{1}{N} \sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z - \bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}) \right| \\ \text{subject to} & L(\boldsymbol{\theta}) \le (1+\gamma) L(\boldsymbol{\theta}^*), \end{array} \tag{3.7}$$

where $L(\boldsymbol{\theta}^*)$ denotes the optimal loss over the training set provided by the unconstrained classifier and $\gamma \ge 0$ specifies the maximum additional loss with respect to the loss provided by the unconstrained classifier. Here, we can ensure maximum nondiscrimination with no loss in accuracy by setting $\gamma = 0$. As in Section 3.1.2, it is possible to specialize problem (3.7) for the same classifiers and show that the formulation remains convex.

Fine-grained accuracy constraints. In many classifiers, including logistic regression and SVMs, the loss function (or the dual of the loss function) is additive over the points in the training set, *i.e.*, $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta})$, where $L_i(\boldsymbol{\theta})$ is the individual loss associated with the i-th point in the training set. Moreover, the individual loss $L_i(\boldsymbol{\theta})$ typically tells us how *close* the predicted label $f(\boldsymbol{x}_i)$ is to the true label y_i, by means of the signed distance to the decision boundary. Therefore, one may think of incorporating loss constraints for a certain set of users, and consequently, prevent individual users originally classified as positive (by the unconstrained classifier) from being classified as negative by the constrained classifier. To do so, we find the decision boundary parameters $\boldsymbol{\theta}$ as:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})\, d_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \right| \\ \text{subject to} & L_i(\boldsymbol{\theta}) \le (1+\gamma_i) L_i(\boldsymbol{\theta}^*) \quad \forall i \in \{1, \ldots, N\}, \end{array} \tag{3.8}$$

where $L_i(\boldsymbol{\theta}^*)$ is the individual loss associated to the i-th user in the training set provided by the unconstrained classifier and $\gamma_i \ge 0$ is her allowed additional loss.

The constraints in Eq. 3.8 can also help ensure that there are no egregious misclassifications while adding nondiscrimination requirements into the classifier training.

3.2 Evaluation

In this section, we experiment with several synthetic and real-world datasets to evaluate the effectiveness of our decision boundary covariance in controlling disparate treatment and disparate impact.

Across this section, we quantify disparate impact (Eq. 2.5) as the absolute difference between the positive class probability for the sensitive feature groups with $z = 0$ and $z = 1$, as in various prior studies, (Calders and Verwer, 2010; Corbett-Davies et al., 2017b; Menon and Williamson, 2017), *i.e.*:

$$DI = \Big|P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)\Big|, \tag{3.9}$$

where a value of DI closer to zero denotes a smaller degree of disparate impact.

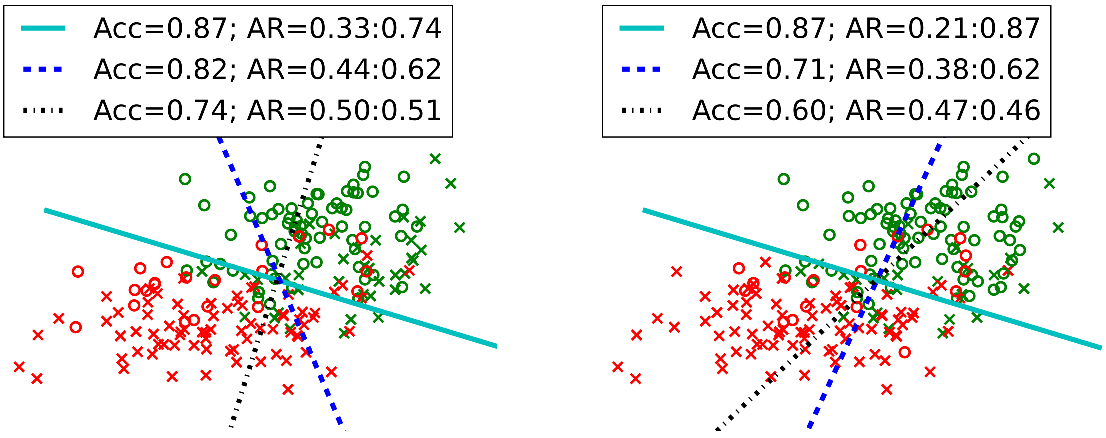
Some studies also adopt another measure of disparate impact, the p-rule. This measure quantifies the differences between positive class probabilities for the two groups using ratios instead of differences.[9] However, as the publication corresponding to this chapter shows (Zafar et al., 2017b), using p-rule as a measure of disparate impact leads to very similar experimental insights.

3.2.1 Synthetic datasets

To simulate different degrees of disparate impact in classification outcomes, we generate two synthetic datasets with different levels of correlation between a single, binary sensitive feature and class labels. We then train two types of logistic regression classifiers: one type maximizes accuracy subject to disparate impact constraints (Section 3.1.2), and the other minimizes disparate impact under fine-grained accuracy constraints (Section 3.1.3).

Specifically, we generate $4{,}000$ binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from two different

[9] The p-rule is defined as: $\min(\frac{P(\hat{y}=1|z=0)}{P(\hat{y}=1|z=1)}, \frac{P(\hat{y}=1|z=1)}{P(\hat{y}=1|z=0)})$. This measure is inspired by the guidelines by the US Equal Employment Opportunity Commission stating that the acceptance ratios between the protected and non-protected groups should be no less than 0.8. However, courts in Europe have been known to use the difference instead of the ratios (Bernard and Hepple, 1999). Since both p-rule and Eq. (3.9) are designed to quantify significant disparities in acceptance rates, we expect both of them to convey similar insights (barring corner cases, such as, when the acceptance rates are very close to 0 or very close to 1.0).



(a) $\phi = \pi/4$

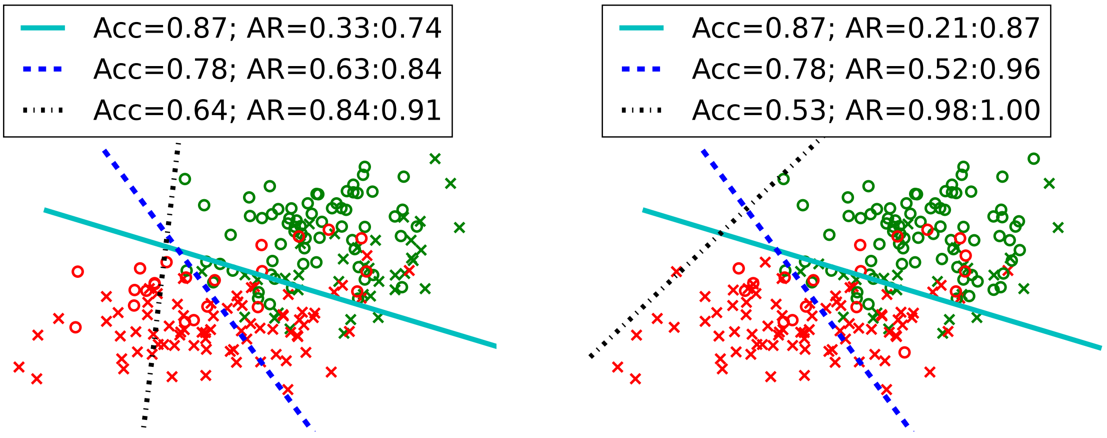
(b) $\phi = \pi/8$

Figure 3.1: [Synthetic data: Maximizing accuracy subject to disparate impact constraints] Performance of different (unconstrained and constrained) classifiers along with their accuracy (Acc) and positive class acceptance rates (AR) for groups $z = 0$ (crosses) and $z = 1$ (circles). Green points represent examples with $y = 1$ and red points represent example with $y = -1$. The solid lines show the decision boundaries for logistic regression classifiers without disparate impact constraints. The dashed lines show the decision boundaries for logistic regression classifiers trained to maximize accuracy under disparate impact constraints (Eq. (3.4)). Each column corresponds to a dataset with different correlation value between sensitive feature values and class labels. Lowering the covariance threshold c towards zero lowers the degree of disparate impact, but causes a greater loss in accuracy. Furthermore, for the dataset with higher correlation between the sensitive feature and class labels ($\pi/8$), the loss in accuracy is greater.

Gaussian distributions:

$$p(\boldsymbol{x}|y = 1) = \mathcal{N}([2; 2], [5,\ 1; 1,\ 5])$$
$$p(\boldsymbol{x}|y = -1) = \mathcal{N}([-2; -2], [10,\ 1; 1,\ 3]).$$

Then, we draw each user's sensitive feature z from a Bernoulli distribution: $p(z = 1) = p(\boldsymbol{x}'|y = 1)/(p(\boldsymbol{x}'|y = 1) + p(\boldsymbol{x}'|y = -1))$, where $\boldsymbol{x}' = [\cos(\phi),\ -\sin(\phi); \sin(\phi),\ \cos(\phi)]\boldsymbol{x}$ is simply a rotated version of the feature vector, $\boldsymbol{x}$. We generate two datasets with different values for the parameter ϕ ($\pi/4$ and $\pi/8$), which controls the correlation between the sensitive feature, z, and the class labels, y (and in turn, the resulting degree of disparate



(a) $\phi = \pi/4$

(b) $\phi = \pi/8$

Figure 3.2: [Synthetic data: Minimizing disparate impact subject to fine-grained accuracy constraints] The dashed lines show the decision boundaries for logistic regression classifiers trained to minimize disparate impact with constraints that prevents users with z = 1 (circles) labeled as positive by the unconstrained classifier from being moved into the negative class in the process (Eq. (3.8)). As compared to the previous experiment in Figure 3.1, the constrained classifier now leads to a *rotations* as well as *shifts* in the unconstrained decision boundaries (in order to prevent the specified points from being classified into the negative class).

impact). Here, the closer ϕ is to zero, the higher the correlation between z and y, and hence, the higher the degree of disparate impact.

Next, we train logistic regression classifiers optimizing for accuracy on both the datasets. The accuracy of the classifiers in both cases is 0.87 (note that the datasets only differ in terms of the correlation between z and y). However, the classifiers lead to $DI = |0.33 - 0.74| = 0.41$ and $DI = |0.21 - 0.87| = 0.66$ on datasets with $\phi = \pi/4$ and $\phi = \pi/8$, respectively. To overcome this discrimination, we train logistic regression classifiers with disparate impact constraints (Eq. 3.4) on both datasets.

Figure 3.1 shows the decision boundaries provided by the classifiers for two (successively decreasing) covariance thresholds, c. We compare these boundaries against the unconstrained decision boundary (solid line). As expected, given the data generation process, the disparate impact constraints map into a rotation of the decision boundary (dashed lines), which is greater as we decrease threshold value c or increase the correlation in the original data (from $\phi = \pi/4$ to $\phi = \pi/8$). This movement of the decision boundaries shows that our disparate impact constraints are successfully undoing (albeit in a highly controlled setting) the rotations we used to induce disparate impact in the dataset. Moreover, a smaller covariance threshold (a larger rotation) leads to a more nondiscriminatory solution, although, it comes at a larger cost in accuracy.

Figure 3.2 shows the decision boundaries provided by the classifiers that minimize disparate impact under fine-grained accuracy constraints (Eq. (3.8)). Here, the fine-grained accuracy constraints ensure that the users with $z = 1$ classified as positive by the unconstrained classifier (circles above the solid line) are not labeled as negative by the constrained classifier. The decision boundaries provided by this formulation, in contrast to the previous one, are rotated *and shifted* versions of the unconstrained boundary. Such shifts enable the constrained classifiers to avoid negatively classifying users specified in the constraints.

Next, we illustrate how the decision boundary of a non-linear classifier, a SVM with radial basis function (RBF) kernel, changes under disparate impact constraints (Eq. (3.6)). To this end, we generate $4{,}000$ user binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from

$$\begin{aligned} p(\boldsymbol{x}|y = 1, \beta) &= \beta\mathcal{N}([2; 2], [5\ 1; 1\ 5]) + (1 - \beta)\mathcal{N}([-2; -2], [10\ 1; 1\ 3]) \\ p(\boldsymbol{x}|y = -1, \beta) &= \beta\mathcal{N}([4; -4], [4\ 4; 2\ 5]) + (1 - \beta)\mathcal{N}([-4; 6], [6\ 2; 2\ 3]) \end{aligned}$$

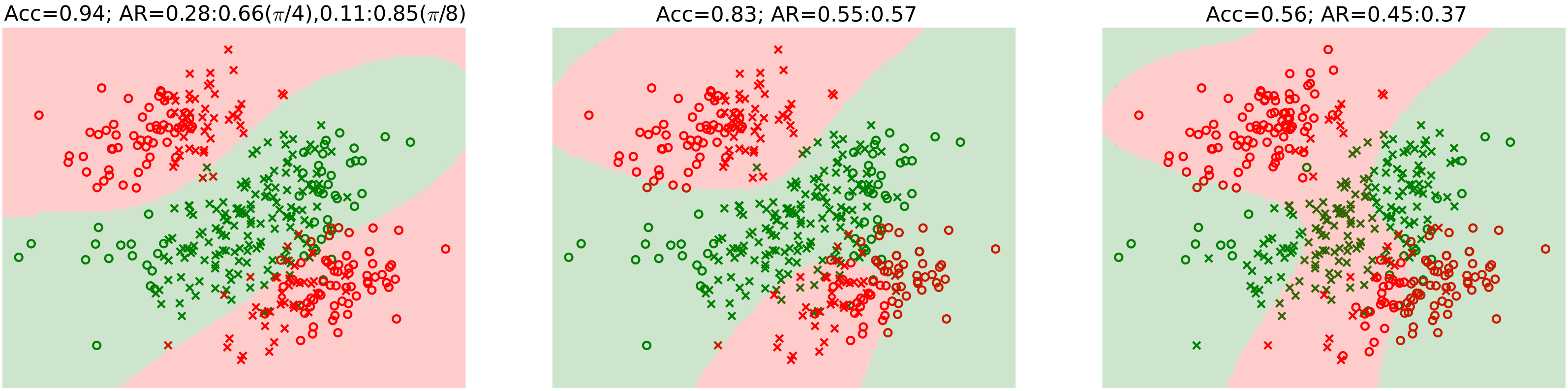
where $\beta \in \{0, 1\}$ is sampled from Bernoulli(0.5). Then, we generate each user's sensitive feature z by applying the same rotation as described earlier.

Figure 3.3 shows the decision boundaries provided by the SVM that maximizes accuracy under disparate impact constraints with $c = 0$ for two different correlation values: $\phi = \pi/4$ and $\phi = \pi/8$, in comparison with the unconstrained SVM. We observe that, in this case, the decision boundaries provided by the constrained SVMs are very different to the decision boundary provided by the unconstrained SVM, and are not just simple shifts or rotations of the latter.

3.2.2 Real-world datasets

We now evaluate the effectiveness of our covariance framework in removing disparate impact on real-world datasets. In doing so, we also compare the performance of our framework to several methods from the non-discriminatory machine learning literature.

In all the experiments, to obtain more reliable estimates of accuracy and disparate impact, we repeatedly split each dataset into a train (70%) and test (30%) set 5 times and report the average statistics for accuracy and disparate impact.



Acc=0.94; AR=0.28:0.66(π/4),0.11:0.85(π/8)

Acc=0.83; AR=0.55:0.57

Acc=0.56; AR=0.45:0.37

(a) Unconstrained

(b) $\phi = \pi/4$

(c) $\phi = \pi/8$

Figure 3.3: [Synthetic data: Maximizing accuracy subject to disparate impact constraints] Decision boundaries for SVM classifier with RBF Kernel trained without disparate impact constraints (left) and with disparate impact constraints (middle and right) on two synthetic datasets. Also shown are the classification accuracy (Acc) and acceptance rate (AR) for each group. The decision boundaries for the constrained classifier are not just the rotated and shifted version of the unconstrained classifier.

Datasets and Experimental Setup

Here, we experiment with two real-world datasets: The Adult income dataset (Adult, 1996) and the Bank marketing dataset (Bank, 2014).

The Adult dataset contains a total of $45{,}222$ subjects, each with 14 features (*e.g.*, age, educational level) and a binary label, which indicates whether a subject's annual income is above (positive class) or below (negative class) 50K USD. With the aim of experimenting with binary as well as non-binary (polyvalent) sensitive features, we consider the features gender and race to be sensitive. Here, gender (with feature values: men and women) serves as an example of binary sensitive feature and race (with feature values: American-Indian, Asian, Black, White and Other) serves as an example of a non-binary sensitive feature.

The Bank dataset contains a total of $41{,}188$ subjects, each with 20 features (*e.g.*, marital status) and a binary label, which indicates whether the client has subscribed (positive class) or not (negative class) to a term deposit. In this case, we consider age as (binary) sensitive feature, which is discretized to indicate whether or not the client's age is between 25 and 60 years.

For detailed statistics about the distribution of different sensitive features in positive class in these datasets, we refer the reader to Appendix A.

For the sake of conciseness, while presenting the results for binary sensitive features, we refer to women and men, respectively, as protected and non-protected groups in Adult data. Similarly, in Bank data, we refer to users between age 25 and 60 as protected and rest of the users as non-protected group.

Methods

In our experiments, we also compare our approach to well-known competing method from discrimination-aware machine learning literature (detailed in Chapter 6). More specifically, we consider the following methods:

- ***Our method (C-LR and C-SVM)*:** Implements our covariance constraints-based methods for controlling disparate impact with a logistic regression classifier (Eq. (3.4)) and a dual-form SVM classifier with a linear kernel (Eq. (3.6)). On the datasets considered here, different choices of kernel (linear vs. RBF) lead to a very similar performance in terms of accuracy and disparate impact. This method does not use the sensitive feature information at decision time.
- ***Preferential sampling (PS-LR and PS-SVM)*:** Implements the data pre-processing technique of Kamiran and Calders (2010) on a logistic regression and a SVM classifier. Specifically, this method operates as follows: (i) We first train a standard

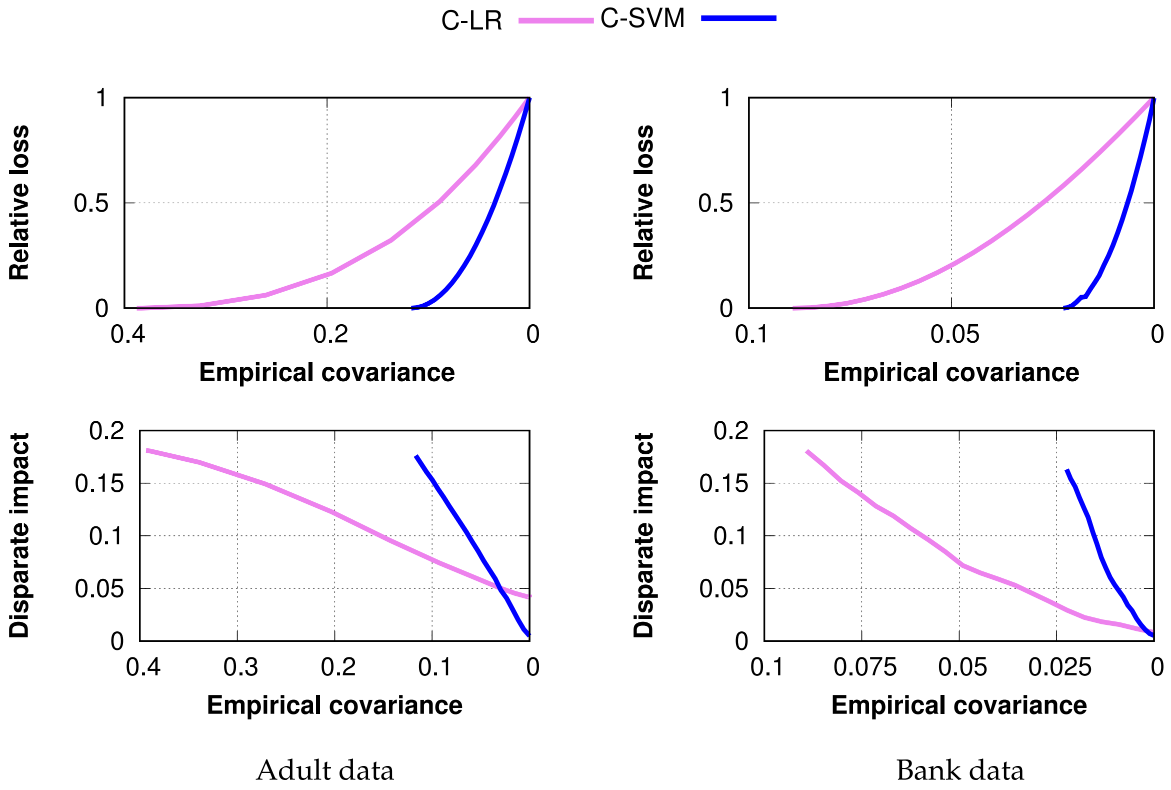


Figure 3.4: [Real-world data: Maximizing accuracy subject to disparate impact constraints on a single, binary sensitive feature] Panels in the top row show the trade-off between the empirical covariance in Eq. (3.2) and the relative loss (with respect to the unconstrained classifier), for the Adult (left) and Bank (right) datasets. Here each pair of (covariance, loss) values is guaranteed to be Pareto optimal by construction. Panels in the bottom row show the correspondence between the empirical covariance and disparate impact in Eq. (3.9) for classifiers trained under disparate impact constraints. The figure shows that a decreasing empirical covariance leads to higher loss but lower disparate impact.

(potentially discriminatory) classifier on the given dataset. (ii) Next, we move / replicate the protected group data points to / on the positive side of the decision boundary (and vice versa for the non-protected group) until the decision boundary leads to zero disparate impact, *i.e.*, until it satisfies Eq. (2.5). (iii) We then train the final (non-discriminatory) classifier on the perturbed dataset. This method does not use the sensitive feature information at decision time.

- ***Regularized logistic regression (R-LR)*:** The in-processing regularized logistic regression technique of Kamishima et al. (2011). This technique is only limited to the logistic regression classification model. This technique works by adding a regularization term in the objective function that penalizes the mutual information between the sensitive feature and the classifier decisions. This technique needs the sensitive feature information at decision time, hence cannot remove disparate treatment.

- ***Post-Processing (PP-LR and PP-SVM)*:** The post-processing technique discussed in Corbett-Davies et al. (2017b). This method works by training a standard logistic regression or SVM classifier on the given dataset. It then finds a pair of acceptance thresholds[10] such that the decisions based on those thresholds lead to maximum accuracy while having no disparate impact. This technique also requires the sensitive feature information at decision time so it cannot avoid disparate treatment.

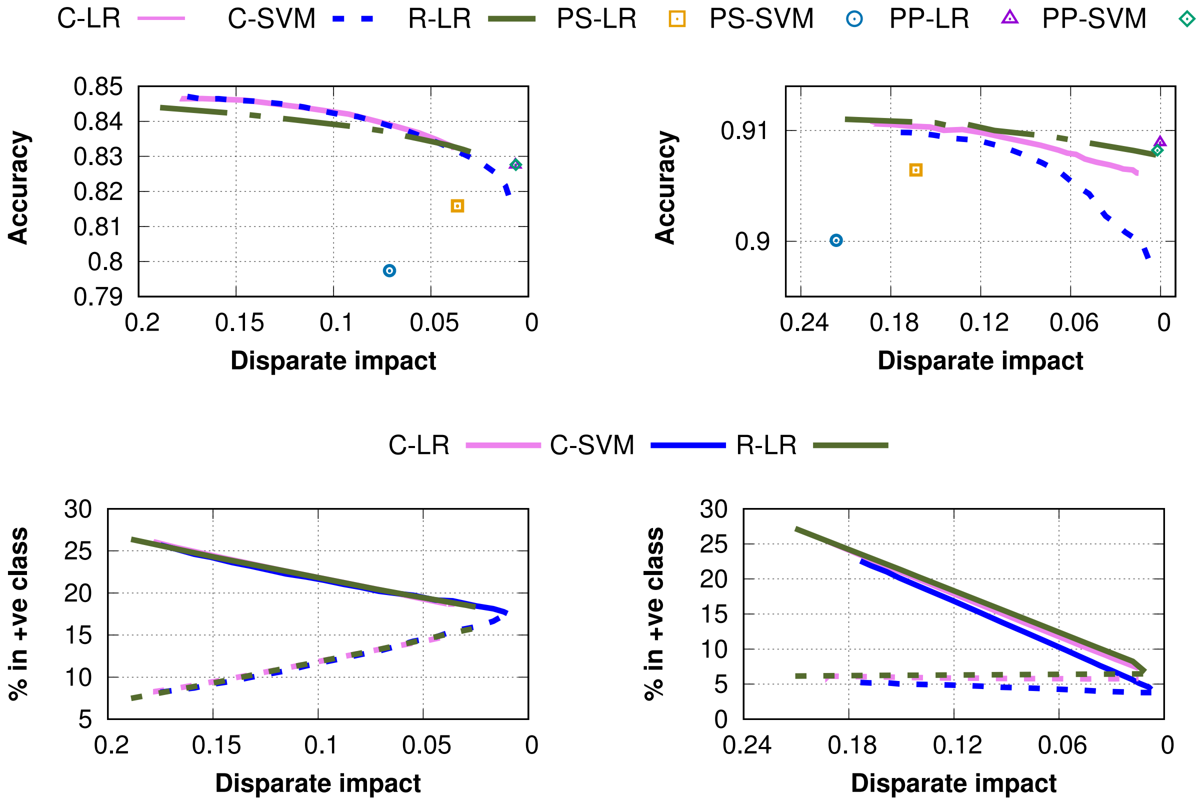
Results

First, we experiment with two standard (unconstrained) logistic regression and SVM classifiers. In the Adult dataset, the logistic regression classifier leads to an accuracy of 0.846. However, the classifier results in highly disparate positive class acceptance rates for protected and non-protected groups: 0.08 and 0.26. The SVM classifier leads to a similar accuracy (0.847) and disparity in positive class acceptance rates (0.08 vs 0.25). In the Bank dataset, the two classifiers lead to accuracies of 0.911 and 0.910, respectively, and acceptance rates of 0.06 vs. 0.25, and 0.05 vs. 0.23 respectively. The high disparity in acceptance rates over the two datasets clearly constitutes a case of disparate impact.

We then apply our framework to eliminate disparate impact with respect to a single binary sensitive feature, gender and age, for respectively, the Adult and Bank datasets. For each dataset, we train several logistic regression and SVM classifiers (denoted by 'C-LR' and 'C-SVM', respectively), each subject to disparate impact constraints with different values of covariance threshold, c (Eqs.(3.4, 3.6)). Next, we study the effect of covariance constraints on the loss function value, level of disparate impact and accuracy of the classifier.

Figure 3.4 (top row) shows the empirical decision boundary covariance against the relative loss incurred by the classifier. The 'relative loss' is normalized between the loss incurred by an unconstrained classifier and by the classifier with a covariance threshold of 0. We notice that as expected, a decreasing value of empirical covariance results in an increasing loss. However, each pair of (covariance, loss) values is guaranteed to be Pareto optimal, since our problem formulation is convex. The bottom row in Figure 3.4 investigates the correspondence between decision boundary covariance and disparate impact (Eq. (3.9)) computed on the training set, showing that, as desired: i) the lower the

[10] The acceptance threshold is 0 for a standard logistic regression or SVM classifier.



Adult data

Bank data

Figure 3.5: [Real-world data: Maximizing accuracy subject to disparate impact constraints on a single, binary sensitive feature] The figure shows the accuracy against disparate impact in Eq. 3.9 (top) and the percentage of protected (dashed) and non-protected (solid) users in the positive class against the disparate impact value (bottom). For all methods, a decreasing degree of disparate impact also leads to a decreasing accuracy. The post-processing technique (PP-LR and PP-SVM) achieves the best disparate impact-accuracy tradeoff. However, this technique as well as R-LR use the sensitive feature information at decision time (as opposed to C-LR, C-SVM, PS-LR and PS-SVM), and would hence lead to a violation of disparate treatment (Eq. 2.6).

covariance, the lower the disparate impact of the classifier and (ii) 0 disparate impact maps to roughly zero covariance.

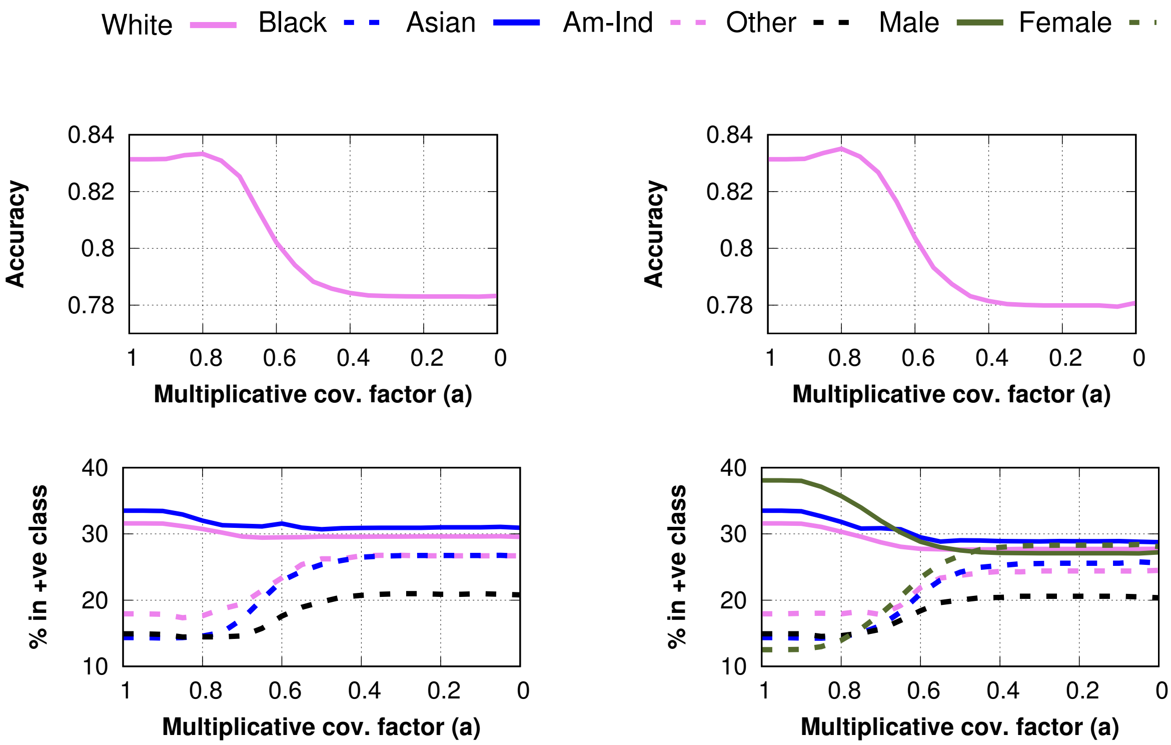
We next compare the performance of our constrained classifiers in terms of disparate impact–accuracy tradeoffs with the baselines methods mentioned above. The results presented in Figure 3.5, top row, show that: i) the performance of our classifiers (C-LR, C-SVM) and regularized logistic regression (R-LR) is comparable, ours are slightly better for Adult data (left column) while slightly worse for Bank data (right column); ii) the preferential sampling presents the worst performance and results in high disparate impact; and, (iii) the post-processing technique leads to the best performance among all methods. However, we note that both R-LR and PP-LR / PP-SVM use the sensitive feature information at the decision time while the other two techniques do not.

For a more fair comparison, we also train our method with access to sensitive features at decision time. Specifically, we train constrained logistic regression classifiers (C-LR) under the same setup as above, with the exception that the non-sensitive ($\boldsymbol{x}$) and sensitive features (z) are not disjoint feature sets—that is, the classifier learns a non-zero weight for the sensitive feature z.

Under this setup, on the Adult dataset, our constrained logistic regression classifier (C-LR) achieves an accuracy of 0.839 and DI of 0.09, as compared to 0.828 accuracy and 0.01 DI achieved by the PP-LR classifier. In this case C-LR achieves a better accuracy than PP-LR, but does not remove DI as well as PP-LR. Next, we adjust the thresholds of PP-LR in a way that the resulting classifier has DI ≤ 0.9 (*i.e.*, it tries to match the DI of C-LR) while maximizing accuracy. Under these thresholds, PP-LR achieves an accuracy of 0.840 and DI of 0.07. On the Bank dataset, C-LR achieves an accuracy of 0.908 (0.909 for PP-LR) and DI of 0.01 (0.0 for PP-LR). On both Bank and Adult datasets, both methods achieve similar accuracy for a similar level of DI (with PP-LR performing marginally better).

The bottom row of Figure 3.5 shows the percentage of users from protected and non-protected groups in the positive class along with the degree of disparate impact. We note that in the Adult data, all classifiers move non-protected users (men) to the negative class and protected users (women) to the positive class to remove disparate impact. In contrast, in the Bank data, they only move non-protected (young and old) users originally labeled as positive to the negative class since it provides a smaller accuracy loss. However, the latter can be problematic: from a business perspective, a bank may be interested in finding potential subscribers rather than losing existing customers. This observation could motivate the business necessity clause of the disparate impact doctrine. To counter such situations, one can use our alternative formulation in Section 3.1.3. We experiment with this formulation later in this section.

Finally, we apply our framework to eliminate disparate impact with respect to non-binary (race) and several (gender and race) sensitive features in the Adult dataset. We do not compare with competing methods since the pre-propressing and in-processing methods described above cannot handle non-binary or several sensitive features, whereas the post-processing technique—which involves trying various combinations of sensitive feature group-conditional thresholds—can become unscalable with an increase in number of groups. Figure 3.6 summarizes the results by showing the accuracy and the percentage of subjects sharing each sensitive feature value classified as positive against a multiplicative covariance factor $a \in [0, 1]$ such that $c = ac^*$, where c^* is the unconstrained



(a) Non-binary (polyvalent) sensitive feature

(b) Multiple sensitive features

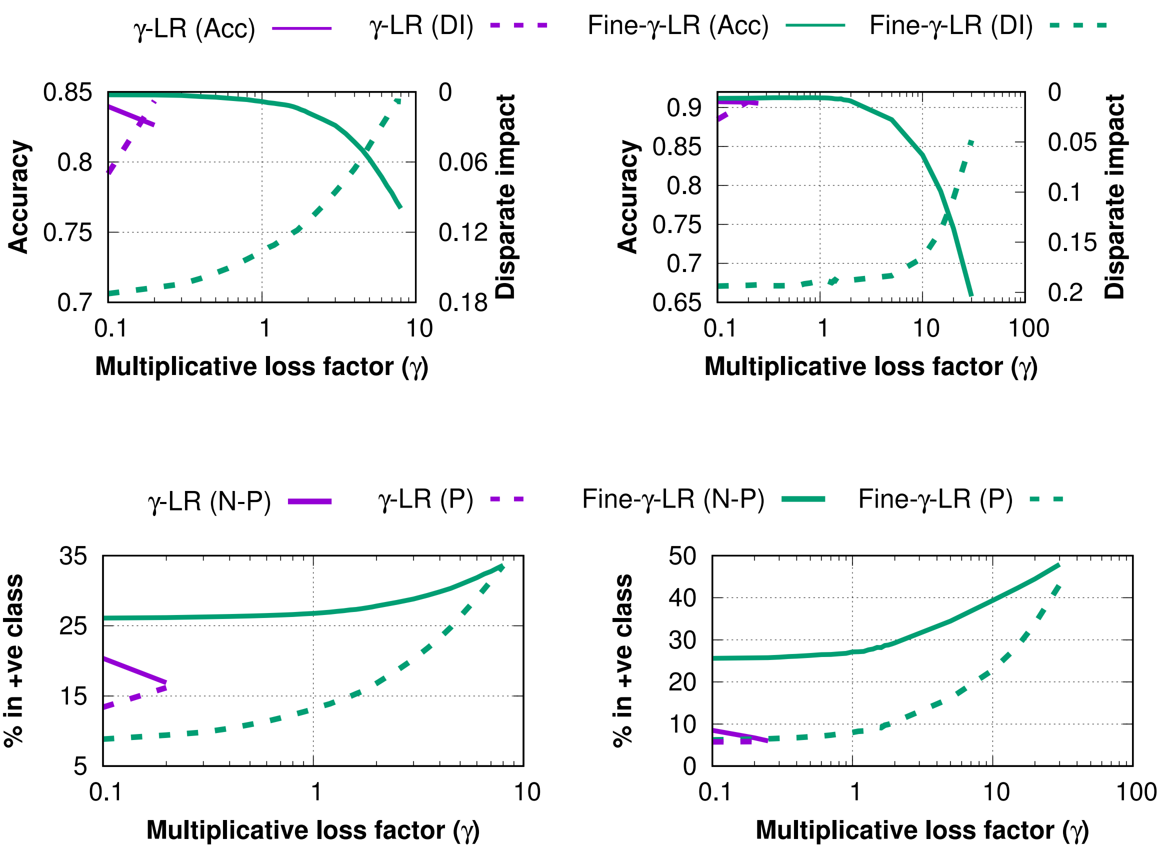
Figure 3.6: [Real-world data: Maximizing accuracy subject to disparate impact constraints on a polyvalent (left) and multiple (right) sensitive features] The figure shows accuracy (top) and percentage of users in positive class (bottom) against a multiplicative factor $a \in [0, 1]$ such that $c = ac^*$, where c^* denotes the unconstrained classifier covariance. Reducing the covariance threshold leads to outcomes with less and less disparate impact, but causes further drops in accuracy.

classifier covariance[11] (note that disparate impact in Eq. (3.9) is only defined for a binary sensitive feature). As expected, as the value of c decreases, the percentage of subjects in the positive class from sensitive feature value groups become nearly equal [12] while the loss in accuracy is modest.

Disparate impact's business necessity clause. We now experiment with our formulation for handling the business necessity clause (Section 3.1.3) to avoid scenarios where removing disparate impact leads to almost all the users being assigned the negative class label (Figure 3.5). Specifically, we demonstrate that our formulation in Section 3.1.3

[11] For several sensitive features, we compute the initial covariance c^*_k for each of the sensitive feature k, and then compute the covariance threshold separately for each sensitive feature as ac^*_k.

[12] With the exception of the race 'Other'. We note that this 'Other' constitutes a very small part of the whole data (0.8%) and among other factors, this exception could have been caused by the inaccurate estimation of decision boundary covariance due to sparse representation of this group.



Adult

Bank

Figure 3.7: [Minimizing disparate impact subject to constraints on accuracy, or on $-ve$ class classification for certain points] Panels in top row show the accuracy (solid) and disparate impact (dashed) against γ. Panels in the bottom row show the percentage of protected (P, dashed) and non-protected (N-P, solid) users in the positive class against γ. Allowing for more loss in accuracy results in a solution with less disparate impact.

can minimize disparate impact while precisely controlling loss in accuracy. We also demonstrate that our formulation can additionally provide guarantees for classifying certain users in the positive class while minimizing disparate impact.

To this end, we first train several logistic regression classifiers (denoted by 'γ-LR'), which minimize the decision boundary covariance subject to accuracy constraints over the entire dataset by solving problem (3.7) with increasing values of γ. Then, we train logistic regression classifiers (denoted by 'Fine-γ-LR') that minimize the decision boundary covariance subject to *fine-grained* accuracy constraints by solving problem (3.8). Here, we prevent the non-protected users that were classified as positive by the unconstrained logistic regression classifier from being classified as negative by constraining that their distance from decision boundary stays positive while learning the nondiscriminatory boundary. We then increase $\gamma_i = \gamma$ for the remaining users. In both cases, we increased the value of γ until we reach 0 disparate impact during training. Figure 3.7 summarizes the results for both datasets, by showing (a) the average accuracy (solid curves) and disparate impact (dashed curves) against γ, and (b) the percentage of non-protected (N-P, solid curves) and protected (P, dashed curves) users in the positive class against γ. We observe that, as we increase γ, the classifiers that constrain the overall training loss (γ-LR) remove non-protected users from the positive class and add protected users to the positive class, in contrast, the classifiers that prevent the non-protected users that were classified as positive in the unconstrained classifier from being classified as negative (Fine-γ-LR) add both protected and non-protected users to the positive class. As a consequence, the latter achieves lower accuracy for the same value of disparate impact.

3.3 Discussion

In this chapter, we introduced a novel measure of decision boundary covariance, which enables us to ensure nondiscrimination with respect to one or more sensitive features, in terms of both disparate treatment and disparate impact, in a variety of linear and non-linear classifiers. We leverage this measure to derive two complementary formulations: one that maximizes accuracy subject to disparate impact constraints, and helps ensure compliance with a non-discrimination policy or law (*e.g.*, limiting disparity in positive class outcome rates between the groups below a certain threshold); and another one that minimized disparate impact subject to accuracy constraints, and ensures fulfilling certain business needs (*e.g.*, disparate impact's business necessity clause).

Comparison with related techniques reveal that our method provides an accuracy comparable to that of other methods for the same degree of disparate impact. Moreover, as compared to the post-processing technique, our method provides an additional flexibility that it can also operate without access to the sensitive feature at decision time. One could potentially "combine" the preferential sampling and the post-processing techniques by learning optimal group-conditional thresholds for removing disparate impact on the training dataset, re-labeling the training dataset according to these thresholds, and then training an accuracy-maximizing classifier on the relabeled dataset. Such a strategy could relax the post-processing technique's requirement of having access to the sensitive feature at decision time. Analysis of such combined techniques would be an interesting avenue for future work.

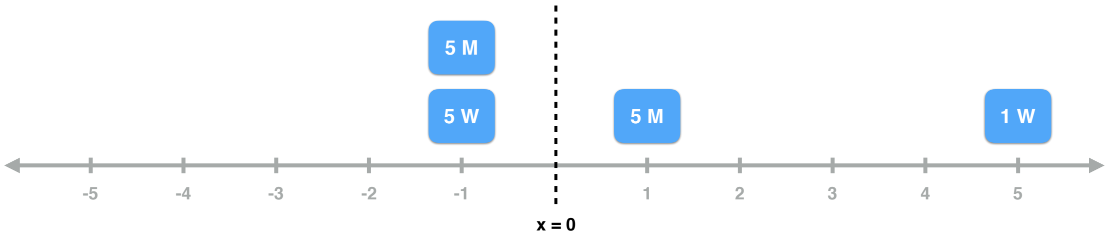


Figure 3.8: Covariance constraints may perform unfavorably in the presence of outliers. The figure shows a hypothetical dataset with just one feature (x) with values ranging form -5 to 5. Data points belong to two groups: men (M) or women (W). Each box shows the number of subjects of from a certain group (M or W) with that feature value. The decision boundary is at $x = 0$. The decision boundary covariance in this case is 0, yet the disparity in positive class outcome rates between men and women (0.5 for men and 0.17 for women) is very high. This situation is caused by one woman with feature value 5—this outlier point cancels out the effect of five normal examples (W with feature value -1) while computing the covariance.

On the negative side, we note that, as opposed to the post-processing scheme, our method does not always fully remove disparate impact, *i.e.*, it does not always drive the disparity in acceptance rates close to zero. Such situations can arise due to various reasons. First, since our mechanisms relies on empirically estimating the decision boundary covariance, very small presence of a certain group in the dataset can lead to poor estimate of the covariance and might not fully remove disparate impact. Furthermore, while the post-processing schemes to remove disparate impact operate on the data of dimensionality 1 (that is, the scalar score assigned to each item by the classifier), our method operates by using all the features used in classification in order to compute the decision boundary covariance. As a result, our method is expected to suffer more from the data sparsity problem.

We also notice that our method might not perform well in the presence of outliers. Consider for instance the example shown in Figure 3.8, where an outlier point causes the decision boundary covariance to be 0, even when the disparity in positive class outcomes caused by the corresponding decision boundary is very high. However, such outliers can in fact deteriorate the performance of any learning task (Bishop, 2006), even when no other constraints are applied, and one might wish to remove such outliers before training any classification model.

Also, while we note that a decreasing covariance threshold corresponds to a decreasing degree of disparate impact (Eq. 3.9), the relation between the two is only empirically observed. A precise mapping between covariance and DI is quite challenging to derive analytically since it depends on the specific classifier and the dataset being used. Such a theoretical analysis would be an interesting future direction.

Finally, as we discussed in Section 2.2.2, a disparity in positive class outcome rates of different groups may not always result in a disparate impact liability. In other words, disparate impact is not always a suitable measure of nondiscrimination. We discuss examples, and ways to address such scenarios in the next chapter.

Chapter 4

Disparate mistreatment: A new measure of discrimination

While disparate impact is an intuitive interpretation of discrimination—especially in scenarios when the training data is suspected to be biased (Siegel, 2014)—in certain other scenarios, its utility can be quite limited. For example, consider *Ricci vs. DeStefano* (Ricci, 2009), the US Supreme Court case mentioned in Section 2.2.2. The court in this case found that since the promotion test was relevant to the job at hand, the apparent disparate impact in the selection outcomes would not cause a discrimination liability. In other words, the disparate impact in this instance would not be deemed as causing wrongful relative disadvantage. This case seems to suggest that in situations when one can ascertain the reliability of the decisions in the training data, disparate impact might *not* be a suitable interpretation (and measure) of wrongful relative disadvantage, and mitigating disparate impact in such cases can instead be interpreted as causing reverse discrimination.

To account for such situations, we propose an alternative measure of discrimination (or an interpretation of wrongful relative disadvantage), ***disparate mistreatment***,[13] especially well-suited for scenarios where ground truth is available for historical decisions used during the training phase. We call a decision making process to be suffering from disparate mistreatment with respect to a given sensitive feature (*e.g.*, race) if the *misclassification rates* (in contrast to beneficial outcome rates under disparate impact) differ for groups of people having different values of that sensitive feature (*e.g.*, African-Americans and whites). For example, in the case of the NYPD Stop-question-and-frisk program (SQF) (Meares, 2014) where pedestrians are stopped on the suspicion of possessing an illegal weapon (Goel et al., 2016), having different prediction accuracy (or equivalently, different misclassification rates) for different races would constitute a case of disparate

[13] In a concurrent work, Hardt et al. (2016) proposed a measure of discrimination which is in essence very close to disparate mistreatment. For details, see Section 4.2.

mistreatment. In this way, *disparate mistreatment interprets the disparity in misclassification rates as imposition of wrongful relative disadvantage*.

In addition to the *overall* misclassification rate in general, depending on the application scenario and the consequences of each type of misclassifications , one might want to measure disparate mistreatment with respect to *different kinds* of misclassification rates. For example, in pretrial risk assessments, the decision making process might only be required to ensure that the false positive rates are equal for all groups, since it may be more acceptable to let a guilty person go, rather than incarcerate an innocent person.[14] On the other hand, in loan approval systems, one might instead favor a decision making process in which the false negative rates are equal, to ensure that deserving (positive class) people with a certain sensitive feature value are not denied (negative class) loans disproportionately. Similarly, depending on the application scenario at hand, and the cost of the type of misclassification, one may choose to measure disparate mistreatment using false discovery and false omission rates, instead of false positive and false negative rates (detailed in Table 4.1).

To train classifiers that are free of disparate mistreatment, we extend our decision boundary covariance mechanism and propose a tractable proxy that can be included in the formulation of convex boundary-based classifiers as a convex-concave constraint. The resulting formulation can be solved efficiently using recent advances in convex-concave programming (Shen et al., 2016b).

Relevant publication

Results presented in this chapter are published in (Zafar et al., 2017a).

4.1 Differentiating disparate mistreatment from disparate treatment and disparate impact

In this section, we use an illustrative example to differentiate our newly proposed measure of disparate mistreatment from existing measures of disparate treatment and disparate impact.

Disparate mistreatment. Intuitively, disparate mistreatment can arise in any automated decision making system whose outputs (or decisions) are not perfectly (*i.e.*, 100%) accurate. For example, consider a decision making system that uses a logistic regression classifier to provide binary outputs (say, positive and negative) on a set of people. If

[14] *"It is better that ten guilty persons escape than that one innocent suffer"*—William Blackstone

User features		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

Figure 4.1: Decisions of three fictitious classifiers ($\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon. Gender is a sensitive feature, whereas the other two features (suspicious bulge in clothing and proximity to a crime scene) are non-sensitive. Ground truth on whether the person is actually in possession of an illegal weapon is also shown.

the items in the training data with positive and negative class labels are not linearly separable, as is often the case in many real-world application scenarios, the system will misclassify (*i.e.*, produce false positives, false negatives, or both, on) some people. In this context, the misclassification rates may be different for groups of people having different values of sensitive features (*e.g.*, males and females; African-Americans and whites) and thus disparate mistreatment may arise.

Figure 4.1 provides an example of decision making systems (classifiers) with and without disparate mistreatment. In all cases, the classifiers need to decide whether to stop a pedestrian—on the suspicion of possessing an illegal weapon—using a set of features such as bulge in clothing and proximity to a crime scene. The "ground truth" on whether a pedestrian actually possesses an illegal weapon is also shown. We show decisions made by three different classifiers $\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$. We deem $\mathbf{C_1}$ and $\mathbf{C_2}$ as discriminatory due to disparate mistreatment because their rate of erroneous decisions for males and females are different: $\mathbf{C_1}$ has different false negative rates for males and females (0.0 and 0.5, respectively), whereas $\mathbf{C_2}$ has different false positive rates (0.0 and 1.0) as well as different false negative rates (0.0 and 0.5) for males and females.

Disparate treatment. As described in Section 2.2.1, disparate treatment arises when a decision making system provides different outputs for groups of people with the same (or similar) values of non-sensitive features but different values of sensitive features.

In Figure 4.1, we deem $\mathbf{C_2}$ and $\mathbf{C_3}$ to be discriminatory due to disparate treatment since **C2**'s (**C3**'s) decisions for $Male$ 1 and $Female$ 1 ($Male$ 2 and $Female$ 2) are different even though they have the same values of non-sensitive features.

Disparate impact. Finally, as mentioned in Section 2.2.2, disparate impact arises when a decision making system provides outputs that benefit (hurt) a group of people sharing a value of sensitive feature more frequently than other groups of people.

In Figure 4.1, assuming that a pedestrian benefits from a decision of not being stopped, we deem $\mathbf{C}_1$ as discriminatory due to disparate impact because the fraction of males and females that were stopped are different (1.0 and 0.66, respectively).

4.1.1 Application scenarios for disparate impact vs. disparate mistreatment

Note that unlike in the case of disparate mistreatment, the notion of disparate impact is independent of the "ground truth" information about the decisions, *i.e.*, whether or not the decisions are correct or valid. Thus, the notion of disparate impact is particularly appealing in application scenarios where ground truth information for decisions does not exist and the historical decisions used during training are not reliable and thus cannot be trusted. Unreliability of historical decisions for automated decision making systems is particularly concerning in scenarios like recruiting or loan approvals, where biased judgments by humans in the past may be used when training classifiers for the future. In such application scenarios, it is hard to distinguish correct and incorrect decisions, making it hard to assess or use disparate mistreatment as a notion of discrimination.

However, in scenarios where ground truth information for decisions can be obtained, disparate impact can be quite misleading as a notion of discrimination. That is, in scenarios where the validity of decisions can be reliably ascertained, it would be possible to distinguish disproportionality in beneficial (or, desirable class) decision outcomes for sensitive feature groups that arises from justifiable reasons (*e.g.*, qualification of the candidates) and disproportionality that arises for non-justifiable reasons (*i.e.*, discrimination against certain groups). By requiring beneficial decision outcomes to be proportional (*i.e.*, requiring Eq. (2.5) to hold), the no-disparate-impact criterion risks introducing reverse-discrimination against qualified candidates. In contrast, when the correctness of decisions can be determined, disparate mistreatment can not only be accurately assessed, but its implementation as a discrimination measure (*i.e.*, its removal from decision making outcomes) can also avoid the above-mentioned reverse-discrimination, making it a more appealing notion of discrimination.

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Table 4.1: In addition to the overall misclassification rate, error rates can be measured in two different ways: false negative rate and false positive rate are defined as fractions over the *class distribution in the ground truth labels*, or true labels. On the other hand, false discovery rate and false omission rate are defined as fractions over the *class distribution in the predicted labels*.

4.1.2 How does disparate mistreatment capture wrongful relative disadvantage?

Consider again the fictitious decision making task presented in Figure 4.1. In this task, classifier $\mathbf{C_1}$—which has a false positive rate of 0 for men and 1.0 for women—can be interpreted as imposing wrongful relative disadvantage on women since it can be thought of as wrongly stereotyping women as carrying an illegal weapon even when in reality they did not. Note that generating a false positive on persons from *any* group, men or women, could be thought of as an imposition of wrongful disadvantage. However, only when the false positive rates of the groups are *different* does the wrongful disadvantage become *relative*.[15]

4.2 Measuring disparate mistreatment

Using the formal setup described in Section 2.3, we now formalize disparate mistreatment in a classification task.

A binary classifier does not suffer from disparate mistreatment if the misclassification rates for different groups of people having different values of the sensitive feature z are the same. Table 4.1 describes various ways of measuring misclassification rates. Specifically, misclassification rates can be measured as fractions over the *class distribution in the ground truth labels*, *i.e.*, as false positive and false negative rates, or over the *class*

[15] Recall from Section 2.1 and Altman (2016) that discrimination is in inherently a relative phenomenon. We will discuss the cases of wrongful disadvantage, without regard to the group membership in Section 6.3.

distribution in the predicted labels, *i.e.*, as false omission and false discovery rates. Consequently, the absence of disparate mistreatment in a binary classification task can be specified with respect to the different misclassification measures as follows:

overall misclassification rate (OMR):

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1), \tag{4.1}$$

false positive rate (FPR):

$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1), \tag{4.2}$$

false negative rate (FNR):

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1), \tag{4.3}$$

false omission rate (FOR):

$$P(\hat{y} \neq y | z = 0, \hat{y} = -1) = P(\hat{y} \neq y | z = 1, \hat{y} = -1), \tag{4.4}$$

false discovery rates (FDR):

$$P(\hat{y} \neq y | z = 0, \hat{y} = 1) = P(\hat{y} \neq y | z = 1, \hat{y} = 1). \tag{4.5}$$

Finally, in a concurrent work, Hardt et al. (2016) proposed measures of discrimination called "equal opportunity" and "equalized odds" which are in essence very similar to our measure(s) of disparate mistreatment. Specifically, a classifier satisfies equal opportunity if Eq. (4.3) holds, and it satisfies equalized odds if Eqs. (4.2-4.3) hold.

A brief detour into the impossibility of nondiscrimination.
In certain application scenarios, one might be interested in satisfying more than one type of nondiscrimination defined in Eqs. (2.5-2.6, 4.1-4.5).

Some recent works (Chouldechova, 2016; Friedler et al., 2016; Kleinberg et al., 2017) have investigated the impossibility of simultaneously satisfying multiple notions of nondiscrimination. Chouldechova (2016) and Kleinberg et al. (2017) show that, when the fraction of users with positive class labels differ between members of different sensitive feature groups, it is impossible to construct classifiers that are equally *well-calibrated* (where well-calibration essentially measures the false discovery and false omission rates of a classifier) and also satisfy the equal false positive and false negative rate criterion (except for a "dumb" classifier that assign all examples to a single class). These results suggest that satisfying all five criterion of disparate mistreatment (Table 4.1) simultaneously is impossible when the underlying distribution of data is different for different groups. Kleinberg et al. (2017) also show the impossibility of simultaneously satisfying disparate impact and disparate mistreatment. However, in practice, it may still be interesting to explore the best, even if imperfect, extent of nondiscrimination a classifier can achieve.

4.3 Training classifiers free of disparate mistreatment

In this section, we devise a mechanism to train classifiers free of disparate mistreatment when it is defined in terms of overall misclassification rate, false positive rate and false negative rate, *i.e.*, Eqs. (4.1-4.3).

To train such a classifier, one could incorporate the appropriate condition from Eqs. (4.1-4.3) (based on which kind of misclassifications disparate mistreatment is being defined for) into the classifier formulation. For example, in order to remove disparity in overall misclassification rates (Eq. (4.1)), one could solve the following optimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\ & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon, \end{array} \tag{4.6}$$

where $\epsilon \in \mathbb{R}^+$ controls the extent of disparate mistreatment.

However, since the conditions in Eqs. (4.1-4.3) are, in general, non-convex, solving the constrained optimization problem defined by (4.6) seems difficult.

To overcome the above difficulty, we propose a tractable proxy, inspired by our disparate impact proxy in Section 3.1. In particular, we propose to measure disparate mistreatment using the covariance between the users' sensitive features and the signed distance between the feature vectors of *misclassified* users and the classifier decision boundary, *i.e.*:

$$\text{Cov}_{OMR}(z, g_{\boldsymbol{\theta}}(y, \boldsymbol{x})) = \mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \boldsymbol{x}))] \tag{4.7}$$

$$\approx \frac{1}{N} \sum_{(\boldsymbol{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \boldsymbol{x}), \tag{4.8}$$

where $g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) = \min(0, y d_{\boldsymbol{\theta}}(\boldsymbol{x}))$ and the term $\mathbb{E}[(z - \bar{z})]\bar{g}_{\boldsymbol{\theta}}(\boldsymbol{x})$ cancels out since $\mathbb{E}[(z - \bar{z})] = 0$.

As in the case of disparate impact, if a decision boundary satisfies Eq. (4.1), then the (empirical) covariance defined above will be (approximately) zero (for a sufficiently large training set) and we can train a classifier free of disparate mistreatment with respect to overall misclassification rate by replacing the (intractable) constraint in Eq. (4.6) by an alternative constraint as follows:

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimize}} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N}\textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c, \\ & \tfrac{1}{N}\textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{aligned} \tag{4.9}$$

where $c \in \mathbb{R}^+$ is a given threshold, which trades off accuracy and disparate mistreatment.

Similarly, we can define the above covariance measure for disparate mistreatment with respect to false positive rates, false negative rates, false omission rates or false discovery rates. For example, for false positive rates, one needs to consider the set of *misclassified* users with (ground-truth) negative labels ($\mathcal{D}^-$), *i.e.*,

$$\text{Cov}_{FPR}(z, g_{\boldsymbol{\theta}}(y,\boldsymbol{x})) \approx \frac{1}{N^-} \sum_{(\boldsymbol{x},y,z)\in\mathcal{D}^-} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}), \tag{4.10}$$

where N^- represents the size of $\mathcal{D}^-$.

However, in contrast with the covariance measure in the case of disparate impact, defined by Eq. (3.2), the above covariance measures are not convex. Fortunately, the covariance constraints for disparate mistreatment with respect to overall misclassification rates, false positive rates and false negative rates can be easily converted into convex-concave constraints, which can be solved efficiently by using recent advances in convex-concave programming (Shen et al., 2016b), as follows.

Consider the constraints in Eq. (4.9), *i.e.*,

$$\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \quad \boxdot \quad c,$$

where '$\boxdot$' denotes '$\geq$' and '$\leq$' and, without loss of generality, we left out the constant term $\frac{1}{N}$. Then, we can split the sum in the above expression into two terms:

$$\sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} (0-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} (1-\bar{z})\, g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \quad \boxdot \quad c, \tag{4.11}$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ are the subsets of the training dataset $\mathcal{D}$ taking values $z = 0$ and $z = 1$, respectively. Define $N_0 = |\mathcal{D}_0|$ and $N_1 = |\mathcal{D}_1|$, then one can write $\bar{z} = \frac{(0\times N_0)+(1\times N_1)}{N} = \frac{N_1}{N}$

and rewrite Eq. (4.11) as:

$$\frac{-N_1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \sim c,$$

which, given that $g_{\boldsymbol{\theta}}(y,\boldsymbol{x})$ is convex in $\boldsymbol{\theta}$, results into a convex-concave (or, difference of convex) function.

Finally, we can rewrite the problem defined by (4.9) as:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c \\ & \frac{-N_1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{array} \tag{4.12}$$

which is a Disciplined Convex-Concave Program (DCCP) for any convex loss $L(\boldsymbol{\theta})$, and can be efficiently solved using well-known heuristics (Shen et al., 2016b).

Proceeding similarly, we can convert the covariance constraints for disparate mistreatment with respect to false positive rates and false negative rates to convex-concave constraints. For example, Eq. (4.6) can be rewritten to reduce disparity in false positive rates as:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\text{minimize}} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c \\ & \frac{-N_1^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{N_0^-}{N^-} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^-} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c, \end{array} \tag{4.13}$$

where $\mathcal{D}_i^-$ is the subset of the training data with $z = i$ and $y = -1$, and $N_i^- = |\mathcal{D}_i^-|$.

Note that unlike in the publication corresponding to this chapter (Zafar et al., 2017a), we define the false positive rate covariance (Eq. (4.10)) only over the ground truth negative dataset instead of the whole dataset. In cases where $\frac{N_0^-}{N_0} \neq \frac{N_1^-}{N_1}$ (or in other words, the base-rates are different for the two sensitive feature groups), the false positive rate covariance as defined by Zafar et al. (2017a) would not fully remove disparate mistreatment.

While the covariance constraints for disparate mistreatment with respect to false omission and false discovery rates can be readily defined, the corresponding constraints cannot be easily converted into convex-concave constraints. Handling such constraints efficiently is left as an interesting avenue for future work.

Finally, just like the disparate impact-free formulation (Section 3.1.2), the above formulation for removing disparate mistreatment provides the flexibility to remove disparate treatment as well. That is, since our formulation does not require the sensitive feature information at decision time, by keeping the features $\boldsymbol{x}$ disjoint from the sensitive feature z, one can remove disparate mistreatment and disparate treatment simultaneously.

Next, we particularize the formulation given by (4.12) for a logistic regression classifier (Bishop, 2006).

Logistic regression without disparate mistreatment. The disparate mistreatment constraints, when disparate mistreatment is defined in terms of false negative rates, can be particularized for a logistic regression classifier as:

$$\begin{array}{lll} \underset{\boldsymbol{\theta}}{\text{minimize}} & -\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) & \left.\vphantom{\sum}\right\} \text{LR formulation} \\ \text{subject to} & \left.\begin{array}{l} \frac{-N_1^+}{N^+} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^+} \min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \\ +\frac{N_0^+}{N^+} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^+} \min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \leq c, \\ \frac{-N_1^+}{N^+} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_0^+} \min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \\ +\frac{N_0^+}{N^+} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_1^+} \min(0, y\boldsymbol{\theta}^T\boldsymbol{x}) \geq -c, \end{array}\right\} & \begin{array}{c}\text{Disparate}\\ \text{mistreatment}\\ \text{constraints}\end{array} \end{array} \tag{4.14}$$

4.4 Evaluation

In this section, we conduct experiments on synthetic as well as real world datasets to evaluate the effectiveness of our scheme in controlling disparate mistreatment. To this end, we first generate several *synthetic* datasets that illustrate different variations of disparate mistreatment and show that our method can effectively remove disparate mistreatment in each of the variations, often at a small cost in accuracy. We then conduct experiments on two real world datasets. In both the synthetic and real-world datasets, we compare the performance of our scheme with different competing methods.

For this evaluation, we aim at removing disparate mistreatment when it is defined in terms of false positive rates (Eq. (4.2)) and false negative rates (Eq. (4.3)). Specifically, we measure the degree of disparate mistreatment as:

$$DM_{FPR} = P(\hat{y} \neq y|z = 0, y = -1) - P(\hat{y} \neq y|z = 1, y = -1), \tag{4.15}$$

$$DM_{FNR} = P(\hat{y} \neq y|z = 0, y = 1) - P(\hat{y} \neq y|z = 1, y = 1), \tag{4.16}$$

where the closer the values of DM_{FPR} and DM_{FNR} to 0, the lower the degree of disparate mistreatment. Note that unlike in the case of disparate impact in Eq. (3.9), we do not use the absolute difference while quantifying disparate mistreatment. As we later show in this section, the (in)equality in the signs of DM_{FPR} and DM_{FNR} carries significant consequences when considering disparate mistreatment with respect to false positive rate and false negative rate simultaneously. In such cases, the sign of the differences should also be taken into account.

4.4.1 Synthetic datasets

In this section, we empirically study the trade-off between nondiscrimination and accuracy in a classifier that suffers from disparate mistreatment. However, disparate mistreatment can arise in multiple different ways, as detailed below. To study these different situations, we first start with a simple scenario in which the classifier suffers from disparate mistreatment in terms of *only* false positive rate *or* false negative rate. Then, we focus on a more complex scenario in which the classifier is discriminatory in terms of *both*.

Disparate mistreatment on *only* false positive rate *or* false negative rate

The first scenario considers a case where a classifier maximizing accuracy leads to disparate mistreatment in terms of only the false positive rate (false negative rate), while being nondiscriminatory with respect to false negative rate (false positive rate), *i.e.*, $DM_{FPR} \neq 0$ and $DM_{FNR} = 0$ (or, alternatively, $DM_{FPR} = 0$ and $DM_{FNR} \neq 0$).

To simulate this scenario, we generate $10{,}000$ binary class labels ($y \in \{-1, 1\}$) and corresponding sensitive feature values ($z \in \{0, 1\}$), both uniformly at random, and assign a two-dimensional user feature vector ($\boldsymbol{x}$) to each of the points. To ensure different distributions for negative classes of the two sensitive feature groups (so that the two groups have different false positive rates), the user feature vectors are sampled from the following distributions (we sample 2500 points from each distribution):

$$\begin{aligned}
p(\boldsymbol{x}|z = 0, y = 1) &= \mathcal{N}([2, 2], [3, 1; 1, 3]) \\
p(\boldsymbol{x}|z = 1, y = 1) &= \mathcal{N}([2, 2], [3, 1; 1, 3]) \\
p(\boldsymbol{x}|z = 0, y = -1) &= \mathcal{N}([1, 1], [3, 3; 1, 3]) \\
p(\boldsymbol{x}|z = 1, y = -1) &= \mathcal{N}([-2, -2], [3, 1; 1, 3]).
\end{aligned}$$

Next, we train a logistic regression classifier optimizing for accuracy on this data. The classifier is able to achieve an accuracy of 0.85. However, due to the differences in feature distributions for the two sensitive feature groups, it achieves $DM_{FNR} = 0.15 - 0.15 = 0$ and $DM_{FPR} = 0.25 - 0.04 = 0.21$, which constitutes a clear case of disparate mistreatment

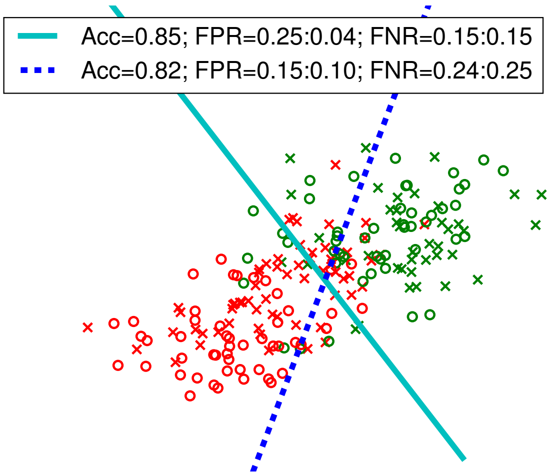


Figure 4.2: [Synthetic data with disparity only in false positive rates] The figure shows the original decision boundary (solid line) and nondiscriminatory decision boundary (dashed line), along with corresponding accuracy and false positive rates for groups $z = 0$ (crosses) and $z = 1$ (circles). Disparate mistreatment constraints cause the original decision boundary to rotate such that previously misclassified subjects with $z = 0$ are moved into the negative class (decreasing false positives), while well-classified subjects with $z = 1$ are moved into the positive class (increasing false positives), leading to similar false positive rates for both groups. The false negative rates disparity in this specific example stay unaffected.

in terms of false positive rate. We then train a logistic regression classifier subject to nondiscrimination constraints on false positive rate, with a covariance threshold $c = 0$.

Figure 4.2 shows the decision boundaries for both the unconstrained classifier (solid) and the classifier with constraints on disparate mistreatment (dashed). We observe that applying the disparate mistreatment constraint successfully causes the false positive rates for both groups ($z = 0$ and $z = 1$) to become similar, and hence, the outcomes of the classifier become more nondiscriminatory, *i.e.*, $DM_{FPR} \to 0$, while DM_{FNR} remains close to zero. We note that the invariance of DM_{FNR} may however change depending on the underlying distribution of the data.

Disparate mistreatment on *both* false positive rate and false negative rate

In this part, we consider a more complex scenario, where the outcomes of the classifier suffer from disparate mistreatment with respect to *both* false positive rate and false negative rate, *i.e.*, both DM_{FPR} and DM_{FNR} are non-zero. This scenario can in turn be split into two cases:

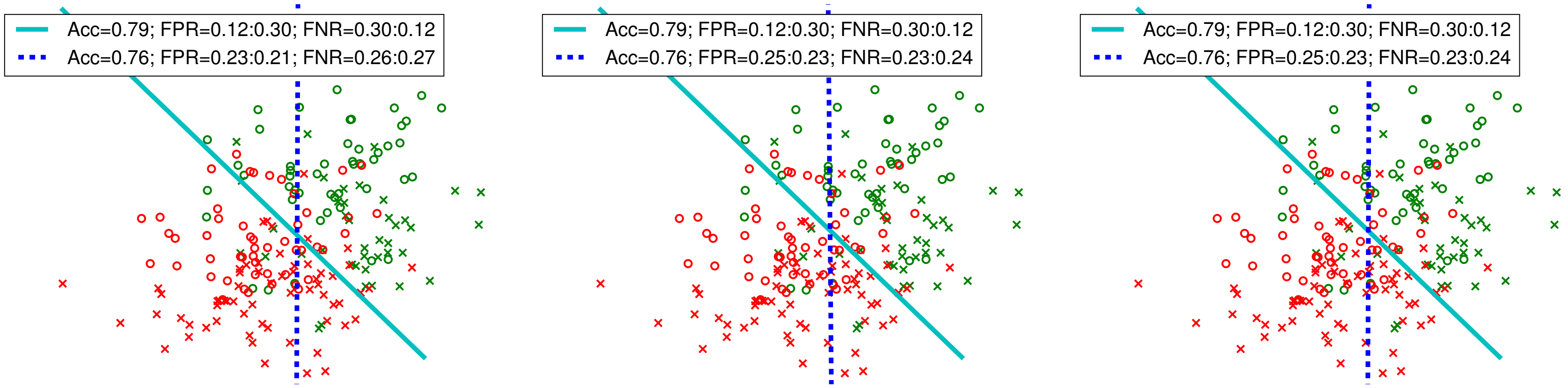
I. DM_{FPR} and DM_{FNR} have *opposite signs*, *i.e.*, the decision boundary disproportionately *favors* subjects from a certain sensitive feature group to be in the positive class (even when such assignments are misclassifications) while disproportionately assigning the subjects from the other group to the negative class. As a result, false positive rate for one group is higher than the other, while the false negative rate for the same group is lower.
II. DM_{FPR} and DM_{FNR} have the *same sign*, *i.e.*, both false positive as well as false negative rate are higher for a certain sensitive feature group. These cases might arise in scenarios when a certain group is harder to classify than the other.
Next, we experiment with each of the above cases separately.

— **Case I:** To simulate this scenario, we first generate $2{,}500$ samples from each of the following distributions:

$$
\begin{aligned}
p(\boldsymbol{x}|z=0,y=1) &= \mathcal{N}([2,0],[5,1;1,5]) \\
p(\boldsymbol{x}|z=1,y=1) &= \mathcal{N}([2,3],[5,1;1,5]) \\
p(\boldsymbol{x}|z=0,y=-1) &= \mathcal{N}([-1,-3],[5,1;1,5]) \\
p(\boldsymbol{x}|z=1,y=-1) &= \mathcal{N}([-1,0],[5,1;1,5])
\end{aligned}
$$

An accuracy-maximizing logistic regression classifier on this dataset attains an overall accuracy of 0.79 but leads to a false positive rate of 0.12 and 0.30 (*i.e.*, $DM_{FPR} = 0.12 - 0.30 = -0.18$) for the sensitive feature groups $z = 0$ and $z = 1$, respectively; and false negative rates of 0.30 and 0.12 (*i.e.*, $DM_{FNR} = 0.30 - 0.12 = 0.18$). To remove this disparate mistreatment, we train three different classifiers, with disparate mistreatment constraints on (i) false positive rates (ii) false negative rates and (iii) on both false positive and false negative rates.

Figure 4.3 summarizes the results for this scenario by showing the decision boundaries for the unconstrained classifier (solid) and the constrained nondiscriminatory classifiers. Here, we can observe several interesting patterns. First, removing disparate mistreatment on only false positive rate causes a rotation in the decision boundary to move previously *misclassified* subjects with $z = 1$ into the negative class, *decreasing* their false positive rate. However, in the process, it also moves previously *well-classified* subjects with $z = 1$ into the negative class, *increasing* their false negative rate. As a consequence, controlling disparate mistreatment on false positive rate (Figure 4.3(a)), also removes disparate mistreatment on false negative rate. A similar effect occurs when we control disparate mistreatment only with respect to the false negative rate (Figure 4.3(b)), and therefore, provides similar results as the constrained classifier for both false positive and false negative rates (Figure 4.3(c)). This effect is explained by the distribution of the

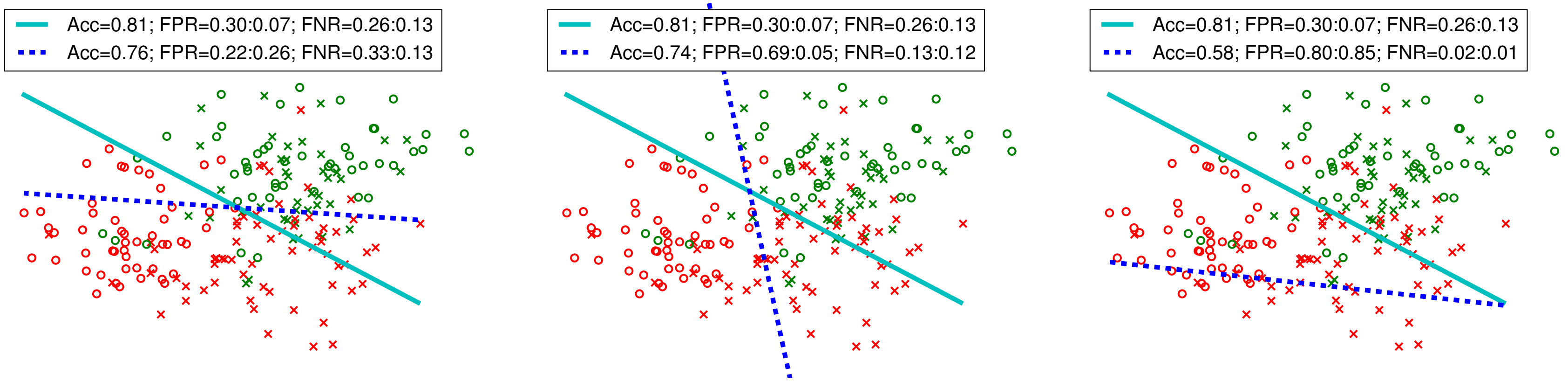


(a) FPR constraints

(b) FNR constraints

(c) Both constraints

Figure 4.3: [Synthetic data with disparity in false positive as well as false negative rates: DM_{FPR} and DM_{FNR} have opposite signs. Removing disparate mistreatment on FPR can potentially help remove disparate mistreatment on FNR. Removing disparate mistreatment on both at the same time leads to very similar results.



(a) FPR constraints

(b) FNR constraints

(c) Both constraints

Figure 4.4: [Synthetic data with disparity in false positive as well as false negative rates: DM_{FPR} and DM_{FNR} have the same sign. Removing disparate mistreatment on FPR can potentially increase disparate mistreatment on FNR. Removing disparate mistreatment on both at the same time causes a larger drop in accuracy.

data, where the centroids of the clusters for the group with $z = 0$ are shifted with respect to the ones for the group $z = 1$.

— **Case II:** To simulate the scenario where both DM_{FPR} and DM_{FNR} have the same sign, we generate $2{,}500$ samples from each of the following distributions:

$$
\begin{aligned}
p(\boldsymbol{x}|z=0, y=1) &= \mathcal{N}([1,2],[5,2;2,5]) \\
p(\boldsymbol{x}|z=1, y=1) &= \mathcal{N}([2,3],[10,1;1,4]) \\
p(\boldsymbol{x}|z=0, y=-1) &= \mathcal{N}([0,-1],[7,1;1,7]) \\
p(\boldsymbol{x}|z=1, y=-1) &= \mathcal{N}([-5,0],[5,1;1,5])
\end{aligned}
$$

We then train an accuracy-optimizing logistic regression classifier on this dataset. It attains an accuracy of 0.81 but leads to $DM_{FPR} = 0.30 - 0.07 = 0.23$ and $DM_{FNR} = 0.26 - 0.13 = 0.13$, resulting in disparate mistreatment in terms of both false positive and negative rates. Then, similarly to the previous scenario, we train three different kind of constrained classifiers to remove disparate mistreatment on (i) false positive rate, (ii) false negatives rate, and (iii) both.

Figure 4.4 summarizes the results by showing the decision boundaries for both the unconstrained classifiers (solid) and the constrained classifier (dashed) when controlling for disparate mistreatment with respect to false positive rate, false negative rate and both, respectively. We observe following noticeable patterns. First, controlling disparate mistreatment for only false positive rate (false negative rate), leads to a relatively minor drop in accuracy, but in contrast to Case I, can exacerbate the disparate mistreatment on false negative rate (false positive rate). For example, while the decision boundary is moved to control for disparate mistreatment on false negative rate, that is, to ensure that more subjects with $z = 0$ are well-classified in the positive class (reducing false negative rate), it also moves previously well-classified negative subjects into the positive class, hence increasing the false positive rate. A similar phenomenon occur when controlling disparate mistreatment with respect to only false positive rate. As a consequence, controlling for both types of disparate mistreatment simultaneously brings DM_{FPR} and DM_{FNR} close to zero, but causes a large drop in accuracy.

4.4.2 Real-world datasets

In this section, we experiment with two real-world datasets to test the effectiveness of our scheme in controlling disparate mistreatment. We also conduct comparisons with two different competing methods.

Datasets and experimental setup. We experiment with two real-world datasets: the ProPublica COMPAS risk assessment dataset (Larson et al., 2016a) and the NYPD stop-question-and-frisk (SQF) dataset (Stop, Question and Frisk Data, 2017).

The ProPublica COMPAS dataset consists of data about $7,215$ pretrial criminal defendants, and contains a number of features such as the age of the criminal defendant, number of prior criminal offenses *etc.*, and a class label indicating whether the person recidivated within two years or their arrest (positive class) or not (negative class). For more information about the data collection, we point the reader to a detailed description (Larson et al., 2016b) and some of the follow-up discussion on this dataset (Angwin and Larson, 2016; Flores et al., 2016). We designate race as the sensitive feature. Following ProPublica's analysis (Larson et al., 2016b), we only consider a subset of offenders whose race (the sensitive feature) is either African-American or white. Recidivism rates for the two groups are shown in Table A.4 in Appendix A. For modeling the classification task, we use the same set of features as used by ProPublica (Larson et al., 2016b).[16] After performing the filtering described above, we obtain $5,287$ subjects and 5 features.

The NYPD SQF dataset consists of $84,868$ pedestrians who were stopped in the year 2012 on the suspicion of having a weapon. The dataset also contains over 100 features (*e.g.*, gender, height, reason for stop) and a binary label which indicates whether (negative class) or not (positive class) a weapon was discovered. For our analysis, we consider the race to be the sensitive feature with values African-American and white. The classes in this dataset are highly imbalanced (97% of subjects in positive class), and as a result, a logistic regression classifier classifies almost all data points into the positive class. To counter this imbalance, we subsample the dataset to have equal number of subjects from each class. Information about weapon discovery rate for both races in included in Tables A.5 and A.6 in Appendix A. Furthermore, for training the classifiers, we consider the same set of features as Goel et al. (2016), with the exception that we exclude the highly sparse features 'precinct' and 'timestamp of the stop'. After performing these two filtering steps, we obtain $5,832$ subjects and 19 features.

Methods. In our experiments, we compare our approach to two baseline methods. More specifically, we consider the following methods:

[16] Notice that goal of this section is not to analyze the best set of features for recidivism prediction, rather, we focus on showing that our method can effectively remove disparate mistreatment in a given dataset. Hence, we chose to use the same set of features as used by ProPublica for their analysis. Moreover, since race is also included in this feature set, we additionally assume that *all* the methods have access to the sensitive features while making decisions. However, we will discuss the results of our method when operating without access to race as well.

- ***Our method***: Implements our scheme to avoid disparate treatment and disparate mistreatment *simultaneously*. Disparate mistreatment is avoided by using covariance constraints on false positive and / or false negative rates. Disparate treatment is avoided by ensuring that sensitive feature information is not used while making decisions, *i.e.*, by keeping user feature vectors ($\boldsymbol{x}$) and the sensitive features (z) disjoint.

- ***Our method$_{sen}$***: Implements our scheme to avoid disparate mistreatment only. The user feature vectors ($\boldsymbol{x}$) and the sensitive features (z) are not disjoint, that is, the classifier learns a non-zero weight for z. Therefore, the sensitive feature information is used for decision making, resulting in disparate treatment.

- ***Hardt et al.*** (Hardt et al., 2016): Operates by post-processing the outcomes of a possibly discriminatory classifier (logistic regression in this case) and using different decision thresholds for different sensitive feature value groups to remove disparate mistreatment. By construction, it needs the sensitive feature information while making decisions, and hence cannot avoid disparate treatment. This method is similar to the post-processing scheme discussed in (Corbett-Davies et al., 2017b).

- ***Baseline***: Baseline introduced by us to felicitate a second comparison method. Tries to remove disparate mistreatment by introducing different penalties for misclassified data points with different sensitive feature values during training phase. Specifically, it proceeds in two steps. First, it trains a possibly discriminatory classifier minimizing a loss function (*e.g.*, logistic loss) over the training data. Next, it selects the set of misclassified data points from the sensitive feature group that presents the higher error rate. For example, if one wants to remove disparate mistreatment with respect to false positive rate and $DM_{FPR} > 0$ (which means the false positive rate for points with $z = 0$ is higher than that of $z = 1$), it selects the set of misclassified data points in the training set having $z = 0$ and $y = -1$. Next, it iteratively re-trains the classifier with increasingly higher penalties on this set of data points until a certain level of nondiscrimination is achieved in the training set (until $DM_{FPR} \leq \epsilon$). The algorithm is summarized in Figure 1, particularized to remove disparate mistreatment defined in terms of false positive rate. This process can be intuitively extended to account for disparate mistreatment in terms of false negative rate or for *both* false positive rate and false negative rate. This method can be trained with or without using sensitive feature information while making decisions. We opt for the latter option.

Algorithm 1: Baseline method for removing disparate mistreatment with respect to FPR.

```
Input: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, z_i)\}_{i=1}^{N}$, $\Delta > 0$ $\epsilon > 0$
Output: Non-discriminatory baseline decision boundary $\boldsymbol{\theta}$
Initialize: Penalty $C = 1$
Train (potentially discriminatory) classifier $\boldsymbol{\theta} = \text{argmin}_{\boldsymbol{\theta}} \sum_{\mathbf{d} \in \mathcal{D}} L(\boldsymbol{\theta}, \mathbf{d})$
Compute $\hat{y}_i = \text{sign}(d_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$ and $DM_{FPR}$ on $\mathcal{D}$.
if $DM_{FPR} > 0$ then $s = 0$
else $s = 1$
$\mathcal{P} = \{\boldsymbol{x}_i, y_i, z_i | \hat{y} \neq y_i, z_i = s\}$, $\bar{\mathcal{P}} = \mathcal{D} \setminus \mathcal{P}$.
while $DM_{FPR} > \epsilon$ do
    Increase penalty: $C = C + \Delta$.
    $\boldsymbol{\theta} = \text{argmin}_{\boldsymbol{\theta}} \quad C \sum_{\mathbf{d} \in \mathcal{P}} L(\boldsymbol{\theta}, \mathbf{d}) + \sum_{\mathbf{d} \in \bar{\mathcal{P}}} L(\boldsymbol{\theta}, \mathbf{d})$
end
```

Results. First, we experiment with a standard logistic regression classifier optimizing for accuracy on both datasets. For the COMPAS dataset, the (unconstrained) logistic regression classifier leads to an accuracy of 0.664. However, the classifier yields false positive rates of 0.35 and 0.17, respectively, for African-Americans and whites (*i.e.*, $DM_{FPR} = 0.18$), and false negative rates of 0.32 and 0.61 (*i.e.*, $DM_{FNR} = -0.29$). These results constitute a clear case of disparate mistreatment in terms of both false positive rate and false negative rate. The classifier puts one group (African-Americans) at relative disadvantage by disproportionately misclassifying negative (did not recidivate) subjects from this group into the positive (did recidivate) class. This disproportional assignment results in a significantly higher false positive rate for African-Americans as compared to whites. On the other hand, the classifier puts the other group (whites) on a relative advantage by disproportionately misclassifying positive (did recidivate) subjects from this group into the negative (did not recidivate) class (resulting in a higher false negative rate). Note that this scenario resembles our synthetic example Case I in Section 4.4.1.

For the SQF data, the (unconstrained) logistic regression classifier leads to an accuracy of 0.751. However, the classifier yields false positive rates of 0.38 and 0.11, respectively, for African-Americans and whites (*i.e.*, $DM_{FPR} = 0.27$), and false negative rates of 0.19 and 0.31 (*i.e.*, $DM_{FNR} = -0.12$). Notice that unlike the COMPAS dataset, being classified positive here is an advantageous outcome—positive class in this case is not being stopped whereas the positive class in the COPMAS dataset is being classified as being a recidivist. This scenario also resembles our synthetic example Case I in Section 4.4.1.

Next, we apply our framework on a logistic regression classifier to eliminate disparate mistreatment with respect to false positive rate, false negative rate, and on both, and compare its performance with the two alternative methods. While controlling for

		FPR constraints			FNR constraints			Both constraints		
		Acc	D_{FPR}	D_{FNR}	Acc	D_{FPR}	D_{FNR}	Acc	D_{FPR}	D_{FNR}
ProPuclica COMPAS	Our method$_{\text{sen}}$	0.653	0.03	−0.10	0.656	−0.05	−0.01	0.654	−0.02	−0.03
	Baseline	0.631	0.01	−0.18	0.656	−0.03	−0.03	0.615	−0.19	0.13
	Hardt et al.	0.661	0.01	−0.08	0.654	−0.06	0.01	0.632	0.02	0.01
NYPD SQF	Our method	0.633	0.06	−0.01	0.705	0.22	−0.07	0.642	0.05	0.04
	Our method$_{\text{sen}}$	0.727	0.08	0.07	0.743	0.18	0.00	0.726	0.07	0.07
	Baseline	0.527	0.02	−0.08	0.734	0.14	0.01	0.435	−0.71	0.95
	Hardt et al.	0.725	0.03	0.12	0.734	0.14	0.04	0.722	0.02	0.06

Table 4.2: Performance of different methods while removing disparate mistreatment with respect to false positive rate, false negative rate and both. When provided with the same amount of information, our technique as well as the post-processing technique of Hardt et al. lead to similar accuracy for the same level of disparate mistreatment. The baseline tends to present the worst results.

disparate mistreatment with respect to FPR and FNR simultaneously, the method of Hardt et al. can be interpreted as finding the optimal point that minimizes the loss on the average of the two group-conditional ROC curves (one curve for each sensitive feature group), or the one that minimizes the loss on the point-wise minimum of the two curves. The optimal point in both cases lies on the point-wise minimum of the two curves. Both variants lead to similar performance, hence we report the results for the former.

Table 4.2 shows the results by showing the trade-off between disparate mistreatment and accuracy achieved by our method, the method by Hardt et al., and the baseline. Similarly to the results in Section 4.4.1, we observe that for all three methods, controlling for disparate mistreatment on false positive rate (false negative rate) also helps decrease disparate mistreatment on false negative rate (false positive rate), at least to some limited extent. Moreover, both our method and the method by Hardt et al. achieve similar accuracy for a given level of disparate mistreatment when provided with the same amount of information (sensitive feature information). We also note that the baseline tends to be somewhat unstable and fails to converge to a nondiscriminatory solution in some cases (*e.g.*, both FPR and FNR constraints on COMPAS and SQF datasets).

Finally, as noted in the beginning of this section, one of the five features considered for the ProPublica COMPAS dataset was race. As as a result, all the methods on this dataset lead to disparate treatment with respect to race since the final outcome of the classifiers directly depends on race. To avoid this disparate treatment, we also train our method on the remaining four features (number of prior offenses, age of the defendant, arrest charge and the degree of the charge) while excluding race.

In this case, while removing disparate mistreatment on false positive rate, our method achieves an accuracy of 0.603 and a DM_{FPR} of 0.06. While removing disparate mistreatment on false negative rate, our method achieves an accuracy of 0.616 and DM_{FNR} of -0.15. Applying constraints on both false positive and false negative rate does not lead to significant removal of disparate mistreatment as compared to the unconstrained classifier—the accuracy is 0.662 while DM_{FPR} and DM_{FNR} are 0.16 and -0.28, respectively.

These results show that for the COMPAS dataset, predictably (as in the case of SQF dataset), our method without access to the sensitive feature at the decision time sacrifices a greater amount of accuracy while removing disparate mistreatment as compared to the case when it has access to the sensitive feature (*i.e.*, **Our method$_{\textbf{sens}}$** in Table 4.2). Additionally, in the case of constraints on both false positive as well as false negative rates, our method without access to sensitive feature does not effectively remove disparate mistreatment. This observation would seem to suggest that using race as a feature would lead to a more effective removal of disparate mistreatment. However, we point out that we do not notice the same issue for the SQF dataset, or the synthetic datasets considered in Section 4.4.1. The problem here may also have been caused by the very small feature set available (only four features), and gathering a larger feature set might help alleviate this issue.

4.5 Discussion

In this chapter, we proposed a new measure of nondiscrimination, disparate mistreatment, that might be a more suitable measure of discrimination as compared to disparate impact in situations when one is learning from datasets with (unbiased) ground truth labels. We also propose mechanisms to remove disparate mistreatment from classification tasks, and compare the performance of our method with two competing techniques.

Experimental results show that when provided with the same amount of information, our method provides a similar accuracy as compared to competing methods for the same level of disparate mistreatment. Additionally, our method provides the possibility to remove both disparate mistreatment as well as disparate treatment simultaneously. However, this removing both kinds of discrimination would naturally lead to a lesser accuracy as compared to the cases when one is concerned with removing just one kind of discrimination.

We also note that our method for removing disparate mistreatment suffers from the similar limitations as the disparate impact-free classification method proposed in Chapter 3. Additionally, our formulation of training classifiers free of disparate mistreatment is not a convex program, but a disciplined convex-concave program (DCCP), which can be efficiently solved using heuristic-based methods (Shen et al., 2016b). While these methods are shown to work well in practice, unlike convex optimization, they do not provide any guarantees on the global optimality of the solution. In such cases, as is often suggested, one can solve the optimization problem with multiple random initialization points, and pick the solution with the best performance (Shen et al., 2016a).

Moreover, we note that in the case of controlling disparate mistreatment with respect to false positive or false negative rates, the corresponding covariance is computed only over the ground truth negative and ground truth positive datasets, respectively (Section 4.3). Since our method operates by estimating these covariances on the given training dataset, in cases when the training dataset consists of a very small negative or very small positive class, the corresponding covariance estimates might be inaccurate and as a result, our method might not be able to remove disparate mistreatment effectively. However, class imbalance problems are not specific to our method only, and are a well-known issue in general classification tasks as well (Bishop, 2006; Japkowicz, 2000).

CHAPTER 5

Discrimination beyond disparity: Preference-based measures of discrimination

Notice that the nondiscrimination measures examined until now quantify wrongful relative disadvantage through the absence of equality or parity (*e.g.*, parity of treatment in the case of disparate treatment, and parity of impact in the case of disparate impact). While the absence of parity is an intuitive way to capture wrongful relative disadvantage, we notice that some interpretations of the discrimination definition (Section 2.1) may argue otherwise. We describe two such interpretations below.

These new interpretations, which we refer to as ***preferred treatment*** and ***preferred impact*** are respectively motivated by the game theoretic notions of ***envy-freeness*** (Varian, 1974) and ***bargaining consensus*** (Nash Jr, 1950). At the core of these interpretations is the idea of *group preferences*: Given the choice between various sets of decision outcomes, any group of users would collectively *prefer* the set that contains *the largest fraction* (or the greatest number) of beneficial decision outcomes for that group.[17] Our new preference-based measures of nondiscrimination use the concept of user groups' preference as follows:

— **Preferred treatment.** A decision making system offers preferred treatment if every sensitive feature group (*e.g.*, men and women) *prefers* the set of decisions they receive over the set of decisions they would have received had they collectively presented themselves to the system as members of a different sensitive group. The preferred treatment interpretation is inspired by the game theoretic notion of envy-freeness. Under an envy-free system, all the parties involved in decision making prefer their own outcomes over the outcomes of the others—even when such outcomes are disparate. Here, preferred

[17] Although it is quite possible that certain *individuals* from the group may not prefer the set that maximizes the benefit for the *group as a whole*. See Section 5.1 for details.

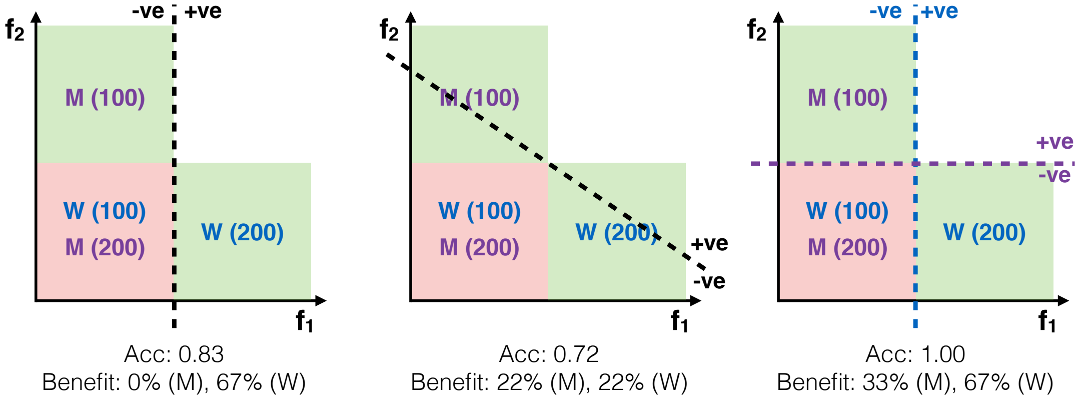


Figure 5.1: A fictitious decision making scenario involving two groups: men (M) and women (W). Feature f_1 (x-axis) is highly predictive for women whereas f_2 (y-axis) is highly predictive for men. Green (red) quadrants denote the positive (negative) class. Within each quadrant, the points are distributed uniformly and the numbers in parenthesis denote the number of subjects in that quadrant. The **left panel** shows the optimal classifier satisfying parity in treatment. This classifier leads to all the men getting classified as negative. The **middle panel** shows the optimal classifier satisfying parity in impact (in addition to parity in treatment). This classifier achieves impact parity by misclassifying women from positive class into negative class, and in the process, incurs a significant cost in terms of accuracy. The **right panel** shows a classifier consisting of group-conditional classifiers for men (purple) and women (blue). Both the classifiers satisfy the preferred treatment criterion since for each group, adopting the other group's classifier would lead to a smaller fraction of beneficial outcomes (refer to Section 5.1 for a discussion on group- vs. individual-level preferences). Additionally, this group-conditional classifier is also a preferred impact classifier since both groups get more benefit as compared to the impact parity classifier. The overall accuracy is better than the parity classifiers.

treatment interprets the presence of envy (where one group prefers another group's outcomes over their own) as imposition of wrongful relative disadvantage.

Notice that the preferred treatment interpretation represents a relaxation of treatment parity (or avoiding disparate treatment). That is, every decision making system that achieves treatment parity also satisfies preferred treatment, which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred treatment condition is at least as high as the one achieved under treatment parity. Additionally, preferred treatment allows group-conditional decision making (not allowed by treatment parity), which might be necessary to achieve high decision accuracy in scenarios when the predictive power of features varies greatly between different sensitive feature groups, as shown in Figure 5.1.

In this way, while preferred treatment is a looser interpretation of nondiscrimination than treatment parity, it retains a core nondiscrimination property embodied in treatment parity, namely, *envy-freeness at the level of user groups*. Under preferred treatment, no group of users (*e.g.*, men or women, African-Americans or whites) would feel that they would be collectively better off by switching their group membership (*e.g.*, gender, race). Thus, preferred treatment decision making, despite allowing disparate treatment, is not vulnerable to being characterized as "reverse discrimination" against, or "affirmative action" for certain groups.

— **Preferred impact.** A decision making system offers preferred impact if every sensitive feature group (*e.g.*, men and women) *prefers* the set of decisions they receive over the set of decisions they would have received under the criterion of impact parity (or avoiding disparate impact). The preferred impact interpretation is inspired by the bargaining problem in game theory where given some limited resources and a base resource allocation, two parties try to agree on a solution that maximizes their respective benefits beyond the base allocation (under the resource constraints).[18] For reaching a preferred impact solution, we take the solution satisfying impact parity to be the base allocation. Here, preferred impact interprets one or more groups not preferring their outcomes to the impact parity solution as imposition of wrongful relative disadvantage on those groups.

Note that the preferred impact criterion represents a relaxation of impact parity. That is, every decision making system that achieves impact parity also satisfies preferred impact, which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred impact condition is at least as high as the one achieved under impact parity. Additionally, preferred impact allows disparity in benefits received by different groups, which may be justified in scenarios where insisting on impact parity would only lead to a reduction in the beneficial outcomes received by one or more groups, without necessarily improving them for any other group (essentially resulting in non Pareto-optimal solutions). In such scenarios, insisting on impact parity can additionally lead to a reduction in the decision accuracy, creating a case of tragedy of impact parity with a worse decision making all round, as shown in Figure 5.1.

In this way, while preferred impact is a looser interpretation of nondiscrimination compared to impact parity, by guaranteeing that every group receives *at least* as many beneficial outcomes as they would have received under impact parity, it retains the core nondiscrimination gains in beneficial outcomes that the historically discriminated groups would have achieved under the nondiscrimination criterion of impact parity.

[18] If no agreement can be reached, then the parties resort to the base allocation .

In the rest of this chapter, we formally describe the preference-based notions of nondiscrimination. To enable decision making that satisfies the preferred treatment and preferred impact interpretations, we extend our decision boundary covariance mechanism and propose tractable proxies that can be encoded into the classifier formulations as convex-concave constraints. We show empirically on various synthetic and real-world datasets that preference-based measure of nondiscrimination can lead to significant gains in accuracy over parity-based measures, hence reducing the cost of nondiscrimination.

Relevant publication

Results presented in this chapter are published in (Zafar et al., 2017c).

5.1 Measures for preference-based nondiscrimination

We now formalize our preference-based measures of nondiscrimination. To that end, we first formalize the notion of group benefits, then revisit the parity-based measures of disparate treatment and disparate impact, and finally formalize the two preference-based measures.

Group benefit ($\mathcal{B}_z$) is the fraction of beneficial outcomes received by users sharing a certain value of the sensitive feature z (*e.g.*, females, males). For example, in a loan approval scenario, the beneficial outcome for a user may be receiving the loan and the group benefit for each value of z can be defined as:

$$\mathcal{B}_z(\boldsymbol{\theta}) = P(\hat{y} = 1 | \boldsymbol{\theta}, z) \tag{5.1}$$

Given this definition of groups benefits, one can re-write the absence of disparate impact—also formulated in Eq. (2.5)—in a classifier $\boldsymbol{\theta}$ as follows:

$$\mathcal{B}_z(\boldsymbol{\theta}) = \mathcal{B}_{z'}(\boldsymbol{\theta}) \quad \forall\, z, z' \in \mathcal{Z}, \tag{5.2}$$

i.e., the probability of the classifier assigning a beneficial outcome to all sensitive feature groups is the same.

In case one aims to train group-conditional classifiers (one classifier for each group), *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$, one can re-write the above parity impact condition as follows:

$$\mathcal{B}_z(\boldsymbol{\theta}_z) = \mathcal{B}_{z'}(\boldsymbol{\theta}_{z'}) \quad \forall\, z, z' \in \mathcal{Z}. \tag{5.3}$$

Satisfying disparate treatment, on the other hand, merely requires that the sensitive feature information is not used in decision making, *i.e.*: no group-conditional classifiers are used ($\boldsymbol{\theta}_z = \boldsymbol{\theta}_{z'} \ \forall\, z, z' \in \mathcal{Z}$) and that the classifier parameters do not include the sensitive feature z (z and $\boldsymbol{x}$ are disjoint sets).

Given the above metrics, we can formalize the preference-based nondiscrimination measures as follows:

Preferred treatment. If a classifier $\boldsymbol{\theta}$ resorts to group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z\in\mathcal{Z}}$, it is a preferred treatment classifier if each group sharing a sensitive feature value z benefits more from its corresponding group-conditional classifier $\boldsymbol{\theta}_z$ than it would benefit if it would be classified by any of the other group-conditional classifiers $\boldsymbol{\theta}_{z'}$, *i.e.*,

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}_{z'}) \quad \forall\, z, z' \in \mathcal{Z}. \tag{5.4}$$

Note that, if a classifier $\boldsymbol{\theta}$ does not resort to group-conditional classifiers, *i.e.*, $\boldsymbol{\theta}_z = \boldsymbol{\theta}$ for all $z \in \mathcal{Z}$, it will be always be a preferred treatment classifier. That is, *a classifier satisfying treatment parity criterion will also be a preferred treatment classifier.* This shows that the set of treatment parity classifiers is a subset of preferred treatment classifiers. In other words, a preferred treatment classifier (in theory) can always have an accuracy which is *at least* as good as that of a treatment parity classifier.

Preferred impact. A classifier $\boldsymbol{\theta}$ offers preferred impact over a classifier $\boldsymbol{\theta}'$ ensuring impact parity if it achieves higher group benefit for each sensitive feature value group, *i.e.*,

$$\mathcal{B}_z(\boldsymbol{\theta}) \geq \mathcal{B}_z(\boldsymbol{\theta}') \quad \forall\, z \in \mathcal{Z}. \tag{5.5}$$

One can also rewrite the above condition for group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z\in\mathcal{Z}}$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_z\}_{z\in\mathcal{Z}}$, as follows:

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}'_z) \quad \forall\, z \in \mathcal{Z}. \tag{5.6}$$

Again, note that *a classifier that satisfies the impact parity condition will also be a preferred impact classifier*. Following this reasoning, it is easy to show that the set of impact parity classifiers is a subset of preferred impact classifiers, and consequently, a preferred impact classifier (in theory), can always achieve *at least* as high an accuracy as the impact parity classifier.

Connection to the fair division literature. Our notion of preferred treatment is inspired by the concept of envy-freeness (Berliant and Thomson, 1992; Varian, 1974) in the fair division literature. Intuitively, an envy-free resource division ensures that no user would *prefer the resources allocated* to another user over their own allocation. Similarly, our notion of preferred treatment ensures envy-free decision making at the level of sensitive feature groups. Specifically, with preferred treatment classification, no sensitive feature group would *prefer the outcomes from the classifier* of another group.

Our notion of preferred impact draws inspiration from the two-person bargaining problem (Nash Jr, 1950) in the fair division literature. In a bargaining scenario, given a base resource allocation (also called the disagreement point), two parties try to divide some additional resources between themselves. If the parties cannot agree on a division, no party gets the additional resources, and both would only get the allocation specified by the disagreement point. Taking the resources to be the beneficial outcomes, and the disagreement point to be the allocation specified by the impact parity classifier, a preferred impact classifier offers enhanced benefits to all the sensitive feature groups. Put differently, the group benefits provided by the preferred impact classifier Pareto-dominate the benefits provided by the impact parity classifier.

On individual-level preferences. Notice that preferred treatment and preferred impact notions are defined based on the group preferences, *i.e.*, whether a *group as a whole* prefers (or, gets more benefits from) a given set of outcomes over another set. However, it is quite possible that a set of outcomes preferred by the group collectively is not preferred by certain *individuals* in the group. An example of such a setup is provided in Figure 5.2, where even though the classifier is a preferred treatment classifier for men at a group-level, it is not a preferred treatment classifier for men at an individual-level.[19] Consequently, one can extend our proposed notions to account for individual preferences as well, *i.e.*, a set of outcomes is preferred over another if *all* the individuals in the group prefer it. In the remainder of the paper, we focus on preferred treatment and preferred impact in the context of group preferences, and leave the case of individual preferences and its implications on the cost of achieving nondiscrimination for future work.

5.1.1 How do preference-based measures capture wrongful relative disadvantage?

As described earlier in this chapter, preferred treatment and preferred impact are inspired by game theoretic notions of envy-freeness and bargaining consensus.

In this context, a decision making process that does not ensure preferred treatment—*i.e.*, where one or more groups are envious of another group's outcomes—can be inter-

[19] On the other hand, the classifier in Figure 5.1 is not only a preferred treatment classifier (for both men and women) at a group-level, but it is also one at the level of the individuals—since no individual from either group would prefer the other group's classifier.

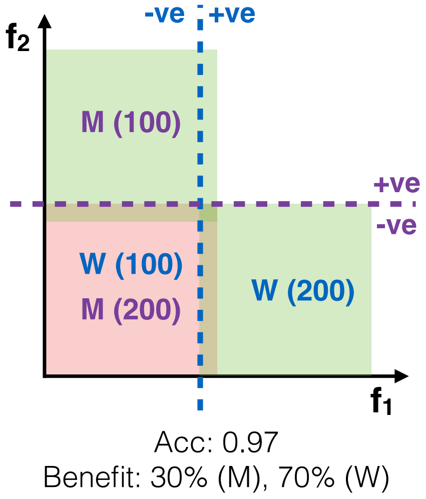


Figure 5.2: [Individual vs. group-level preferences] A fictitious dataset with optimal (group-conditional) classifiers. This dataset is a slight variant of the one in Figure 5.1, with the difference being that the (positive and negative) classes are not perfectly separable in this case (even with group-conditional classifiers). On this dataset, 30% of the men receive beneficial outcomes with their own classifier whereas 10% receive beneficial outcomes with the classifier of women. So the preferred treatment criterion (for group-level preferences) is satisfied, as men would prefer their own classifier *as a group*. However, some of the men who did not receive beneficial outcomes under their own classifier, receive beneficial outcomes when using the classifier of women, *i.e.*, the men inside the bottom left (red) quadrant who are on the right side of the classifier for women (blue line). So these men would *individually* prefer women's classifier, even though the men's group as a whole prefers their own classifier. Hence, while this setup provides preferred treatment for men at a group-level, it does not provide preferred treatment at an individual-level. (For women, the setup provides preferred treatment both at a group as well as at an individual-level.)

preted as causing wrongful relative disadvantage on the groups that get the envious outcomes, as these groups feel that they would get better outcomes had they been the part of another group (with all other features being the same).

On the other hand, a decision making process that does not ensure preferred impact—*i.e.*, one or more groups get lower benefits than the impact parity solution (or the base allocation)—can be interpreted as causing wrongful relative disadvantage since it denies the groups in questions of the nondiscrimination gains that they would have received under an impact parity situation by decreasing their beneficial outcomes even further.

5.2 Mechanisms for training classifiers with preferred treatment & preferred impact

For training a classifier adhering to preferred treatment and preferred impact, one can add the appropriate condition from Eqs. (5.4) and (5.6) in the classifier formulation.

For example, one can train a preferred impact classifier as follows:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} & \sum_{\boldsymbol{\theta}_z \in \mathcal{Z}} L(\boldsymbol{\theta}_z) \\ \text{subject to} & \mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}'_z) \quad \text{for all } z \in \mathcal{Z}, \end{array} \tag{5.7}$$

where $\mathcal{D}_z = \{(\boldsymbol{x}_i, y_i, z_i) \in \mathcal{D} | z_i = z\}$ denotes the set of users in the training set sharing the sensitive feature value z. The constant term $\mathcal{B}_z(\boldsymbol{\theta}'_z)$ is the group benefits for group z as defined by Eq. (5.1) and can be precomputed for a given parity impact classifier.

Unfortunately, it is quite challenging to solve the above optimization problem since the constraints (specified using probabilities defined in Eq. (5.1)) are non-convex for many well-known classifiers (*e.g.*, SVM). Hence, we approximate the group benefits by using the ramp (convex) function $r(a) = \max(0, a)$. The optimization problem hence becomes:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} & \sum_{\boldsymbol{\theta}_z \in \mathcal{Z}} L(\boldsymbol{\theta}_z) \\ \text{subject to} & \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, {\boldsymbol{\theta}'_z}^T \boldsymbol{x}) \quad \text{for all } z \in \mathcal{Z}, \end{array} \tag{5.8}$$

which is a disciplined convex-concave program (DCCP) for convex decision boundary-based classifiers and can be efficiently solved using well-known heuristics such as the one proposed by Shen et al. (2016b). For example, if we particularize the above formulation to group-conditional (standard) logistic regression classifiers $\boldsymbol{\theta}'_z$ and $\boldsymbol{\theta}_z$ and L_2-norm regularizer, then, Eq. (5.8) adopts the following form:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} & -\frac{1}{N} \sum_{(\boldsymbol{x}, y, z) \in \mathcal{D}} \log p(y | \boldsymbol{x}, \boldsymbol{\theta}_z) + \sum_{z \in \mathcal{Z}} \lambda_z ||\boldsymbol{\theta}_z||^2 \\ \text{subject to} & \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, {\boldsymbol{\theta}'_z}^T \boldsymbol{x}) \quad \text{for all } z \in \mathcal{Z}. \end{array} \tag{5.9}$$

where $p(y = 1 | \boldsymbol{x}, \boldsymbol{\theta}_z) = \frac{1}{1 + e^{-\boldsymbol{\theta}_z^T \boldsymbol{x}}}$. One can similarly particularize the formulation for other convex boundary-based classifiers like squared loss, linear / non-linear SVMs, *etc.*

One can train a preferred treatment classifier by solving the following optimization problem:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} & \sum_{\boldsymbol{\theta}_z \in \mathcal{Z}} L(\boldsymbol{\theta}_z) \\ \text{subject to} & \mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z' \in \mathcal{Z}. \end{array} \tag{5.10}$$

where the preferred treatment constraints, defined by Eq. (5.4), use empirical estimates of the group benefits, defined by Eq. (5.1). Note unlike in the case of preferred impact (Eq. (5.7)), in this case, both the left and right hand sides of the inequalities contain optimization variables.

However, the constraints in the above problem are non-convex and thus we adopt a similar strategy as in the case of preferred impact classifiers. More specifically, we solve instead the following tractable problem:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} & \sum_{\boldsymbol{\theta}_z \in \mathcal{Z}} L(\boldsymbol{\theta}_z) \\ \text{subject to} & \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, {\boldsymbol{\theta}_{z'}}^T \boldsymbol{x}) \quad \text{for all } z \in \mathcal{Z}, \end{array} \tag{5.11}$$

which is also a disciplined convex-concave program (DCCP) for convex boundary-based classifiers.

5.3 Evaluation

In this section, we compare the performance of preferred treatment and preferred impact classifiers against unconstrained, treatment parity and impact parity classifiers on a variety of synthetic and real-world datasets. More specifically, we consider the following classifiers, which we train to maximize utility subject to the corresponding constraints:

- ***Uncons***: an unconstrained classifier that resorts to group-conditional classifiers. It violates treatment parity—it trains a separate classifier per sensitive feature value group—and potentially violates impact parity—it may lead to different benefits for different groups.

- ***Parity***: a parity classifier that does not use the sensitive feature group information in the decision making, but only during the training phase, and is constrained to satisfy both treatment parity—its decisions do not change based on the users' sensitive feature value as it does not resort to group-conditional classifiers—and impact parity—it ensures that the benefits for all groups are the same. We train this classifier using the methodology proposed in Section 3.1.2.

- ***Preferred treatment***: a classifier that resorts to group-conditional classifiers and is constrained to satisfy preferred treatment—each group gets higher benefit with its own classifier than any other group's classifier.

- ***Preferred impact***: a classifier that resorts to group-conditional classifiers and is constrained to be preferred over the *Parity* classifier.

- ***Preferred both***: a classifier that resort to group-conditional classifiers and is constrained to satisfy both *preferred treatment* and *preferred impact*.

For the experiments in this section, we use logistic regression classifiers with L_2-norm regularization. We randomly split the corresponding dataset into 70%-30% train-test folds 5 times, and report the average accuracy and group benefits in the test folds.

5.3.1 Synthetic datasets

Experimental setup. Following the setup in Section 3.2.1, we generate a synthetic dataset in which the unconstrained classifier (*Uncons*) offers different benefits to each sensitive feature group. In particular, we generate $20{,}000$ binary class labels $y \in \{-1, 1\}$ uniformly at random along with their corresponding two-dimensional feature vectors sampled from the following Gaussian distributions:

$$p(\boldsymbol{x}|y = 1) = \mathcal{N}([2; 2], [5, 1; 1, 5])$$
$$p(\boldsymbol{x}|y = -1) = \mathcal{N}([-2; -2], [10, 1; 1, 3])$$

Then, we generate each sensitive feature from the Bernoulli distribution $p(z = 1) = p(\boldsymbol{x}'|y = 1)/(p(\boldsymbol{x}'|y = 1) + p(\boldsymbol{x}'|y = -1))$, where $\boldsymbol{x}'$ is a rotated version of $\boldsymbol{x}$, *i.e.*, $\boldsymbol{x}' = [\cos(\pi/8), -\sin(\pi/8); \sin(\pi/8), \cos(\pi/8)]$. Finally, we train the five classifiers described above and compute their overall (test) accuracy and (test) group benefits.

Results. Figure 5.3 shows the trained classifiers, along with their overall accuracy and group benefits. We can make several interesting observations:

The ***Uncons*** classifier leads to an accuracy of 0.87, however, the group-conditional boundaries and high disparity in treatment for the two groups (0.16 vs. 0.85) mean that it satisfies neither treatment parity nor impact parity. Moreover, it leads to only a small violation of preferred treatment—benefits for group-0 would increase slightly from 0.16 to 0.20 by adopting the classifier of group-1. However, this will not always be the case, as we will later show in the experiments on real data.

The ***Parity*** classifier satisfies both treatment and impact parity, however, it does so at a large cost in terms of accuracy, which drops from 0.87 for *Uncons* to 0.57 for *Parity*.

The ***Preferred treatment*** classifier (not shown in the figure), leads to a minor change in decision boundaries as compared to the *Uncons* classifier to achieve preferred treatment. Benefits for group-0 (group-1) with its own classifier are 0.20 (0.84) as compared to 0.17 (0.83) while using the classifier of group-1 (group-0). The accuracy of this classifier is 0.87.

The ***Preferred impact*** classifier, by making use of a looser notion of nondiscrimination compared to impact parity, provides higher benefits for both groups at a much smaller cost in terms of accuracy than the *Parity* classifier (0.76 vs. 0.57). Note that, while the *Parity* classifier achieved equality in benefits by misclassifying *negative examples from group-0* into the positive class and misclassifying *positive examples from group-1* into the negative class, the *Preferred impact* classifier only incurs the former type of misclassifications. However, the outcomes of the *Preferred impact* classifier do not satisfy the preferred treatment criterion: group-1 would attain higher benefit if it used the classifier of group-0 (0.96 as compared to 0.86).

Finally, the classifier that satisfies preferred treatment and preferred impact (***Preferred both***) achieves an accuracy and benefits at par with the *Preferred impact* classifier.

Next, experiment with a non linearly-separable dataset with a SVM classifier using radial basis function (RBF) kernel.

Following the setup of Section 3.1.2, we generated a synthetic dataset consisting of $4{,}000$ user binary class labels uniformly at random. We then assign a 2-dimensional user feature vector to each label by drawing samples from the following distributions:

$$p(\boldsymbol{x}|y=1,\beta) = \beta N([2;2],[5\ 1;1\ 5]) + (1-\beta)N([-2;-2],[10\ 1;1\ 3])$$
$$p(\boldsymbol{x}|y=-1,\beta) = \beta N([4;-4],[4\ 4;2\ 5]) + (1-\beta)N([-4;6],[6\ 2;2\ 3])$$

where $\beta \in \{0,1\}$ is sampled from Bernoulli(0.5). We then generate the corresponding user sensitive features z by applying the same rotation as for the synthetic dataset in Figure 5.3.

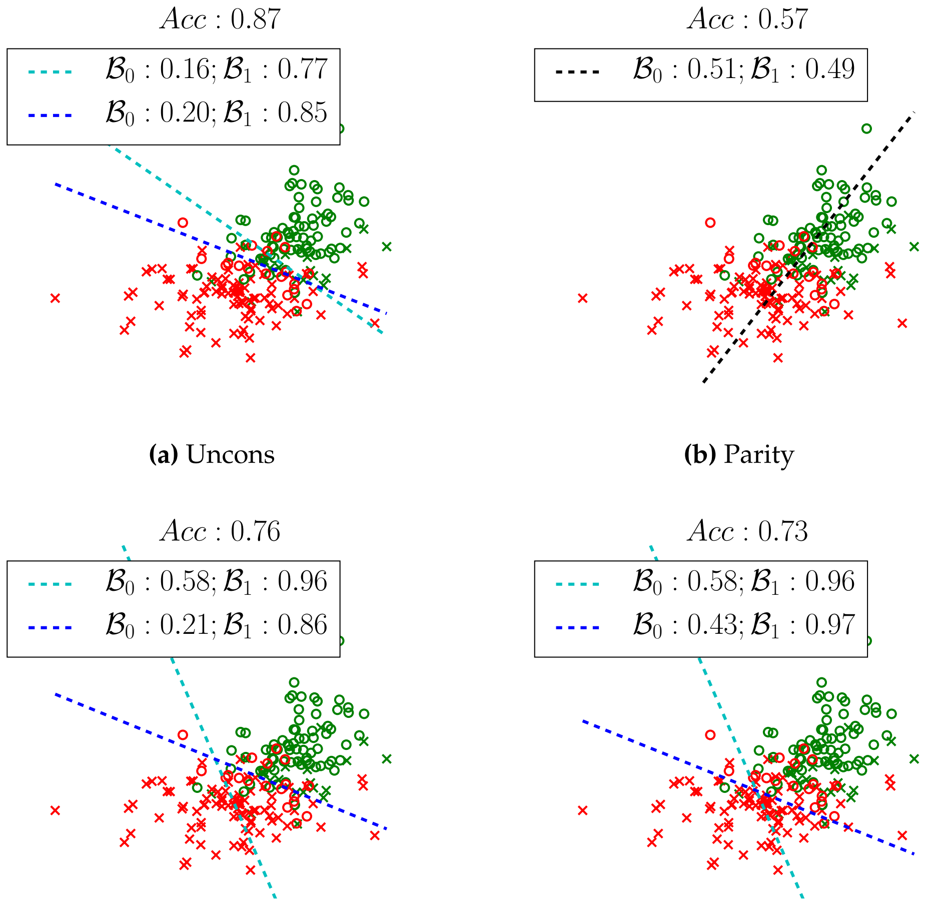
We then train the various classifiers described at the beginning of the section. The results are shown in Figure 5.4. Top row in the figure shows the group-conditional classifiers for group-0, whereas, the bottom row shows the ones for group-1. For the case of parity classifier, due to treatment parity condition, both groups use the same classifier.

The ***Uncons*** classifier leads to an accuracy of 0.96, however, the group-conditional classifiers lead to high disparity in beneficial outcomes for both groups (0.07 vs. 0.87). The classifier also leads to a violation of preferred treatment—the benefits for group-0 would increase from 0.07 with its own classifier to 0.17 with the classifier of group-1.

The ***Parity*** classifier satisfies both treatment and impact parity, however, it does so at a large cost in terms of accuracy, which drops from 0.96 for *Uncons* to 0.61 for *Parity*.

The ***Preferred treatment*** classifier, adjusts the decision boundary for group-0 to remove envy and does so at a small cost in accuracy (from 0.96 to 0.93).

The ***Preferred impact*** classifier, by making use of the relaxed parity-nondiscrimination conditions, provides higher or equal benefits for both groups at a much smaller cost in



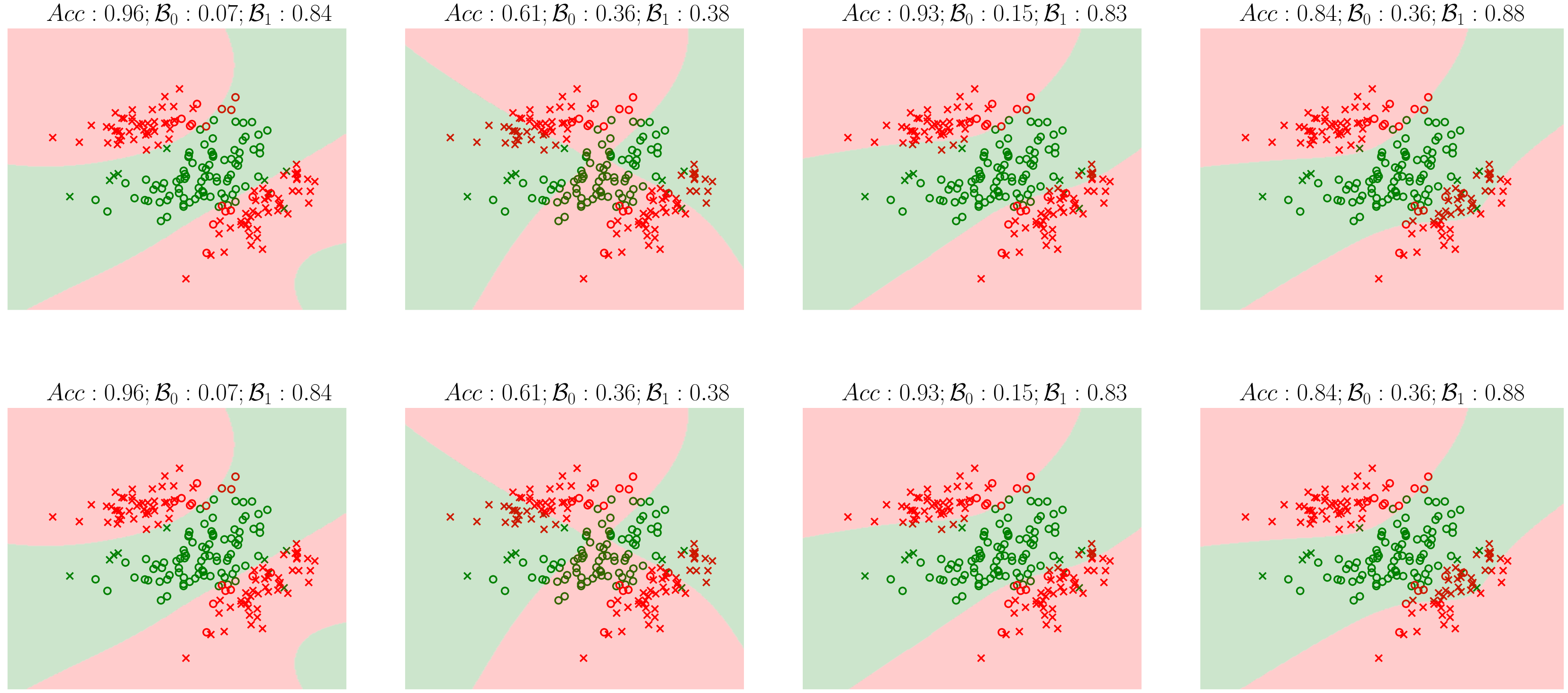
(a) Uncons

(b) Parity

(c) Preferred impact

(d) Preferred both

Figure 5.3: [Linearly separable synthetic data] Crosses denote group-0 (points with $z = 0$) and circles denote group-1. Green points belong to the positive class in the training data whereas red points belong to the negative class. Each panel shows the accuracy of the decision making scenario along with group benefits ($\mathcal{B}_0$ and $\mathcal{B}_1$) provided by each of the classifiers involved. For group-conditional classifiers, cyan (blue) line denotes the decision boundary for the classifier of group-0 (group-1). Parity case (panel (b)) consists of just one classifier for both groups in order to meet the treatment parity criterion. Preference-based measures can significantly lower the cost of nondiscrimination.



(a) Uncons

(b) Parity

(c) Uncons

(d) Parity

Figure 5.4: [Non- linearly-separable synthetic data] Crosses denote group-0 (points with $z = 0$) and circles denote group-1. Green points belong to the positive class in the training data whereas red points belong to the negative class. Each panel shows the classifiers with top row containing the classifiers for group-0 and the bottom for group-1, along with the overall accuracy as well as the group benefits ($\mathcal{B}_0$ and $\mathcal{B}_1$) provided by each of the classifiers involved. For parity classifier, no group-conditional classifiers are allowed, so both top and bottom row contain the same classifier.

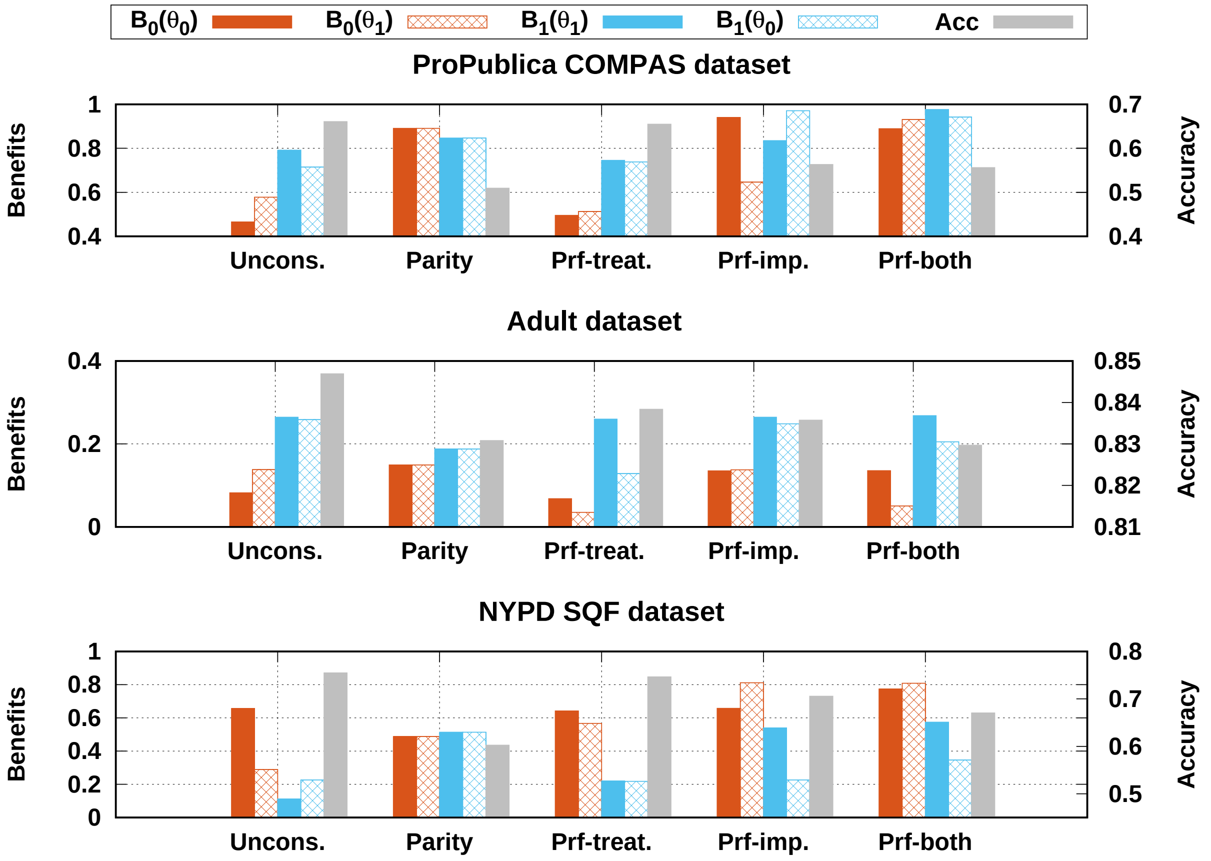


Figure 5.5: [Real-world datasets] The figure shows the accuracy and benefits received by the two groups for various decision making scenarios. 'Prf-treat.', 'Prf-imp.', and 'Prf-both' respectively correspond to the classifiers satisfying preferred treatment, preferred impact, and both preferred treatment and impact criteria. Sensitive feature values 0 and 1 denote blacks and whites in ProPublica COMPAS dataset and NYPD SQF datasets, and women and men in the Adult dataset. $\mathcal{B}_i(\boldsymbol{\theta}_j)$ denotes the benefits obtained by group i when using the classifier of group j. For the *Parity* case, we train just one classifier for both the groups, so the benefits do not change by adopting other group's classifier.

terms of accuracy than the *Parity* classifier (0.84 vs. 0.61). The preferred impact classifier in this case also satisfies the preferred treatment criterion.

5.3.2 Real-world datasets

We experiment with three real-world datasets: the COMPAS recidivism prediction dataset, the Adult income dataset, and the New York Police Department (NYPD) Stop-question-and-frisk (SQF) dataset.

Results. Figure 5.5 shows the accuracy achieved by the five classifiers described above along with the benefits they provide for the three datasets. We can draw several interesting observations:[20]

In all cases, the ***Uncons*** classifier, in addition to violating treatment parity (a separate classifier for each group) and impact parity (high disparity in group benefits), also violates the preferred treatment criterion (in all cases, at least one of group-0 or group-1 would benefit more by adopting the other group's classifier). On the other hand, the ***Parity*** classifier satisfies the treatment parity and impact parity but it does so at a large cost in terms of accuracy.

The ***Preferred treatment*** classifier provides a much higher accuracy than the *Parity* classifier—its accuracy is at par with that of the *Uncons* classifier—while satisfying the preferred treatment criterion. However, it does not meet the preferred impact criterion. The ***Preferred impact*** classifier meets the preferred impact criterion but does not always satisfy preferred treatment. Moreover, it also leads to a better accuracy then *Parity* classifier in all cases. However, the gain in accuracy is more substantial for the SQF datasets as compared to the COMPAS and Adult dataset.

The classifier satisfying preferred treatment and preferred impact (***Preferred both***) has a somewhat underwhelming performance in terms of accuracy for the Adult dataset. While the performance of this classifier is better than the *Parity* classifier in the COMPAS dataset and NYPD SQF dataset, it is slightly worse for the Adult dataset.

In summary, the above results show that ensuring either preferred treatment or preferred impact is less costly in terms of accuracy loss than ensuring parity-based nondiscrimination, however, ensuring both preferred treatment and preferred impact can lead to comparatively larger accuracy loss in certain datasets. We hypothesize that this loss in accuracy may be partly due to splitting the number of available samples into groups during training—each group-conditional classifier use only samples from the corresponding sensitive feature group—hence decreasing the effectiveness of empirical risk minimization.

[20] The directionality of discrimination in the SQF dataset is different from what one would expect (NY-CLU, 2018)—an unconstrained classifier gives more benefits to African-Americans as compared to whites. This is due to the fact that a larger fraction of stopped whites were found to be in possession on an illegal weapon (Tables A.5 and A.6 in Appendix A).

5.4 Discussion

In this chapter, we introduced two preference-based notions of nondiscrimination—preferred treatment and preferred impact—establishing a previously unexplored connection between discrimination-aware machine learning and the economics and game theoretic concepts of envy-freeness and bargaining. Then, we proposed tractable proxies to design boundary-based classifiers satisfying these notions and experimented with a variety of synthetic and real-world datasets, showing that preference-based nondiscrimination often allows for greater decision accuracy than existing parity-based notions.

Our work opens many promising avenues for future work. For example, our methodology, just like the previous chapters, is limited to convex boundary-based classifiers. A natural follow up would be to extend our methodology to other types of classifiers, *e.g.*, neural networks and decision trees.

Further refinements to our notions based on corresponding ideas from fair-division literature (*i.e.*, envy-freeness and bargaining) are also possible. For example, we defined preferred treatment and preferred impact in the context of group preferences, however, it would be worth revisiting the proposed definitions in the context of individual preferences (*e.g.*, envy-freeness at the level of individuals). Similarly, while we only explored group preferences without considering the qualifications of the users, one could extend these notions to take into account the qualifications when satisfying these preferences. For example, in envy-free rent division (Gal et al., 2016), while different users have certain room preferences, one aims at satisfying the envy-freeness criterion *while* taking into account the price each user is willing to pay for their preferred rooms (*i.e.*, one also considers the "user qualifications").

The fair division literature also establishes a variety of fairness axioms (Nash Jr, 1950) such as Pareto-optimality and scale invariance. It would be interesting to study such axioms in the context of discrimination-aware machine learning.

We also note that while moving from parity to preference-based nondiscrimination offers many attractive properties, we acknowledge it may not always be the most appropriate notion—in some scenarios, parity-based nondiscrimination may very well present the eventual goal and be more desirable. An example of such cases would be the diversity-enhancing schemes that aim at redressing historical discrimination by encouraging proportionality in beneficial outcome rates for different groups (MacCarthy, 2017; Siegel, 2014).

Chapter 6

Related work

In this chapter, we review work from various fields related to the area of discrimination-aware algorithmic decision making.

6.1 A brief overview of algorithmic decision making in social domains

Usage of algorithmic decision making in social domains has a long history.

For example, the first studies on usage of algorithmic decision making in predicting parole violations dates back to the 1920s. Hart (1923), Burgess (1928) and Tibbitts (1931) conducted one of the very first studies to evaluate the potential of predicting the risk of parole violation based on several related factors such as the type of offense committed by a defendant, employment status, *etc.*[21] The first examples of real-world deployment algorithms for predicting parole date back to the 1970s (Hoffman and Beck, 1974). Since then, the usage of algorithmic decision making in criminal risk assessment has risen significantly, with a number of jurisdictions in the US deploying automated software for risk prediction. For more details, we point the interested reader to Kehl and Kessler (2017). At this point, it is important to note that parole or recidivism risk assessment algorithms are mostly used as a tool to *assist* human decision makers, rather than *entirely replacing* them (Kehl and Kessler, 2017).

Similarly, the use of algorithmic decision making in credit scoring also goes back around six decades. For example, FICO scores have been being used in the US since the 1950s (FICO, 2018a). FICO (and similar) scores are used by a large number of financial institutions to assess the creditworthiness of their clients (FICO, 2018a) and are based on factors such as the payment history of the client, debt burden *etc.* (FICO, 2018b).

[21] Interestingly, "national or racial origin"—an attribute now regarded as protected—was also a factor in these early models (Tibbitts, 1931).

While certain applications such as criminal risk assessment and credit have a long history of usage of algorithmic decision making, the number of applications where algorithmic decision making is now being used to assist or replace human decision making has risen significantly in past few years. These applications span both offline as well as online worlds.

For example, in the offline world, predictive policing algorithms, such as PredPol, are increasingly being used across the US (Mohler et al., 2015; Perry, 2013). These algorithms operate by analyzing the historical data about crimes in a set of locations and determine how to allocate police officers in different locations to reduce crime. Usage of these algorithms follows the idea that concentrated police deployment in crime "hotspot areas" can help reduce crime (Mohler et al., 2015). In the online world, algorithmic decisions are also used for tasks such as matching potential job seekers with employers (Chandler, 2017; Posse, 2016; Woods, 2011) based on factors such as technical skills mentioned in resume, and recommending online content to web users (Covington et al., 2016; Graepel et al., 2010) based on factors such as users' query history.

6.2 Avoiding discrimination in classification

In this section, we will discuss techniques that aim to remove disparate treatment, disparate impact or disparate mistreatment from classification outcomes. To the best of our knowledge, no related techniques have been proposed to control for preferred treatment and preferred impact.

The first study on discrimination-free classification dates back to 2008 when Pedreschi et al. (2008) proposed techniques to avoid discrimination in classification rule mining. In the years that followed, a number of studies proposed techniques to remove discrimination from classification outcomes. Especially, last year or so has seen a flurry of methods proposed to control discrimination in classification. These studies operate by first specifying one or more measures of discrimination that they aim to control, *i.e.*, disparate treatment, disparate impact or disparate mistreatment, and then propose techniques to control for the selected measure(s).

These techniques can be divided into three different categories: *pre-processing*, *in-processing* and *post-processing*. Below, we discuss each of these categories separately.

6.2.1 Pre-processing

This technique consists of pre-processing the training data that would later be fed to a training algorithm (Calmon et al., 2017; Feldman et al., 2015; Kamiran and Calders, 2010;

Luong et al., 2011). The goal is to pre-process the training data such that *any* classification algorithm trained on this data would generate discrimination-free outcomes. This strategy can be roughly divided into two different sub-categories. Below, we briefly discuss these subcategories:

The first sub-category involves changing the values of class labels for certain data points (Kamiran and Calders, 2010; Luong et al., 2011). For example, Kamiran and Calders (2010) propose a pre-processing technique that operates by first training an unconstrained classifier, and then moving / duplicating the data points from the group with lower acceptance rate (as compared to the other group) until the classification outcomes are free of disparate impact.

The second sub-category involves perturbing the non-sensitive features (Feldman et al., 2015), or mapping the data to a transformed space (Calmon et al., 2017). For example, building on ideas in the area of privacy-preserving data analysis (specifically t-closeness), Feldman et al. (2015) "repair" the non-sensitive features such that it is impossible to predict the sensitive features from non-sensitive features (which in turn means that the classifier trained on this data will not incur disparate impact), while ensuring that the resulting distribution is close to the original data distribution.

On the plus side, the pre-processing techniques have an advantage that the transformed dataset can be used to train any downstream algorithm.

However, these techniques also suffer from some disadvantages. First, since these techniques are not optimized for any specific classification model, and treat the learning algorithm as a black box, as a consequence, the pre-processing can lead to unpredictable loss in accuracy or may not remove discrimination on the test data (as we saw in Section 3.2.2). Furthermore, transforming the dataset might also affect the explainability of the classifier—*e.g.*, since the feature values were transformed during pre-processing, the feature weights of a linear classifiers might not be interpretable anymore.

6.2.2 In-processing

The second strategy consists of modifying the training procedure of the classifier. Examples of this scheme include Calders and Verwer (2010); Goh et al. (2016); Kamiran et al. (2010); Kamishima et al. (2011); Quadrianto and Sharmanska (2017); Woodworth et al. (2017). Our proposed covariance constraints in Chapters 3 and 4 also fall under this category.

For example, the technique by Kamishima et al. (2011)—which is only limited to a logistic regression classifier—works by adding a regularization term in the objective that penalizes the mutual information between the sensitive feature and the classifier

Method	**Type**	**DT**	**DI**	**DM**	**BN**	**Polyvalent sens.**	**Multiple sens.**	**Range of classifiers**
Our framework (Chapters 3 and 4)	In	✓	✓	✓	✓	✓	✓	Any convex margin-based
Kamiran and Calders (2010)	Pre	✓	✓	✗	✗	✗	✗	Any score-based
Calders and Verwer (2010)	In/Post	✓	✓	✗	✗	✗	✗	Naive Bayes
Kamiran et al. (2010)	In	✓	✓	✗	✗	✗	✗	Decision tree
Luong et al. (2011)	Pre	✓	✗	✗	✗	✗	✗	Any
Kamishima et al. (2011)	In	✗	✓	✗	✗	✗	✗	Logistic regression
Zemel et al. (2013)	Pre/In	✓	✓	✗	✗	✗	✗	Log loss
Feldman et al. (2015)	Pre	✓	✓	✗	✗	✓	✓	Any (only numerical features)
Goh et al. (2016)	In	✓	✓	✓	✗	✓	✓	Ramp loss
Hardt et al. (2016)	Post	✗	✗	✓	✗	✓	✓	Any score-based
Corbett-Davies et al. (2017b)	Post	✗	✓	✓	✗	✓	✓	Any score-based
Woodworth et al. (2017)	In	✗	✗	✓	✗	✗	✗	Any convex linear
Quadrianto and Sharmanska (2017)	In	✓	✓	✓	✗	✗	✗	Hinge loss
Calmon et al. (2017)	Pre	✓	✓	✗	✗	✓	✓	Any
Dwork et al. (2018)	In/Post	✗	✓	✓	✗	✓	✓	Any score-based
Menon and Williamson (2018)	Post	✗	✓	✓	✗	✗	✗	Any score-based

Table 6.1: Capabilities of different methods in eliminating disparate treatment (DT), disparate impact (DI) and disparate mistreatment (DM). We also show the type of each method: pre-processing (pre), in-processing (in) and post-processing (post). None of the prior methods addresses disparate impact's business necessity (BN) clause. Many of the methods do not generalize to multiple (*e.g.*, gender and race) or polyvalent sensitive features (*e.g.*, race, that has more than two values). The strategy by (Feldman et al., 2015) is limited to only numerical non-sensitive features.

decisions. The method of Kamiran et al. (2010), which is limited to a decision tree classifier, operates by changing the splitting or the leaf node labeling criterion of the tree learning phase to remove disparate impact.

Goh et al. (2016), Woodworth et al. (2017) and Quadrianto and Sharmanska (2017) on the other hand suggest adding constraints similar to ours to the classification model. However, their works are only limited to a single specific loss function (Goh et al., 2016; Quadrianto and Sharmanska, 2017) or to a single notion of discrimination (Woodworth et al., 2017).

Finally, Zemel et al. (2013), building on Dwork et al. (2012), combined pre-processing and in-processing by jointly learning a 'fair' representation of the data and the classifier parameters. The joint representation is learnt using a multi-objective loss function that ensures that (i) the resulting representations do not lead to disparate impact, (ii) the reconstruction loss from the original data and intermediate representations is small and (iii) the class label can be predicted with high accuracy. This approach has two main limitations: i) it leads to a non-convex optimization problem and does not guarantee optimality, and ii) the accuracy of the classifier depends on the dimension of the fair representation, which needs to be chosen rather arbitrarily.

6.2.3 Post-processing

The third and final strategy consists of post-processing the classifier scores such that the new outcomes contain no disparate impact or disparate mistreatment (Corbett-Davies et al., 2017b; Dwork et al., 2018; Hardt et al., 2016; Menon and Williamson, 2018).

This approach usually involves learning different decision thresholds for a given score function to remove discrimination (specifically, disparate impact or disparate mistreatment). However, since these strategies require the sensitive feature information at the decision time, they cannot be used in cases where sensitive feature information is unavailable (*e.g.*, due to privacy reasons) or prohibited from being used due to disparate treatment laws (Barocas and Selbst, 2016). For further discussion on using the sensitive feature information at the decision time, see Sections 7.1 and 7.2.

Dwork et al. (2018) combine the in-processing and post-processing scheme by first training a number of classifiers for each group (with each classifier having different acceptance rate for the given group), and then selecting the group-conditional classifiers that minimize a certain loss function. The loss function is formulated as a combination of the loss in accuracy and a penalty term penalizing the deviation from the nondiscrimination goal. Like Hardt et al. (2016) and Corbett-Davies et al. (2017b), this method too requires access to the sensitive feature information at the decision time.

In addition to the issues discussed above, prior studies suffer from one or more of the following limitations: (i) they only accommodate a single, binary sensitive feature, (ii) they are restricted to a narrow range of classifiers, and, (iii) they cannot accommodate multiple discrimination notions *simultaneously*. Table 6.1 compares the capabilities of different methods in meeting different nondiscrimination goals.

Finally, some recent studies (Kilbertus et al., 2017; Kusner et al., 2017) focus on detecting and removing discrimination by leveraging causal inference techniques. However, these studies often require access to causal graphs specifying causal relationships between different features, which can be quite challenging to obtain in practice.

6.3 Fairness beyond discrimination

Notice that in this thesis, our focus was mostly on discrimination. As noted in Section 2.1, discrimination inherently involves imposition on wrongful relative disadvantage on "salient social groups". However, unfair treatment of persons can be carried out without regard to their salient social group membership. This kind of disadvantageous treatment is often referred to as *individual unfairness* in the machine learning literature. On the other hand, the discrimination measures discussed in this thesis (*e.g.*, disparate impact and disparate mistreatment) are often referred to as *group fairness* measures.

Dwork et al. (2012) were the first to formalize the idea of individual fairness. Their idea of individual fairness follows the insight that similar individuals must be treated similarly by the machine learning algorithm. A task specific measure is required to compute the similarity of individuals. Then, they formalize their individual fairness notion using a Lipschitz condition. They also propose mechanisms to achieve individual fairness. In a recent work, Rothblum and Yona (2018) propose mechanisms to alleviate the generalizability and computational intractability problems faced by the method of Dwork et al. (2012).

In a recent study, Speicher et al. (2018) propose another measure for individual fairness. This study argues that while two classifiers violating the Lipschitz condition of Dwork et al. would be deemed individually unfair, it is not clear *which of the two classifiers is more unfair*. They also note that while Dwork et al.'s notions of individual fairness aims at treating similar individuals similarly, it does not take into account the actual qualification (or degree of deservedness) of the individuals. Speicher et al. then propose a new measure of individual fairness that can potentially overcome these issues. Their measure uses inequality indices (specifically, generalized entropy indices) to quantify individual unfairness in the outcomes of a classifier. Using the subgroup decomposibility property of inequality indices, Speicher et al. also formalize the link between the individual unfairness, group unfairness and between-group unfairness of a classifier.

Finally, while discrimination is related to a very specific notion (discussed in detail in Section 2.1), the idea of fairness or justice in law, moral philosophy and ethics spans a much broader ground (Arneson, 2015; Gosepath, 2011; Miller, 2017; Rawls, 2009).

6.4 Connecting various notions of fairness and nondiscrimination

Since we have discussed a number of notions of fairness and nondiscrimination leading up to this point, in this section, we provide a high level view of these notions, and compare / contrast them with each other.

Table 6.2 provides an overview of these notions. As the table shows, the fairness notions can be divided into individual-level fairness notions or group-level notions. As discussed in Section 6.3 individual unfairness can be detected / measured using the methods described in Dwork et al. (2012) or Speicher et al. (2018).

The group unfairness on the other hand can be measured using parity-based notions of disparate treatment, disparate impact or disparate mistreatment. The table also shows the conditions imposed by each of these notions. For example, removing disparate treatment requires that changing the sensitive feature (with all other features being the same) should not change the decision of the classifier for an individual.[22]

The group unfairness can also be measured using preference-based notions of preferred treatment and preferred impact. Note that while the parity-based notions draw inspiration from anti-discrimination legislation in various countries (Chapters 2, 4), the preference-based notions are inspired by ideas in economics and game theory (Chapter 5). Moreover, the preference-based notions also provide better accuracy than their parity-based counterparts.

Finally, while we do not focus on that in this thesis, defining and characterizing preference-based counterparts of disparate mistreatment would be an interesting avenue for future work.

[22] For a discussion into implicit disparate treatment, *i.e.*, disparate treatment via proxy features, see Section 2.4.

Type	Notions	
Individual fairness	Similar individuals receive similar outcomes (Dwork et al., 2012), or, individuals deserving similar outcomes receive similar outcomes (Speicher et al., 2018)	
Group fairness	**No disparate treatment / Parity treatment** Changing the sensitive feature does not *change* the chance of positive class outcome.	**Preferred treatment** Changing the sensitive feature does not *improve* the chance of positive class outcome (at the level of groups). Provides better accuracy than parity treatment.
	No disparate impact / Parity impact Positive class outcome rate *similar* for all groups.	**Preferred impact** Positive class outcome rate *at least as much as parity impact* for all groups. Provides better accuracy than parity impact.
	No disparate mistreatment Prediction accuracy or its components, *i.e.*, false positive rate, false negative rate, false omission rate and false discovery rate are the same for all groups.	

Table 6.2: A broad overview of different notions of fairness / nondiscrimination in the machine learning literature.

6.5 Distributive vs. procedural fairness

Note that almost all the fairness and discrimination notions discussed until this point are concerned with the *distribution* of outcomes (among groups in the case of discrimination or group fairness, and among individuals in the case of individual fairness).

Drawing inspiration from the rich literature in organizational justice (Greenberg, 1987), some recent studies shed light on other aspects of fairness, such as *procedural fairness* (Grgic-Hlaca et al., 2018a,b). While distributive fairness refers to the fairness of *outcomes*, procedural fairness relates to the *process* that leads to these outcomes.

Grgic-Hlaca et al. (2018b) note that using certain features might be deemed procedurally unfair (*e.g.*, criminal history of a defendant's father while accessing the recidivism risk of the defendant) even when this usage leads to fair outcomes. They further note that the machine learning models aiming to achieve fairness in outcomes might overlook the other important properties of the features that might cause those features to be deemed as unfair. Some examples of these properties are: feature volitionality, *i.e.*, whether the feature value presents the volition of the person under consideration (*e.g.*, the criminal history of the father is a non-volitional feature which the defendant might not have any control over); feature privacy, *i.e.*, whether the collection of the feature violates the privacy of the person under consideration, *etc.* Grgic-Hlaca et al. (2018b) then propose methods to quantify the procedural fairness of a classification task, and also propose mechanisms to train procedurally fair classifiers.

In a follow on work, Grgic-Hlaca et al. (2018a) conduct studies to analyze why, in addition to the reasons mentioned above, people deem usage of certain features as unfair. Some of these additional reasons are: the usage of the feature perpetuating a vicious cycle of trapping people in risky behavior, the feature in fact being caused by the sensitive feature of the person itself, *etc.*

6.6 Fairness beyond binary classification

While the focus of this thesis has been on discrimination in binary classification, there have been a number of studies in the broader area of data mining and machine learning that tackle fairness and discrimination issues. We review some of this work below.

Pedreschi et al. (2008) focus on discrimination in **classification rule mining**. They first define the notions of direct and indirect discrimination in the context of rule mining, and then propose schemes to eliminate discrimination. The latter is achieved by distorting the training dataset such that the classification rules learned on this dataset would be non-discriminatory. The aim of this scheme is to generate and release a “cleaned” dataset to third parties. Hajian and Domingo-Ferrer (2013) propose new methods that overcome the limitations of Pedreschi et al. (2008). More details about this line of work can be found in Hajian et al. (2016).

Yang and Stoyanovich (2017) focus on measuring fairness in **ranking** outcomes. Their measure follows the intuitive idea that in the output of a ranking system, items ranked towards the top tend to receive greater attention. Consequently, a fairness-aware ranker might want to ensure equal representation from different socially salient groups at the top ranks. They also propose mechanisms to learn fair rankings. A number of other studies have since proposed mechanisms for learning group-fair rankings (Singh and Joachims, 2018; Zehlike et al., 2017). In a recent work, (Biega et al., 2018) move beyond the ideas of group fairness in ranking and focus on individual fairness.

Berk et al. (2017) formalize fairness in the context of **regression** tasks. Specifically, they present different measure of fairness in regression tasks that are inspired by the ideas of individual fairness, disparate impact and disparate mistreatment in a binary classification setting. They further propose convex mechanisms to include these fairness criteria in the training of regressions tasks as regularization terms. Finally, they study the tradeoffs between fairness and accuracy in the regression setting.

Furthermore, other studies have also looked at fairness issues in **voting** (Celis et al., 2017), **recommendations and personalization** (Burke et al., 2018; Celis and Vishnoi, 2017; Yao and Huang, 2017), **clustering** (Chierichetti et al., 2017), **representation learning** (Bolukbasi et al., 2016; Edwards and Storkey, 2016; Louizos et al., 2016; Madras et al., 2018), **data summarization** (Celis et al., 2018; Kazemi et al., 2018), **bandits** (Joseph et al., 2016) and **reinforcement learning** (Doroudi et al., 2017; Jabbari et al., 2016).

6.7 Fairness over time

Some recent studies have also looked at the temporal aspect of fairness and discrimination in algorithmic decision making.

Ensign et al. (2018) study the problem of feedback loops in predictive policing. Specifically, by making a distinction between the reported crime incidents and the discovered crime incidents, they show that deploying police personnel based on crime history of a specific neighborhood can potentially lead to feedback loops that might result in over-policing of certain areas. Lum and Isaac (2016) show a similar insight.

Liu et al. (2018) study the effect of applying nondiscrimination mechanisms on algorithmic decision making outcomes. They show that while the goal of these corrective measures might be to remove the effects of historical discrimination, depending on various underlying factors, in the long run, these corrective measures can have a positive, neutral or even negative impact on the benefits received by the historically discriminated groups.

These studies point to the need of performing careful domain-specific analysis before applying (1) algorithmic decision making and (2) nondiscrimination mechanisms in real-world applications.

Chapter 7

Discussion, limitations & future work

In this section, we discuss some consequential points that follow from the prior chapters, mention the limitations of our work, and explore avenues for future work.

7.1 Achieving optimal tradeoffs between nondiscrimination and accuracy

In this thesis, we proposed constraints based on distance from decision boundary for controlling various forms of discrimination (*e.g.*, disparate impact, disparate mistreatment). However, we note that these constraints are merely proxies for the positive class acceptance rate (in the case of disparate impact, preferred treatment and preferred impact) and misclassification rates (in the case of disparate mistreatment), and might not lead to optimal results in terms of tradeoffs between nondiscrimination and accuracy. In this section, we discuss some result from the machine learning literature regarding the optimality of these tradeoffs.

Corbett-Davies et al. (2017b) were the first to theoretically quantify the tradeoffs between nondiscrimination and accuracy.[23] Specifically, they show that to achieve optimal classification performance (*i.e.*, the immediate utility) under disparate impact and disparate mistreatment constraints, the classifier should apply separate thresholds for each sensitive feature group. Lipton et al. (2017) and Menon and Williamson (2018) derive the same result for classification accuracy (instead of immediate utility), and show that constraint-based mechanisms, such as ours, that do not use the sensitive feature

[23] Corbett-Davies et al. (2017b) study the problem of algorithmic nondiscrimination from the perspective of bail decisions and public safety. Public safety here is related to whether or not a defendant would go on to commit a crime, if released. Consequently, instead of using the classification accuracy, they study a slightly different objective that they refer to as 'immediate utility'. This objective is formulated as a combination of the utility of a classifier and the cost of detaining individuals. As Corbett-Davies et al. (2017b) mention, and Lipton et al. (2017) show, this objective can also be reformulated as the classification accuracy, and the takeaways would stay the same.

will have an accuracy that is lower than, or at best equal to the one achieved by setting different thresholds.

These results suggest that to get optimal tradeoffs between nondiscrimination and accuracy, one should first train an accuracy-maximizing classifier, and then set the separate decision thresholds in a post-processing step. However, we note two important points regarding these results:

First, the results assume that the decision maker has access to the Bayes optimal predictor. However, access to Bayes optimal predictors with finite datasets (as is often the case in real-world) might not be possible. In fact, Woodworth et al. (2017) show that in the absence of the Bayes optimal predictor, the post-processing scheme can lead to non-optimal results. They further argue that for achieving optimal tradeoffs, one would need to include the nondiscrimination criteria in the design of the learning algorithm itself. Woodworth et al. (2017) also present an in-processing training scheme with finite sample guarantees. However, as they discuss, the guarantees only hold under strong distributional assumptions.

Second, since the post-processing schemes achieve nondiscrimination by setting different thresholds for different sensitive feature groups, they need access to the sensitive feature value at the decision time. However, the sensitive feature value might not always be available at the decision time due to privacy reasons, or might be prohibited from being used due to disparate treatment laws. Specifically, when using the post-processing schemes, a black-box audit of the algorithm to check if it satisfies the disparate treatment criterion (Eq. 2.6) would show that the classifier gives different outcomes to persons who are the same along all features except for their sensitive features.[24]

On the other hand, the in-processing schemes proposed in this thesis can operate without using the sensitive feature at the decision time. Rather than setting different decision thresholds for different groups, these schemes *readjust the weights on the non-sensitive features* such that the final outcomes satisfy the given nondiscrimination criteria (*e.g.*, disparate impact, disparate mistreatment). Of course, an adversarial system designer with the intent to discriminate can use a shadow feature that is highly predictive of the sensitive feature (*e.g.*, using neighborhood to predict race, also known as redlining) to pass the audit for *explicit* or *formal* disparate treatment. However, these scenarios can be prevented by judging the procedural fairness (or the fairness of feature usage) as suggested by Grgic-Hlaca et al. (2018a,b). Additionally, as Siegel (2014) and Primus (2003) note, disparate impact might be used as a tool to root out such facially neutral,

[24] A similar audit is suggested by Kroll et al. (2016).

yet covert intentional discrimination. We further expand on the legal aspects of this discussion in Section 7.2.

Finally, coming back to optimal nondiscrimination accuracy tradeoffs, as the experiments in Sections 3.2.2 and 4.4.2 show, our constraint based schemes achieve similar accuracy for the same level of discrimination as the post-processing schemes when using the sensitive feature at the decision time. However, we note that for some datasets (Section 3.2.2), as opposed to the post-processing schemes, our scheme does not always completely remove discrimination. This may be due to the fact that our scheme relies on a proxy (covariance between the sensitive feature value and distance from decision boundary) to achieve nondiscrimination, while the post-processing schemes directly adjust the per-group decision thresholds to satisfy nondiscrimination criteria. Regardless, exploring further proxies, possibly with guarantees regarding nondiscrimination-accuracy tradeoffs would definitely be an important future research direction.

We next discuss some potential legal issues related to the usage of sensitive features at the decision time.

7.2 Directly using sensitive features to avoid disparate impact or disparate mistreatment

Continuing our discussion on in-processing vs. post-processing schemes, in this section we discuss the legal issues that might arise as a result of setting different decision thresholds for different sensitive feature groups in order to remove disparate impact or disparate mistreatment.

While some studies argue for explicitly using sensitive features in making decisions (Berk, 2009; Lipton et al., 2017), such schemes might face legal issues due to violation of disparate treatment and equal protection laws (Barocas and Selbst, 2016). Specifically, in the context of discrimination-aware decision making, explicitly using sensitive features would be a subject to a strict scrutiny (Wex, 2018) by courts, even if the goal of these schemes is to remove the effects of historical discrimination and "even if the consideration of race is but one factor in a holistic review" (MacCarthy, 2017).

However, as MacCarthy (2017) notes, using other (non-sensitive) features to achieve the same goals may not trigger strict scrutiny. Specifically, while reviewing various recent US Supreme Court verdicts concerning anti-discrimination practices, MacCarthy (2017) states:

> "The implication Inclusive Communities has for designing or modifying algorithms to avoid a disparate impact seems clear: if the objective is to close

> a racial gap, or by implication a gap with respect to any protected class, then modification of the algorithm with variables that do not explicitly refer to group membership would not trigger strict scrutiny. This would be true even if the variables correlated with group membership. The examples used in this case clarify the kind of variable that would not trigger strict scrutiny of a modified algorithm: low income areas, the financial feasibility of the development project, the income level of tenants, neighborhoods with good schools, high crime areas, and neighborhoods near landfills.
>
> The Court seems to be concerned about variables that explicitly refer to group membership. Case commentary also suggests incorporating variables explicitly referring to group membership into an algorithm for the purpose of making it less impactful on protected groups would trigger strict scrutiny."

Siegel (2014) expresses a similar opinion on the matter as well.

However, it is important to note that such debates on whether or not the sensitive feature should be used directly for the sake of redressing historical discrimination are still ongoing, and as noted by the authors, the current studies are by no means the final word on the matter (MacCarthy, 2017). Further progress on policy front would be required to solve this, and several other issue related to the broader topic of (discrimination in) algorithmic decision making (Barocas and Selbst, 2016; Corbett-Davies et al., 2017a; Kim, 2017; Kroll et al., 2016). Another related issue could be to clarify whether or not the explicit usage of sensitive feature at only the training stage, but not at the decision time, would trigger strict scrutiny.

7.3 Achieving nondiscrimination without sacrificing accuracy

Notice that until now, we only discussed mechanisms to remove discrimination (via pre/in/post-processing) while using the *same training dataset*.

However, various authors note that apparent disparities in beneficial outcomes may also be caused by training datasets that may be non-representative of the groups under consideration (Barocas and Selbst, 2016; Corbett-Davies et al., 2017b; Hardt, 2014; Hardt et al., 2016) and the best course of action in such cases might be to gather appropriate training datasets.

For example, one might be learning from a training dataset with very few examples from certain minority groups (these scenarios can arise easily since minorities or protected groups by definition tend to be represented less in certain domains). Having fewer examples from these groups would mean that the standard empirical risk minimization algorithms would lead to poorly fit models for these groups (Amodei et al., 2016; Bishop, 2006; Hashimoto et al., 2018). Gathering more data in such cases could potentially alleviate discrimination issues, without even needing any discrimination-aware learning scheme, and consequently, without having to sacrifice the classification accuracy.

Another issue that can potentially arise is the use of inadequate feature sets, that is, using features that are too coarse-grained or have vastly different predictive power for different groups. A classical example of the usage of coarse-grained features could be a scenario where a creditor could deny loans to whole neighborhoods based on the fact that people from that neighborhood tend not to return their loans. However, since the location in many cases can correlate with the racial makeup of a community, this practice could result in the decision making (intentionally or unintentionally) disproportionately denying loans to certain racial groups. In this case, using more fine-grained features such as individual attributes rather than neighborhood risk averages could potentially reduce such discrimination (Barocas and Selbst, 2016; MacCarthy, 2017). Similarly, discarding features with vastly different predictive power for different groups,[25] and gathering alternative feature sets can also help reduce discrimination.

Similarly, deferring decisions and gathering more information about training examples (Madras et al., 2017; Nan and Saligrama, 2017; Trapeznikov and Saligrama, 2013) based on the confidence of the algorithmic decision making system can also help reduce discrimination without having to sacrifice classification accuracy.

Further exploration of schemes related to augmenting existing training datasets with the goal of reducing nondiscrimination—perhaps along the lines of active learning (Tong and Koller, 2001), or using models with good uncertainty estimates and leveraging this uncertainty to gather more data in parts of feature space with the most variance (Rasmussen and Williams, 2005)—would be a very interesting avenue for future work.

7.4 Suitability of different measures of fairness and nondiscrimination

As noted throughout this thesis (see, *e.g.*, Sections 2.2.2, 4.1 and 5.4), discrimination is a highly domain- and context-specific notion, and consequently, *different notions* of discrimination are suitable for *different application scenarios*. As a result, a careful analysis should be carried out before a specific measure of nondiscrimination is chosen, and

[25] As shown by a 2007 FTC analysis of credit-based insurance scores, only dropping features in order to reduce outcome disparities may severely reduce the performance of the predictor (FTC, 2007).

before an algorithm is modified to satisfy that measure.[26] In this regard, some recent studies have explored potential issues that might arise when applying various notions of discrimination and unfairness without careful analysis of the underlying context.

Corbett-Davies and Goel (2018) argue that under certain situations, various discrimination measures such as well-calibration and false positive rates might not align well with the intended policy objectives. For example, they note that when the base rates for various sensitive feature groups differ, even the very high quality (accuracy maximizing) predictions could lead to differences between the false positive rates of the groups (of course, given that the predictions are not perfect).

Similarly, Speicher et al. (2018) show that removing discrimination—*e.g.*, removing disparity in false positive rates—could potentially make the outcomes more unfair at the level of individuals (for background on individual unfairness, see Section 6.3). They also show that unless a classifier separates the positive and negative class perfectly, maximizing prediction accuracy would *not* lead to the classifier maximizing individual level unfairness.

Addressing these limitations and tradeoffs would probably require efforts on both legal and technical front, and would be an interesting avenue for future work.

[26] In fact, as mentioned in Section 7.3, careful gathering of training datasets might already obviate the need for such corrective measures.

Conclusion

In this thesis, we tried to address the problem of discrimination in algorithmic decision making. First, we proposed mechanisms to limit discrimination in algorithmic decision outcomes. These mechanisms can be configured to operate with a wide range of classification models, and also provide the flexibility to accommodate other useful properties such as ensuring nondiscrimination with respect to multiple groupings of users (*e.g.*, along gender and race simultaneously), and preventing misclassifications for certain sets of users while training nondiscriminatory models. After noticing that existing measures of discrimination might not be suitable for certain application scenarios, we also propose new measures of nondiscrimination (and propose mechanisms for these new measures). One important takeaway that comes out is that there will probably not be a single universal measure of nondiscrimination in machine learning, and different measures will likely need to be applied in different situations.

We notice that several open challenges remain both on the technical as well as the policy fronts (discussed in detail in Chapter 7). Solving these challenges will require (a possibly interdisciplinary) effort on both fronts.

Finally, a code implementation of all the mechanisms proposed in this thesis is available at:

https://github.com/mbilalzafar/fair-classification

Appendices

Appendix A

Dataset statistics

—**The Adult dataset** (Adult, 1996): We consider gender and race as sensitive features.

Gender	Income $\leq 50K$	Income $> 50K$	Total
Males	20,988(69%)	9,539(31%)	30,527
Females	13,026(89%)	1,669(11%)	14, 695
Total	34,014(75%)	11,208(25%)	45,222

Table A.1: [Adult dataset] Class distribution for different genders. The classes are: whether a person earns more than $50K$ USD per year or not.

Race	Income $\leq 50K$	Income $> 50K$	Total
American-Indian/Eskimo	382	53	435
Asian/Pacific-Islander	934	369	1,303
White	28,696	10,207	38, 903
Black	3,694	534	4, 228
Other	308	45	353
Total	34,014(75%)	11,208(25%)	45,222

Table A.2: [Adult dataset] Class distribution for different races. The classes are: whether a person earns more than $50K$ USD per year or not.

—**The Bank Marketing dataset** (Bank, 2014): We consider age as the sensitive feature.

Age	Term deposit: No	Term deposit: Yes	Total
$25 \leq$ age ≤ 60	35,240(90%)	3,970(10%)	39,210
age < 25 or age > 60	1,308(66%)	670(34%)	1,978
Total	36,548(89%)	4,640(11%)	41,188

Table A.3: [Bank dataset] Class distribution for different races. The classes are: whether a person would subscribe for a term deposit or not.

—**ProPublica COMAPS dataset** (Larson et al., 2016a): We consider race as the sensitive feature.

Race	Recidivate	Did not recidivate	Total
Black	$1,661(52\%)$	$1,514(48\%)$	$3,175$
White	$8,22(39\%)$	$1,281(61\%)$	$2,103$
Total	$2,483(47\%)$	$2,795(53\%)$	$5,278$

Table A.4: [ProPublica COMPAS dataset] Class distribution for different races. The classes are: whether a defendant would receidivate within two years or not.

—**NYPD SQF dataset** (Stop, Question and Frisk Data, 2017): We consider race as the sensitive feature.

Since the NYPD SQF policy changed over time, with significantly different number of stops per year (NYCLU, 2018), we only use the data from the year 2012 for the sake of consistency. As explained in Section 4.4.2, since the original dataset (Table A.5) is highly skewed towards the positive class (person not found in posession of a weapon), we subsample the majority class (positive) to match the size of the minority (negative) class.

Race	Weapon discovered: Yes	Weapon discovered: No	Total
Black	$2,113(3\%)$	$77,337(97\%)$	$79,450$
White	$803(15\%)$	$4,616(85\%)$	$5,419$
Total	$2,916(3\%)$	$81,953(97\%)$	$84,869$

Table A.5: [NYPD SQF dataset—original] Class distribution for different races. The classes are: whether or not an illegal weapon would be recovered on a pedestrian stopped at the suspicion of carrying one.

Race	Weapon discovered: Yes	Weapon discovered: No	Total
Black	$2,113(43\%)$	$2,756(57\%)$	$4,869$
White	$803(83\%)$	$160(17\%)$	963
Total	$2,916(50\%)$	$2,916(50\%)$	$5,832$

Table A.6: [NYPD SQF dataset—with balanced classes] Class distribution for different races. The classes are: whether or not an illegal weapon would be recovered on a pedestrian stopped at the suspicion of carrying one.

Bibliography

Adult (1996). http://tinyurl.com/UCI-Adult.

Altman, A. (2016). Discrimination. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*.

Angwin, J. and Larson, J. (2016). Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say. Accessed: 2018-06-22.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arneson, R. (2015). Equality of opportunity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition.

Avraham, R., Logue, K. D., and Schwarcz, D. (2014). Understanding insurance antidiscrimination laws. *Southern California Law Review*, 87(2):195–274.

Badger, E. (2016). We're all a little biased, even if we don't know it. https://www.nytimes.com/2016/10/07/upshot/were-all-a-little-biased-even-if-we-dont-know-it.html. Accessed: 2018-06-02.

Bagenstos, S. R. (2015). Disparate impact and the role of classification and motivation in equal protection law after inclusive communities. *Cornell Law Review*, 101.

Bank (2014). `http://tinyurl.com/UCI-Bank`.

Barocas, S. and Hardt, M. (2017). NIPS 2017 Tutorial on Fairness in Machine Learning. `http://mrtz.org/nips17/`. Accessed: 2018-06-11.

Barocas, S. and Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and social problems*, 1(4):231.

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

Berliant, M. and Thomson, W. (1992). On the Fair Division of a Heterogeneous Commodity. *Journal of Mathematics Economics* .

Berman, J. B. (2000). Defining the" Essence of the Business": An Analysis of Title VII's Privacy BFOQ after Johnson Controls. *The University of Chicago Law Review*, 67(3):749–775.

Bernard, C. and Hepple, B. (1999). Indirect discrimination: Interpreting seymour-smith. *The Cambridge Law Journal*, 58(2):399–412.

Bhardwaj, R., Nambiar, A. R., and Dutta, D. (2017). A study of machine learning in healthcare. In *Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual*, volume 2, pages 236–241. IEEE.

Biddle, D. (2005). *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower.

Biega, A. J., Gummadi, K. P., and Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. *arXiv preprint arXiv:1805.01788*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS*.

Brown (1954). Brown vs. Board of Education. Supreme Court of the United States.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

Burgess, E. W. (1928). Factors Determining Success or Failure on Parole. *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois*.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*, pages 202–214.

Calders, T. and Verwer, S. (2010). Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*.

Celis, L. E., Huang, L., and Vishnoi, N. K. (2017). Group fairness in multiwinner voting. *arXiv preprint arXiv:1710.10057*.

Celis, L. E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., and Vishnoi, N. K. (2018). Fair and diverse dpp-based data summarization. *arXiv preprint arXiv:1802.04023*.

Celis, L. E. and Vishnoi, N. K. (2017). Fair personalization. *arXiv preprint arXiv:1707.02260*.

Chandler, S. (2017). The AI Chatbot will Hire You Now. `https://www.wired.com/story/the-ai-chatbot-will-hire-you-now/`. Accessed: 2018-06-02.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037.

Chouldechova, A. (2016). Fair Prediction with Disparate Impact:A Study of Bias in Recidivism Prediction Instruments. *arXiv preprint, arXiv:1610.07524*.

Civil Rights Act (1964). Civil Rights Act of 1964, Title VII, Equal Employment Opportunities.

Cook, R. (2015). Discrimination revised: reviewing the relationship between social groups, disparate treatment, and disparate impact. *Moral Philosophy and Politics*, 2(2):219–244.

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

Corbett-Davies, S., Goel, S., and Gonzólez-Bailón, S. (2017a). Even imperfect algorithms can improve the criminal justice system. `https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html`. Accessed: 2018-06-02.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017b). Algorithmic Decision Making and the Cost of Fairness. In *KDD*.

Covington, P., Adams, J., and Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *RecSys*.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and Attacking the Saddle Point Problem in High-dimensional Non-convex Optimization. In *NIPS*.

Dickson, B. (2017). How Artificial Intelligence Is Shaping the Future of Education. `https://www.pcmag.com/article/357483/how-artificial-intelligence-is-shaping-the-future-of-educati`. Accessed: 2018-06-04.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40.

Doroudi, S., Thomas, P. S., and Brunskill, E. (2017). Importance sampling for fair policy selection. In *UAI*.

Dwork, C., Hardt, M., Pitassi, T., and Reingold, O. (2012). Fairness Through Awareness. In *ITCSC*.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

Edwards, H. and Storkey, A. (2016). Censoring representations with an adversary. In *ICLR*.

Eidelman, V. (2017). Secret Algorithms Are Deciding Criminal Trials and We're Not Even Allowed to Test Their Accuracy. **https://www.aclu.org/blog/privacy-technology/surveillance-technologies/secret-algorithms-are-deciding-criminal-trials-and**. Accessed: 2018-06-04.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 160–171.

FDIC's Compliance Examination Manual (2017). FDIC's Compliance Examination Manual. **https://www.fdic.gov/regulations/compliance/manual/**. Accessed: 2018-06-12.

Federal Reserve (2016). Consumer compliance handbook. **https://www.federalreserve.gov/publications/supervision_cch.htm**. Accessed: 2018-08-31.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In *KDD*.

FICO (2018a). FICO At a Glance. **http://www.fico.com/en/about-us#at_glance**. Accessed: 2018-06-16.

FICO (2018b). How Credit History Impacts Your Credit Score. **https://www.myfico.com/credit-education/whats-in-your-credit-score/**. Accessed: 2018-06-25.

Fisher (2016). Fisher vs. University of Texas. Supreme Court of the United States.

Flores, A. W., Lowenkamp, C. T., and Bechtel, K. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.".

Fribergh, E. and Kjaerum, M. (2010). Handbook on European Non-discrimination Law. **http://fra.europa.eu/sites/default/files/fra_uploads/1510-FRA-CASE-LAW-HANDBOOK_EN.pdf**. Accessed: 2018-06-11.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

FTC (2007). Credit-based Insurance Scores: Impacts on Consumers of Automobile Insurance. A Report to Congress by the Federal Trade Commission.

FTC (2016). Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.

Fullinwider, R. (2018). Affirmative action. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition.

Furletti, M. J. (2002). An Overview and History of Credit Reporting. http://dx.doi.org/10.2139/ssrn.927487.

Fussell, S. (2017). Why can't this soap dispenser identify dark skin? https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773. Accessed: 2018-06-30.

Gal, Y. K., Mash, M., Procaccia, A. D., and Zick, Y. (2016). Which is the fairest (rent division) of them all? In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 67–84. ACM.

Gano, A. (2017). Disparate impact and mortgage lending: A beginner's guide. *U. Colo. L. Rev.*, 88:1109.

Gentle, J. E., Härdle, W. K., and Mori, Y. (2012). *Handbook of Computational Statistics: Concepts and Methods*. Springer Science & Business Media.

GINA (2008). Genetic Information Nondiscrimination Act of 2008.

Goel, S., Rao, J. M., and Shroff, R. (2016). Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. (2016). Satisfying Real-world Goals with Dataset Constraints. In *NIPS*.

Gold, M. E. (2004). Disparate impact under the Age Discrimination in Employment Act of 1967. *Berkeley J. Emp. & Lab. L.*, 25:1.

Goodman, B. and Flaxman, S. (2016). EU Regulations on Algorithmic Decision-making and a "Right to Explanation". In *ICML WHI Workshop*.

Goodman, B. W. (2016). A step towards accountable algorithms? algorithmic discrimination and the european union general data protection. NIPS Symposium of Machine Learning and the Law.

Gosepath, S. (2011). Equality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2011 edition.

Graepel, T., Candela, J. Q., Borchert, T., and Herbrich, R. (2010). Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*.

Green, R. C., Lautenbach, D., and McGuire, A. L. (2015). GINA, genetic discrimination, and genomic medicine. *New England Journal of Medicine*, 372(5):397–399.

Greenberg, J. (1987). A taxonomy of organizational justice theories. *Academy of Management review*, 12(1):9–22.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. (2018a). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *WWW*.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2018b). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*.

Griggs (1971). Griggs vs. Duke Power Co. Supreme Court of the United States.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126. ACM.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459.

Hardt, M. (2014). How Big Data is Unfair: Understanding Sources of Unfairness in Data Driven Decision Making. *Medium*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In *NIPS*.

Hart, H. (1923). Predicting parole success. *J. Am. Inst. Crim. L. & Criminology*, 14:405.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*.

Hoffman, P. B. and Beck, J. L. (1974). Parole decision-making: A salient factor score. *Journal of criminal justice*, 2(3):195–206.

Inclusive Communities (2015). Texas Department of Housing and Community Affairs vs. Inclusive Communities Project, Inc. Supreme Court of the United States.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. (2016). Fairness in reinforcement learning. *arXiv preprint arXiv:1611.03071*.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in Learning: Classic and Contextual Bandits. In *NIPS*.

Kamiran, F. and Calders, T. (2010). Classification with No Discrimination by Preferential Sampling. In *BENELEARN*.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2011). Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*.

Kazemi, E., Zadimoghaddam, M., and Karbasi, A. (2018). Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. In *ICML*.

Kehl, D. L. and Kessler, S. A. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. In *NIPS*.

Kim, P. T. (2017). Data-driven discrimination at work. *William and Mary Law Review*, 58(3):857.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

Krieger, L. H. and Fiske, S. T. (2006). Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment. *California Law Review*, 94(4):997–1062.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, 165:633.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., and Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6):1060.

Kuncel, N. R., Ones, D. S., and Klieger, D. M. (2014). In hiring, algorithms beat instinct. https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct. Accessed: 2018-06-02.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual Fairness. In *NIPS*.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016a). https://github.com/propublica/compas-analysis.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016b). How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed: 2018-06-22.

Lippert-Rasmussen, K. (2006). The Badness of Discrimination. *Ethical Theory and Moral Practice*, 9(2):167–185.

Lipton, Z. C., Chouldechova, A., and McAuley, J. (2017). Does mitigating ml's disparate impact require disparate treatment? *arXiv preprint arXiv:1711.07076*.

Liu, L., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3156–3164.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., Hipp, J. D., Peng, L., and Stumpe, M. C. (2017). Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv preprint arXiv:1703.02442*.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2016). The variational fair autoencoder. In *ICLR*.

Lukyanenko, R., Evermann, J., and Parsons, J. (2014). Instantiation Validity in IS Design Research. In *DESRIST*.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*.

MacCarthy, M. (2017). Standards of fairness for disparate impact assessment of big data algorithms. *Cumb. L. Rev.*, 48:67.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3381–3390.

Madras, D., Pitassi, T., and Zemel, R. (2017). Predict responsibly: Increasing fairness by learning to defer. *arXiv preprint arXiv:1711.06664*.

Maliszewska-Nienartowicz, J. (2014). Direct and indirect discrimination in european union law–how to draw a dividing line. *International Journal of Social Sciences*, 3(1):41–55.

Meares, T. L. (2014). The law and social science of stop and frisk. *Annual review of law and social science*, 10:335–352.

Menon, A. K. and Williamson, R. C. (2017). The Cost of Fairness in Classification. *arXiv:1705.09055*.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, D. (2017). Justice. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., and Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411.

Muñoz, C., Smith, M., and Patil, D. (2016). Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective.

Nan, F. and Saligrama, V. (2017). Adaptive classification for prediction under a budget. In *Advances in Neural Information Processing Systems*, pages 4727–4737.

Nash Jr, J. F. (1950). The Bargaining Problem. *Econometrica: Journal of the Econometric Society*.

Niklas, J., Sztandar-Sztanderska, K., and Szymielewicz, K. (2015). Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making. ***https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf***. Accessed: 2018-06-02.

NYCLU (2018). Stop-and-Frisk Data. ***https://www.nyclu.org/en/stop-and-frisk-data***. Accessed: 2018-06-24.

OŃeil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. The Crown Publishing Group.

Pachal, P. (2015). Google Photos Identified Two Black People as 'Gorillas'. ***http://mashable.com/2015/07/01/google-photos-black-people-gorillas/***.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware Data Mining. In *KDD*.

Perry, W. L. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation.

Podesta, J., Pritzker, P., Moniz, E., Holdren, J., and Zients, J. (2014). Big data: Seizing opportunities, preserving values. *Executive Office of the President. The White House.*

Posse, C. (2016). Cloud Jobs API: Machine Learning Goes to Work on Job Search and Discovery. ***https://cloud.google.com/blog/big-data/2016/11/cloud-jobs-api-machine-learning-goes-to-work-on-job-search-and-discovery***. Accessed: 2018-05-23.

Primus, R. A. (2003). Equal protection and disparate impact: Round three. *Harv. L. Rev.*, 117:494.

Quadrianto, N. and Sharmanska, V. (2017). Recycling Privileged Learning and Distribution Matching for Fairness. In *NIPS*.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rawls, J. (2009). *A theory of justice: Revised edition*. Harvard university press.

Ricci (2009). Ricci vs. DeStefano. Supreme Court of the United States.

Romei, A. and Ruggieri, S. (2014). A Multidisciplinary Survey on Discrimination Analysis. *KER*.

Rothblum, G. N. and Yona, G. (2018). Probably approximately metric-fair learning. *arXiv preprint arXiv:1803.03242*.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Shen, X., Diamond, S., Gu, Y., and Boyd, S. (2016a). DCCP. **https://github.com/cvxgrp/dccp**. Accessed: 2018-07-07.

Shen, X., Diamond, S., Gu, Y., and Boyd, S. (2016b). Disciplined Convex-Concave Programming. *arXiv:1604.02639*.

Siegel, R. B. (2014). Race-conscious but race-neutral: The constitutionality of disparate impact in the roberts court. *Ala. L. Rev.*, 66:653.

Singh, A. and Joachims, T. (2018). Fairness of exposure in rankings. *arXiv preprint arXiv:1802.07281*.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *KDD*.

Stop, Question and Frisk Data (2017). **http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page**. Accessed: 2018-06-22.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *ACM Queue*.

Tatum, B. D. (2017). *Why are all the Black kids sitting together in the cafeteria?: And other conversations about race*. Basic Books.

Tibbitts, C. (1931). Success or failure on parole can be predicted: A study of the records of 3,000 youths paroled from the illinois state reformatory. *Am. Inst. Crim. L. & Criminology*, 22:11.

Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Trapeznikov, K. and Saligrama, V. (2013). Supervised sequential classification under budget constraints. In *Artificial Intelligence and Statistics*, pages 581–589.

Varian, H. R. (1974). Equity, Envy, and Efficiency. *Journal of Economic Theory*.

Wex (2018). Strict Scrutiny. https://www.law.cornell.edu/wex/strict_scrutiny. Accessed: 2018-06-22.

Woods, D. (2011). LinkedIn's Monica Rogati On "What Is A Data Scientist?". https://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist/. Accessed: 2018-06-25.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning Non-Discriminatory Predictors. In *COLT*.

Yang, K. and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, page 22. ACM.

Yao, S. and Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2925–2934.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017a). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017b). Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., Gummadi, K. P., and Weller, A. (2017c). From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*.

Zarsky, T. Z. (2014). Understanding Discrimination in the Scored Society. *Washington Law Review*, 89(4):1375.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578. ACM.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning Fair Representations. In *ICML*.