# Limitations of Fairness in Machine Learning

**Dissertation**
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Michael Lohaus
aus Ochtrup

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

| | |
|---|---|
| Tag der mündlichen Qualifikation | 30.05.2022 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatterin: | Prof. Dr. Ulrike von Luxburg |
| 2. Berichterstatter: | Prof. Dr. hon. Gjergji Kasneci |

# Abstract

The issue of socially responsible machine learning has never been more pressing. An entire field of machine learning is dedicated to investigating the societal aspects of automated decision–making systems and providing technical solutions for algorithmic fairness. However, any attempt to improve the fairness of algorithms must be examined under the lens of potential societal harm. In this thesis, we study existing approaches to fair classification and shed light on their various limitations.

First, we show that relaxations of fairness constraints used to simplify the learning process of fair models are too coarse, since the final classifier may be distinctly unfair even though the relaxed constraint is satisfied. In response, we propose a new and provably fair method that incorporates the fairness relaxations in a strongly convex formulation.

Second, we observe an increased awareness of protected attributes such as race or gender in the last layer of deep neural networks when we regularize them for fair outcomes. Based on this observation, we construct a neural network that explicitly treats input points differently because of protected personal characteristics. With this explicit formulation, we can replicate the predictions of a fair neural network. We argue that both the fair neural network and the explicit formulation demonstrate *disparate treatment*—a form of discrimination in anti-discrimination laws.

Third, we consider fairness properties of the majority vote—a popular ensemble method for aggregating multiple machine learning models to obtain more accurate and robust decisions. We algorithmically investigate worst-case fairness guarantees of the majority vote when it consists of multiple classifiers that are themselves already fair. Under strong independence assumptions on the classifiers, we can guarantee a fair majority vote. Without any assumptions on the classifiers, a fair majority vote cannot be guaranteed in general, but different fairness regimes are possible: on the one hand, using fair classifiers may improve the worst-case fairness guarantees. On the other hand, the majority vote may not be fair at all.

# Zusammenfassung

Die Frage des sozial verantwortlichen maschinellen Lernens ist so dringlich wie nie zuvor. Ein ganzer Bereich des maschinellen Lernens hat es sich zur Aufgabe gemacht, die gesellschaftlichen Aspekte automatisierter Entscheidungssysteme zu untersuchen und technische Lösungen für algorithmische Fairness bereitzustellen. Jeder Versuch, die Fairness von Algorithmen zu verbessern, muss jedoch unter dem Blickwinkel eines möglichen gesellschaftlichen Schadens untersucht werden. In dieser Arbeit untersuchen wir bestehende Ansätze für faire Klassifikationsverfahren und beleuchten deren unterschiedliche Einschränkungen.

Als Erstes zeigen wir, dass Relaxierungen von Fairness, die verwendet werden, um den Lernprozess von fairen Modellen zu vereinfachen, zu grob sind, da der endgültige Klassifikator unfair sein kann, obwohl die relaxierte Bedingung erfüllt ist. Als Antwort darauf schlagen wir eine neue und beweisbar faire Methode vor, die die Fairness-Relaxierungen in einer stark konvexen Formulierung wiederverwendet.

Zweitens beobachten wir ein erhöhtes Bewusstsein für geschützte Merkmale wie Rasse oder Geschlecht in der letzten Schicht tiefer neuronaler Netze, wenn wir sie für faire Ergebnisse regularisieren. Auf Basis dieser Beobachtung konstruieren wir ein neuronales Netz, das die Eingabepunkte wegen geschützter persönlicher Merkmale explizit unterschiedlich behandelt. Mit dieser expliziten Formulierung können wir die Vorhersagen eines fairen neuronalen Netzwerks replizieren. Wir behaupten, dass sowohl das faire neuronale Netzwerk als auch die explizite Formulierung *Disparate Treatment* aufzeigen—eine Form der Diskriminierung in vielen Antidiskriminierungsgesetzen.

Drittens betrachten wir die Fairness-Eigenschaften des Mehrheitsvotums - einer beliebten Ensemble-Methode zur Aggregation mehrerer Modelle maschinellen Lernens. Wir untersuchen algorithmisch Worst-Case-Garantien für die Fairness des Mehrheitsvotums, wenn es aus mehreren Klassifikatoren besteht, die selbst schon fair sind. Unter starken Unabhängigkeitsannahmen an die Klassifikatoren können wir ein faires Mehrheitsvotum garantieren. Ohne irgendwelche Annahmen an die Klassifikatoren kann ein faires Mehrheitsvotum im Allgemeinen nicht garantiert werden, aber es sind verschiedene Fairness-Regime möglich: Einerseits kann die Verwendung fairer Klassifikatoren die Fairness-Garantien für den Worst-Case verbessern. Andererseits kann es sein, dass das Mehrheitsvotum überhaupt nicht fair ist.

# Acknowledgements

As I am writing this thesis, my PhD years are almost over. This is a good moment to reflect on the four-and-a-half (not five) years in Tübingen, two of which have been quite unusual. Ironically, I am writing this piece of text at home–a couple of days before my contract ends–in a ten-day quarantine. Ideal circumstances to think about the people that have helped me to get this far.

In 2017 I met Ulrike von Luxburg during interviews in Tübingen by chance. Luckily, we talked about machine learning, after which I did not hesitate to apply for a PhD position. I can count myself very lucky that I met Ulrike and I am very grateful to her for the opportunity she provided me and for the trust she put in me from the beginning. When close to giving up Ulrike has supported me with my interests at heart and a hopeful outlook for the road ahead. Thank you, Ulrike.

The best and most intense work that I have done was in collaboration with others. In particular, I thank Michaël Perrot without whom the dam would not have broken. Everthing that I have learnt during this collaboration has helped me for all other projects afterwards. I am also thankful to my other brilliant coauthors Matthäus Kleindessner, Chris Russell, Leena C Vankadara, Faiz Ul Wahab, and Siavash Haghiri. Thank you for sticking with me and accompanying me on my journey.

There was another small group of people who I was lucky to meet on a daily basis and who have become the backbone of my PhD. Thanks for the hundreds of coffee breaks, the hikes, and the natural willingness to help wherever possible: my office mates Luca Rendsburg and Moritz Haas, Solveig Klepper, Sascha Meyen, David Künstle, and Debarghya Ghoshdastidar. And thank you Patrizia Balloch for your unconditional help.

I want to thank IMPRS-IS–my TAC, Leila, and everyone who made arriving in Tübingen easy and a lot of fun. Thank you Jonas Peters for showing me this path.

The quite unusual time of social isolation made the PhD even harder in certain regards than it usually is. But special people have always kept my spirits up: Sebastian, Martin and Katie, Dominik, Mara, Caro, Damien, and the choir. Thank you for your friendship and support. The same goes out to my Cologne Crew who I can always count on, even digitally. I especially want to thank my family for being the unshakable foundation in my life. Finally, I am deeply indebted to Lamy for giving me strength every single day.

# Contents

# Chapter 1

# Introduction

In 2018 the German Parliament initiated an Enquete Commission to identify the challenges and benefits of artificial intelligence on the well-being of society and each individual (BT Drucksache 19/2978, 2018). The commission's goal was to develop a national strategy by formulating specific recommendations for future action. To that end, the commission consulted experts in the fields of economics, health, law, ethics, and computer science to lay the foundations for an informed political debate. The guiding principle of the commission was a human-centered AI that should respect human dignity and benefit society as a whole.

The commission was founded due to a growing concern about the impact of AI systems. Automated decision-making systems (ADMs) are increasingly assisting or replacing human decision-makers in deciding who gets a loan, who is pulled over by the police, or who is investigated for health insurance fraud. Although ADMs often outperform humans in evaluating large amounts of data, they do not fulfill the hope of providing decisions that are free of human biases. On the contrary, they have been shown to behave unfairly or discriminatory toward certain groups of people, often reinforcing historical patterns of discrimination or adopting implicit biases of past human decision-makers. If used carelessly, ADMs can deprive certain groups of important goods and opportunities.

Two years after its founding, the commission published its findings and policy recommendations in a final report (BT Drucksache 19/23700, 2020). The report specifically addresses discrimination in AI: "In recent years, much research has been done on discrimination detection and prevention in AI systems. The next step, the transfer of these findings into everyday software development, should be promoted so that the findings can be implemented as quickly and as widely as possible [. . .]." (BT Drucksache 19/23700, 2020, p. 63). As an example for policy makers worldwide, this report expresses the urgent need to regulate artificial intelligence in a way that allows society to leverage the benefits and avoid potential harm. With a call for explainable, transparent, and privacy–preserving AI, the report references recent advances in machine learning research[1]. In particular, it points out the challenges of discrimination and unfair treatment posed by automated decision-making systems.

Fundamentally, fairness–aware machine learning is concerned with the question about how we want to make consequential decisions about humans and what we consider to be "good" decisions (Barocas et al., 2019). Typically, fairness–aware machine learning evolves around technical solutions that aim to satisfy some notion of justice or fairness, but it also opens up new ways of reasoning about what just and fair decisions are. As the report above shows, policy makers as well as philosophers and lawyers rely heavily on future research about fair AI systems from computer scientists. In the end, society and policy makers need to decide how and if we deploy automated systems and which notions of justice and fairness are to be respected. From a machine learning perspective, it is essential to understand suggested fairness–aware methods as solutions to unfair decision–making and test them under the same standards that they seek to meet.

In this thesis, we investigate existing fairness methods in automated decision-making. We identified three limitations of fairness, which we present in three chapters. Our focus is on fairness notions that are derived from legal anti-discrimination theories that can be found for example in European and U.S. American law (Altman, 2020; Barocas and Selbst, 2016).

1. In Chapter 2, we consider fairness approaches that use relaxed formulations of a given fairness notion to reduce computational costs. We find that these relaxations are too relaxed to guarantee a fair decision-making processes.

2. We show in Chapter 3 that neural networks, which were trained to provide fair outcomes, treat people differently based on implicitly learned protected personal traits.

3. When several decision-making processes are available, we can make more accurate decisions by aggregating the outcomes in a majority vote. In Chapter 4 we provide worst-case fairness guarantees and show that the decision of the majority vote is not guaranteed to be fair.

We begin this introductory chapter by discussing a few examples from recent years of high-stakes decisions which sparked discussions about their fairness and societal impact (Section 1.1). As a starting point for fairness–aware machine learning, we use existing anti-discrimination legislation. In Section 1.2 we shortly present the legal frameworks of disparate treatment and disparate impact. Motivated by the legal domain, we formulate statistical fairness notions and consider the question how to do machine learning such that we satisfy a given fairness notion (Section 1.3). We close with the contributions of this thesis in Section 1.5.

[1] For example the ACM Conference on Fairness, Accountability, and Transparency.

## 1.1 Discriminatory Automated Decision-making

Automated decision-making has become a central tool in several human-centered applications such as financial lending, employment, public services, health insurance, or education. In the following, we present a few examples of discriminatory or harmful decision-making starting with the well-known example of recidivism risk scores.

**Pretrial recidivism scores.** In 2016 ProPublica, a non-profit newspaper, published an article about a system that estimates the risk of a defendant to commit new crimes in the future (Angwin et al., 2016). The assessments about the risk of *recidivism* can be used to decide if a defendant is set free or detained until the final trial. Angwin et al. (2016) found that these risk scores falsely categorize black defendants as high risk more often than white defendants. On the other hand, white defendants are falsely labeled low risk more often than black defendants. Northpointe (now Equivant, the owner of the system called COMPAS) answered by pointing out that the *interpretation* of the risk scores is the same for black and white defendants: Given a risk score the probability to re-offend is the same for both groups (Dieterich et al., 2016). Interestingly, both sides of the argument are true since they are based on two different statistical fairness notions, which have since been proven to be mutually exclusive (Chouldechova, 2017; Kleinberg et al., 2017).

**University admission.** In 2014 the Students for Fair Admissions (SFFA) representing a group of anonymous Asian-Americans filed a lawsuit against Harvard College. The SFFA claims that Harvard's admission process violates Title VI of the Civil Rights Act by using racial quotas to the disadvantage of Asian-Americans. The key argument of the expert report by the SFFA (Arcidiacono, 2019) is based on counterfactual reasoning. If we consider an Asian-American student, such that the chance of admission is 25%, the same student would have a chance of 36%, if we would only change the race attribute to 'white', but fix all other features of the student. Harvard, on the other hand, argues that due to the large pool of talented students, it is important to consider a range of valuable information other than test scores, which the statistical model of the SFFA does not deem relevant for admission. In addition, Harvard argues with the notion of demographic parity (see Section 1.3.1) by pointing out that 23% of the student body is Asian-American while Asian-Americans represent 6% of the U.S. population.

**Computer Vision.** Buolamwini and Gebru (2018) have shown that facial analysis algorithms are performing worse for darker skin tones than for brighter skin tones. Buolamwini and Gebru (2018) evaluate commercially available facial analysis algorithms in gender classification tasks and observe large disparities in classification accuracy. In particular, darker-skinned females are misclassified much more often than lighter males. The study notes that the datasets are mostly composed of light-skinned subjects, but even after balancing accuracy disparities can be observed.

A recent example of the disparate impact of facial detection algorithms was discovered in the wake of the Covid pandemic. Feathers (2021) reports that facial detection algorithms used for online exam surveillance are much less likely to detect the faces of darker-skinned students than of light-skinned students. As a result, it becomes a challenge for them to take high-stakes tests without triggering the surveillance software.

A multitude of high-stakes decisions and applications of automated systems have been discussed in recent years. For a comprehensive collection of harmful and discriminatory decision-making, we refer the reader to various accessible books such as Eubanks (2018); Kearns and Roth (2019); Noble (2018); O'Neil (2016). A detailed and technical overview of fairness in machine learning is provided in Barocas et al. (2019).

## 1.2 Fairness Notions from Anti-discrimination Laws

To determine if an automated system is discriminatory, we need to define what constitutes discrimination and how to measure it. An obvious, but not extensive source for notions of fairness are existing anti-discrimination laws. In the following section, we familiarize the reader with common terminology from the legal domain that has inspired many notions of fairness in machine learning. Typically, these laws address high-stakes decisions about humans such as employment, college admission, lending, criminal justice, and access to welfare and social security.

Anti-discrimination is often tested along two popular discrimination theories: (1) *disparate impact* and (2) *disparate treatment* (Barocas and Selbst, 2016). The European counterparts are called *direct* and *indirect* discrimination (Altman, 2020; Lidell and O'Flaherty, 2018). These concepts can be found in Title VII of the Civil Rights Act in the US (Statute, 1991), which protects from employment discrimination by private parties (Harned and Wallach, 2019), but also in other anti-discrimination laws such as the European Racial Equality Directive (2000/43/EC), the British Race Relations Act (1965), or the German Allgemeines Gleichbehandlungsgesetz (2006).

A popular article by Barocas and Selbst (2016), cited in both law and machine learning, lays out the possible liability of automated decision-processes with respect to Title VII. For more details, we refer the reader to Barocas and Selbst (2016) and for an overview of European anti-discrimination theories we refer to the handbook on European non-discrimination law (Lidell and O'Flaherty, 2018). For a philosophical treatise on the topic, we refer to Altman (2020). In the following, we stick to the U.S. American terminology and provide a short summary of the essential terms. First, we discuss who these laws are meant to protect.

**Protected Groups.** According to anti-discrimination laws a treatment is discriminatory when it is based on certain protected attributes (AGG, 2006; Lidell and O'Flaherty, 2018; Statute, 1991). The handbook on European non-discrimination law (Lidell and O'Flaherty, 2018) defines a *protected ground* as an objective and identifiable trait, or a status that distinguishes one person from another, and it may not be the basis for differences in treatment. As protected grounds the handbook lists sex, racial or ethnic origin, age, disability, religion or belief and sexual orientation. In this thesis, we will often refer to the protected attribute without further specifiying it. It is important to note, however, that it is common to assume that every individual can be uniquely sorted into these traits based on some underlying truth. Conceptually, this might be desirable, but this overlooks the fact that these traits are not binary, or categorical, or if there even exists such an objective property. Race, for example, was historically used to define social groups as a basis for social injustices and oppression. The definition of race as a protected attribute reinforces this social construct, even though there is no biological footing (see Barocas and Selbst (2016, Chapter 5) and references therein).

### 1.2.1 Disparate Treatment

A decision-making process exhibits disparate treatment if the decisions are based on a protected attribute. Either (a) similarly situated people are *formally or explicitly* treated differently, or (b) there is an *intent to discriminate* (Barocas and Selbst, 2016) with the chosen decision-making process.

The first case is straight forward. When an employer differentiates applicants by explicitly using the protected attribute, the policy itself proves the differential treatment (Harned and Wallach, 2019). The same argument can be applied for machine learning methods if they formally require the protected attribute as input. Even if harmful discrimination was not intended, for example in affirmative action measures, it can constitute a disparate treatment violation[2].

In practice, however, it is more relevant to prove the intent to discriminate rather than showing formal discrimination, since it can easily be hidden. It is hard to prove that an employer rejected a person because of race if the decision process is not written down or the employer simply denies using race as a selection criterion. For machine learning algorithms, it is widely known that the explicit use of the protected attribute can be masked by inferring the protected attribute from proxy variables, such as a photograph in a CV, or the postal code (Supreme Court, 1999). It is easy to construct a system that does not formally require the protected attribute, but internally uses the inferred protected attribute (Lipton et al., 2018). In Chapter 3, we will show that in fact neural networks implicitly learn to predict the protected attribute in order to obtain fair outcomes.

In addition to masking the intent do discriminate, there might not even be a conscious plan to discriminate. People can also be treated differently due to unconscious human biases. This can happen, for example, in a working environment, where an employer requires a higher performance from one gender for the same salary, or simply the same kind of praise. The employer itself might not be aware of the implicit

[2] Title VII states that explicitly considering race can be sometimes allowed for affirmative action. The EEOC states that employers may take affirmative action if an adverse impact, that does not constitute a business necessity, is proved. However, Barocas and Selbst (2016) note that under current political circumstances in the US, any affirmative action measure is under close scrutiny and likely to be decided unconstitutional.

bias, but the employees are treated differently. Barocas and Selbst (2016) note that unconscious disparate treatment is not appropriately addressed in law and therefore, proving the intent to discriminate is a hard legal exercise.

### 1.2.2 Disparate Impact

If there is no intent to discriminate, or it is hard to prove, we can instead focus on the outcome of the decision process with disparate impact concept. A (facially neutral) decision-making process suffers from disparate impact if it disproportionately harms a protected group, for example when the acceptance rate for a job is lower in one protected group than the other. If an applicant for a job can show that the employer could have used an alternative employment procedure, which meets the requirements of a successful business, but has a less discriminatory adverse impact, the original employment procedure is unlawful. However, this also means that a certain degree of disparate impact is allowed if the employment practice is related to job performance and thus a *business necessity* (Barocas and Selbst, 2016).

As a guideline, the U.S. Equal Employment Opportunity Commission (EEOC), which is responsible for enforcing Title VIIs mandate, suggests the *four-fifth rule*, which would require an employer to make sure that the acceptance rate of a protected group is not lower than 80% of the acceptance rate of another group (Equal Employment Opportunity Commission (EEO), 1978). In machine learning, a large focus has been on mitigating disparate impact since it is easier to observe the outcomes instead of understanding the process itself. In Section 1.3.1, we will look at several fairness notions related to disparate impact and machine learning approaches that aim to fulfill them.

**General Note on Fairness.** The term ’fairness’ which has come to describe this field, is a placeholder term for different concepts around fairness, justice, and discrimination. Above we described legal concepts that can give rise to notions of algorithmic fairness. However, legislation is not a framework for morality and laws alone cannot offer the full picture of fairness. That is why philosophical theories about justice and fairness have influenced the literature on fair decision-making as well. Popular theories on distributive justice are John Roemer’s Equality of Opportunity (Roemer, 1998) and Theories of Distributive Justice (Roemer, 1996), as well as John Rawls’ A Theory of Justice (Rawls, 1996, 2001), or H. Peyton Young’s Equity (Young, 1995). For an overview on justice theories, we refer the reader to Sandel (2009) or the corresponding lectures. A connection to computer science and a roadmap of fairness notions and philosophical theories is laid out in Binns (2018).

## 1.3 Binary Classification and Observational Fairness

Impactful decision-making can often be formulated in terms of a binary classification task like granting a loan, admitting someone for college, or detaining someone in prison. Typically, one of the two outcomes is more beneficial than the other and disparate impact can occur when the beneficial outcome is disproportionately granted to one group. In this thesis we consider the supervised learning task of classification and focus on observational fairness notions that aim to mitigate disparate impact. We define four observable random variables that represent the input features, the protected attribute, the target label, and the prediction.

$X \in \mathcal{X}$ are the input features to the model. We assume that $\mathcal{X} \subset \mathbb{R}^d$ only includes non-protected information, which can be used by the machine learning model without restrictions. In case of granting a loan, this information can include the income, education, or previous payment history.

$Y \in \mathcal{Y}$ is the target label. The goal is to predict $Y$ from given input features $X$ as well as possible. Since we consider binary classification, we have $\mathcal{Y} = \{-1, 1\}$. Typically, $Y = 1$ denotes the more desirable outcome like being granted the loan.

$S \in \mathcal{S}$ is the protected attribute, such as race or gender, which could be determined by the discussed anti-discrimination laws. We assume $\mathcal{S} = \{-1, 1\}$ throughout this thesis, but typically, an extension to multiple labels is straight-forward. We evaluate the fairness of a given classifier with respect to the protected attribute $S$. If not mentioned specifically, the labels of the protected attribute do not codify an advantage or disadvantage or a 'good' or 'bad' group membership.

$\hat{Y} \in \mathcal{Y}$ is the predicted label of a binary classification model $h : \mathcal{X} \to \mathcal{Y}$ with $\hat{Y} = h(X)$.

For this section we assume that a binary classifier $h : \mathcal{X} \to \mathcal{Y}$ and thus its predictions are already given. In order to determine the fairness of $h$ we formulate observational fairness notions using only the joint distribution of the random variables above. In the next section, we shortly address how to learn a fair classifier.

### 1.3.1 Mitigating Disparate Impact through Observational Fairness

The notion of disparate impact refers to the idea that facially neutral decision processes should not disproportionately affect one of the protected groups. It is up to interpretation how to actually measure the disparate impact of a machine learning model. The popular class of *observational fairness criteria* considers (conditional) independence statements of the three random variables $Y, S$, and $\hat{Y}$ (Barocas et al., 2019). This class summarizes a variety of fairness notions that have been suggested in recent years. By neglecting the classifier itself and focusing only on the observable predictions $\hat{Y}$, we can formulate criteria about how the output should be related to the protected attribute and the target label.

- **Independence:** The variables $Y$ and $S$ satisfy *independence* if $\hat{Y} \perp\!\!\!\perp S$.
- **Separation:** The variables $\hat{Y}, Y$ and $S$ satisfy *separation* if $\hat{Y} \perp\!\!\!\perp S \mid Y$.

- **Sufficiency:** The variables $\hat{Y}, Y$ and $S$ satisfy *sufficiency* if $Y \perp\!\!\!\perp S \mid \hat{Y}$.

In the following, we discuss popular fairness notions as a derivation from the independence statements. We formulate them over the joint distribution of $X, S$ and $Y$, hence, we assume that there exists a joint distribution $\mathcal{D}_{\mathcal{Z}}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, from which we can draw examples $(X, S, Y) \sim \mathcal{D}_{\mathcal{Z}}$. Often, we use the shorthand $\mathbb{P}$ for $\mathbb{P}_{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}$.

**Demographic Parity.** A binary classifier $h : \mathcal{X} \to \mathcal{Y}$ fulfills **demographic parity** (or statistical parity) (Calders and Verwer, 2010; Feldman et al., 2015; Kamishima et al., 2012; Zafar et al., 2017a) if for all $s \in \mathcal{S}$

$$\underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=s] = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1]\,. \tag{1.1}$$

Demographic parity requires that the probability of obtaining the more beneficial outcome be equal for every protected group. In our setting, where the label is binary, demographic parity implies that the probability of a negative outcome is also equal for every group. Since we assume that the protected attribute is binary, we can say that a classifier $h : \mathcal{X} \to \mathcal{Y}$ is demographic–parity–fair when

$$\underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=1] = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=-1]\,. \tag{1.2}$$

In practice, perfect fairness will hardly occur and several classifiers will vary in their degree of fairness. We measure the violation of demographic parity with the *difference of demographic parity* (DDP) (Calders and Verwer, 2010; Wu et al., 2019):

$$\mathrm{DDP}(h) = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=1] - \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=-1]\,. \tag{1.3}$$

The DDP is positive when the favored group is $S=1$ and negative when it is $S=-1$. When we have $\mathrm{DDP}(h) = 0$, the classifier $h$ is demographic–parity–fair.

It is also common to consider the ratio of the positive rates, also known as the p%-rule (Biddle, 2005; Zafar et al., 2017a), which has its root in the four-fifth rule mentioned above (Equal Employment Opportunity Commission (EEO), 1978):

$$\frac{p}{100} = \frac{\mathbb{P}\left[h(X) = 1|S = -1\right]}{\mathbb{P}\left[h(X) = 1|S = 1\right]}.$$

In this case, $p = 100$ corresponds to a demographic–parity–fair classifier. If we have $p \geq 80$ corresponding to the 80% rule (Equal Employment Opportunity Commission (EEO), 1978), the positive rate of group $S = -1$ is at least as high as 80% of the other group's positive rate.

**Disadvantages.** The notion of demographic parity is often aspired to because of its technical, but also intuitive simplicity. We might believe in a state of the world where the ability to pay back a loan or perform well in a job does not depend on certain personal traits. If, however, these personal traits actually matter in the current state of the world due to historical discrimination, demographic parity might express our belief about how the state of world *should* be.

Nevertheless, several disadvantages of demographic parity have been discussed (Hardt et al., 2016). The goal of demographic parity can be undermined by applying two differently accurate models on the protected groups. In hiring, for example, the acceptance rate could be equal for two protected groups, but the company hires qualified candidates from group A and only random candidates from group B. By accepting unqualified candidates, the drop out rate will be higher for group B than for group A due to unsatisfactory performance. Additionally, the data that is created will feed back into the bias of the data.

Even if the company does not act with ill intent and wants to correct historical disadvantages with demographic parity, the final job performance of group B might be worse than group A because the lack of opportunities has led to less qualified candidates. Again, the drop out rate for group B will be higher. It is unclear in these scenarios what the long–term impact of demographic parity will be. Liu et al. (2018) showed that blindly enforcing demographic parity can harm the group that was meant to be protected.

Finally, note that demographic parity does not allow the perfect classifier $\hat{Y} = Y$, if the rate of qualification is different in each group: $\mathbb{P}[Y{=}1|S{=}1] \neq \mathbb{P}[Y{=}1|S{=}-1]$. Again, the long–term impact is unclear if demographic parity were enforced in such a scenario.

**Equal Odds and Equal Opportunity.** In the example of granting a loan, the company is generally interested in granting loans to people who can pay them back. If the bank acknowledges that the probability of paying back the loan is higher for one of the protected groups, the bank can argue that for profitability the ability to pay back the loan has to be taken into account. However, apart from differences in the outcome caused by the true target label, the decision should not depend on the protected group.

In the case of binary classification, the criterion of *separation* is equivalent to **equal odds** (Hardt et al., 2016). A classifier $h$ satisfies equal odds if

$$\mathbb{P}[h(X) = 1|Y = y, S = 1] = \mathbb{P}[h(X) = 1|Y = y, S = -1] \text{ for all } y \in \mathcal{Y}. \tag{1.4}$$

In other words, equal odds is fulfilled if both the true positive and the false positive rates are equal across protected groups. This directly implies equal true negative and false negative rates.

We can relax equal odds if the bank determines it more important to obtain a loan when it is deserved than to obtain a loan by mistake. Then, we can require that the chance to obtain a loan deservedly should be equal for all protected groups. A classifier $h$ is fair with respect to **equality of opportunity** (Hardt et al., 2016) if

$$\mathbb{P}[h(X) = 1|Y = 1, S = 1] = \mathbb{P}[h(X) = 1|Y = 1, S = -1]\,. \tag{1.5}$$

Again, instead of only considering exact equality of opportunity, we use a fairness score called *difference of equality of opportunity* (DEO) (Donini et al., 2018):

$$\mathrm{DEO}(h) = \mathbb{P}[h(X) = 1|Y = 1, S = 1] - \mathbb{P}[h(X) = 1|Y = 1, S = -1] \,. \tag{1.6}$$

This quantity is positive when the favoured group is $s = 1$ and negative when it is $s = -1$. When $\mathrm{DEO}(h) = 0$, the classifier $h$ is fair with respect to equality of opportunity.

**Predictive Parity.** The last observational fairness criteria *sufficiency* in binary classification is referred to as **predictive parity** (Zafar et al., 2017b). A classifier $h$ satisfies predictive parity if

$$\mathbb{P}[Y = 1|h(X) = y, S = 1] = \mathbb{P}[Y = 1|h(X) = y, S = -1] \text{ for all } y \in \mathcal{Y}. \tag{1.7}$$

For the other fairness notions above, we have seen that they can be defined in terms of the confusion table between target labels $Y$ and predictions $\hat{Y}$. Similarly, predictive parity requires equal positive predictive values and equal negative predictive values. Predictive Parity can be relaxed into *positive predictive parity* (equal positive predictive values) or *negative predictive parity* (equal negative predictive values).

In this spirit, we can formulate more observational fairness criteria by matching other rates that can be formed from the confusion table (Barocas et al., 2019), for example equal (balanced) classification rates (Chouldechova, 2017; Friedler et al., 2019) or equal false discovery rate (Zafar et al., 2017b).

### 1.3.2 Discussion

**Fairness through Unawareness.** It is well known that the naive idea of removing the protected attribute from the input features is not helpful to construct classifiers that are in some sense fair or objective. The main reasons are *proxies* or *redundant encodings* that can be used to predict the protected attribute (Dwork et al., 2012). Avoiding redundant encodings in the features can be desirable since it would be impossible to determine the protected attribute and hence, construct a biased classifier $f$. In practice, however, this does not happen. A few proxy variables that are only slightly correlated with the protected attribute are enough to predict it with high accuracy (Barocas et al., 2019). For that reason new methods have been suggested to artificially construct a representation of the features that does not contain redundant encodings (Alvi et al., 2019; Madras et al., 2018; Zemel et al., 2013). Nevertheless, Grgić-Hlača et al. (2018) argue for fairness through unawareness in certain scenarios. They propose to evaluate human judgments about the use of individual input features in order to measure the procedural unfairness (a form of avoiding disparate treatment) and remove undesirable features.

Often, fairness through unawareness is interpreted to avoid disparate treatment since the classifier's decisions are not based on the protected attribute. Harned and Wallach (2019) extend this idea with the distinction between training a model and its deployment. If the decision-making process is "unaware" of the protected attribute during deployment, that is when the decisions are made, the decision-making process is not based on the protected attribute. Therefore, it does not constitute disparate treatment. During training, on the other hand, the protected attribute can be used, for example in a regularizer or in fairness constraints, to find a fair classifier. Lipton et al. (2018) calls such approaches disparate learning processes (DLPs) and questions their treatment parity. In Chapter 3 we cover this topic in more depth.

In the next section, we think about how to find fair classifiers. In this thesis we consider mostly DLPs (Donini et al., 2018; Goh et al., 2016; Manisha and Gujar, 2020; Wu et al., 2019; Zafar et al., 2017a,b) such as regularizer approaches or constraint optimization.

**Impossiblilty Results.** In the recidivism example in Section 1.1 we have seen that the opposing parties were arguing on the basis of two different observational fairness notions. ProPublica's Angwin et al. (2016) argument was based on *separation* by comparing the false positive rates between the protected groups. Equivant on the other hand defended with *sufficiency* by pointing out that the COMPAS risk scores reflect the risk of recidivism equally well for both protected groups. The idea to find risk scores that would satisfy both sides would be in vain: since the ProPublica article various works have found that separation and sufficiency are impossible to fulfill simultaneously (Chouldechova, 2017; Kleinberg et al., 2017). Other impossibility results in the general terms of independence, sufficiency, and separation can be found in Barocas et al. (2019). Kim et al. (2020) provide a tool to diagnose the trade-offs of different combinations of group fairness criteria that are derived from the confusion table such as equal odds, calibration, or demographic parity. This tool allows a general understanding of the incompatibility of group fairness definitions.

## 1.4 Learning Fair Classifiers–Regularizers and Relaxations

Our goal in fair binary classification is to obtain a mapping $h : \mathcal{X} \rightarrow \mathcal{Y}$ that is fair with respect to the protected attribute $S$ while remaining accurate on the target label $Y$. The problem of learning fair classifiers has mainly been addressed in three ways. First, pre-processing approaches alter the labels of the examples or their representation to increase the intrinsic fairness of a dataset. A classifier learned on this modified data is then more likely to be fair (Calmon et al., 2017; Dwork et al., 2012; Feldman et al., 2015; Kamiran and Calders, 2012; Madras et al., 2018; Zemel et al., 2013). Second, post-hoc procedures transform existing accurate but unfair classifiers into fair classifiers (Chzhen et al., 2019; Hardt et al., 2016; Kamiran et al., 2010; Menon and Williamson, 2018; Woodworth et al., 2017).

Finally, direct in-processing methods learn a fair and accurate classifier in a single step (Agarwal et al., 2018; Calders and Verwer, 2010; Cotter et al., 2019; Donini et al., 2018; Goh et al., 2016; Kamishima et al., 2012; Manisha and Gujar, 2020; Wu et al., 2019; Zafar et al., 2017a,b). Among these a straightforward approach is to add a regularizer to the standard training objective (Bechavod and Ligett, 2017; Bendekgey and Sudderth, 2021; Beutel et al., 2019; Kleindessner et al., 2021; Lohaus et al., 2020; Manisha and Gujar, 2020; Risser et al., 2021; Wick et al., 2019). In this chapter, we focus on the in-processing approaches with a focus on regularizers and relaxed formulations of the fairness constraints.

We seek to find an accurate classifier $h^*$ with a fairness disparity $|\Lambda(h)|$ less than a tolerance $\tau \in [0,1]$, where $\Lambda(h)$ is a fairness measure such as the DDP or DEO. Given a function class $\mathcal{F}$, the fair and accurate classifier $h^*$ is the solution to the following problem:

$$\min_{h \in \mathcal{F}} L(h)\,, \text{ such that } |\Lambda(h)| \leq \tau,$$

where $L(h) = \mathbb{E}_{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}}\left[\ell(h(X)\,,Y)\right]$ is the true risk of $h$ for some loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In practice, we only have access to a set $\widehat{\mathcal{D}}_\mathcal{Z} = \{(x_i, s_i, y_i)\}_{i=1}^n$ of $n$ i.i.d. examples drawn from $\mathcal{D}_\mathcal{Z}$. Hence, we consider the empirical version of this problem:

$$\min_{h \in \mathcal{F}} \widehat{L}(h)\,, \text{ such that } \left|\widehat{\Lambda}(h)\right| \leq \tau, \tag{1.8}$$

where $\widehat{L}(h) = \frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_\mathcal{Z}} \ell(h(x)\,,y)$ is the empirical risk and $\widehat{\Lambda}(h)$ the empirical fairness measure. The constrained problem can be rewritten into its Lagrangian (Bendekgey and Sudderth, 2021) with the KKT conditions:

$$\min_{h \in \mathcal{F}} L(h) + \lambda\Lambda(h). \tag{1.9}$$

The main difficulty involved in learning a fair classifier is to ensure that $|\Lambda(h)| \leq \tau$ since the constraints $\Lambda(h)$ are typically discontinuous, although note that there always exists a trivial solution to (1.8) since the constant classifier, i.e. only positive predictions, is fair with respect to the observational criteria above. In order to reduce the computational complexity of (1.8), the constraints can be replaced by continuous and convex *fairness relaxations*. In Chapter 2 we analyze how well the relaxations achieve fairness; in Chapter 3 we employ a relaxation as a regularizer in the form of (1.9).

**Decision Boundary Covariance.** Zafar et al. (2017a) learn convex margin-based classifiers $f_\theta : \mathcal{X} \to \mathcal{Y}$, such as a logistic regression or an SVM, by minimizing a convex loss $L(f_\theta)$ over the parameters $\theta$. In order to avoid increasing the complexity with a non-convex fairness constraint, Zafar et al. (2017a) propose the *decision boundary covariance*, a convex and linear relaxation for demographic parity (specifically, the p%-rule (Biddle, 2005; Equal Employment Opportunity Commission (EEO), 1978)). To that end, Zafar et al. (2017a) consider the signed distance to the decision boundary $d_\theta : \mathcal{X} \to \mathbb{R}$, where $d_\theta \geq 0$ corresponds to $f_\theta = 1$ and $d_\theta < 0$ corresponds to $f_\theta = -1$. Decision boundary fairness introduced by Zafar et al. (2017a) states that the distance to the decision boundary should not correlate with the protected attribute. We require the covariance between decision boundary distance and protected attribute to be small:

$$\begin{aligned} \mathrm{Cov}(S, d_\theta(X)) &= \mathbb{E}\left[(S - \mathbb{E}\left[S\right]) d_\theta(X)\right] - \mathbb{E}\left[S - \mathbb{E}\left[S\right]\right] \mathbb{E}\left[d_\theta(X)\right] \\ &= \mathbb{E}\left[(S - \mathbb{E}\left[S\right]) d_\theta(X)\right] \\ &\propto \mathbb{E}\left[d_\theta(X) \mid S = 1\right] - \mathbb{E}\left[d_\theta(X) \mid S = -1\right]. \end{aligned} \tag{1.10}$$

We can reformulate optimization problem (1.8) and replace the constraint $|\mathrm{DDP}(f)| \leq \tau$ with $|\mathrm{Cov}(S, d_\theta(X))| \leq \tau$ in order to obtain a convex optimization problem in $\theta$. For example in logistic regression or linear SVM, where we have $d_\theta(x) = \theta^T x$, the constraint $\theta^T \left(\frac{1}{n} \sum_{i=1}^n \left[(s_i - \bar{s})\, x_i\right]\right)$ is linear in $\theta$.

**Surrogate Functions for Fairness.** In the following we define a class of fairness relaxations for classifiers of the form $h(x) = \mathbb{1}(d_\theta(x) > 0)$. The DDP measure for demographic parity for instance is then equivalent to

$$\mathrm{DDP}(d_\theta) = \underset{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{1}(d_\theta(X) > 0) | S{=}1\right] - \underset{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{1}(d_\theta(X) > 0) | S{=}{-1}\right]. \tag{1.11}$$

In order to relax the discontinuous indicator function, we simply replace it by a continuous and monotonically non-decreasing surrogate function $g : \mathbb{R} \to \mathbb{R}$ (Agarwal et al., 2018; Bendekgey and Sudderth, 2021; Cotter et al., 2019; Goh et al., 2016; Wu et al., 2019; Zafar et al., 2017b), for example for demographic parity we have

$$\Lambda_g(d_\theta) = \underset{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[g(d_\theta(X))\, | S{=}1\right] - \underset{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[g(d_\theta(X))\, | S{=}{-1}\right]. \tag{1.12}$$

Several other methods can be recovered by choosing $g$. With $g(s) = s$, we recover Equation (1.10). With $g(s) = \min(0, s)$ for instance, we recover the relaxation for equality of opportunity by Zafar et al. (2017b) which is motivated by the decision boundary covariance by Zafar et al. (2017a):

$$\mathrm{Cov}(S, g(d_\theta(X)) \mid Y = 1) \propto \mathbb{E}\left[g(d_\theta) \mid Y = 1, S = 1\right] - \mathbb{E}\left[g(d_\theta) \mid Y = 1, S = -1\right]. \tag{1.13}$$

The constraint is not convex, but if the loss function is convex, a local optimum can be found with convex-concave programming (Zafar et al., 2017b). Zafar et al. (2017b) propose several other surrogate functions as relaxations for matching criteria of misclassification rates.

Based on a theoretical analysis of previous work using relaxations, Bendekgey and Sudderth (2021) propose to use the *sigmoid* $g(s) = \sigma(s)$ with $\sigma(s) = 1/(1 + e^{-s})$ and the *log-sigmoid* $g(s) = -\log \sigma(-s)$. The sigmoid function $\sigma(s) = 1/(1 + e^{-s})$ also recovers the regularizer by Wick et al. (2019) which we use in Chapter 3.

## 1.5 Thesis Contributions

In the introduction, we have discussed the need of policy makers for future research to enforce regulations that are often based on rather vague terms: in a second recommendation, the Enquete commission of the German Parliament demands a "requirement for Transparency, Interpretability and Explainability of AI decisions" (BT Drucksache 19/23700, 2020). To this end, it is essential to understand the proposed solutions to unfair decision-making. In this doctoral thesis, we focus on limitations of fairness methods using relaxed fairness constraints or regularizers, and of observational fairness guarantees when we aggregate various fair models.

In Chapter 2, we investigate various fairness approaches that are formulated as a constrained optimization problem using relaxations of the fairness constraints. We show that many existing relaxations are unsatisfactory: even if a model satisfies the relaxed constraint, it can be surprisingly unfair. We propose a principled framework to solve this problem. This new approach uses a strongly convex formulation and comes with theoretical guarantees on the fairness of its solution. In practice, we show that this method gives promising results on real data.

In Chapter 3, we apply fairness regularizers and a preprocessing method to neural networks on binary classification tasks. We show that deep neural networks that satisfy demographic parity do so through a form of protected group awareness, and that the more we force a network to be fair, the more accurately we can recover the protected attribute from the internal state of the network. Based on this observation, we propose a simple two-stage solution for enforcing fairness. First, we train a two-headed network to predict the protected attribute (such as race or gender) alongside the original task, and second, we enforce demographic parity by taking a weighted sum of the heads. Our two-headed approach has near identical performance compared to the regularization-based or preprocessing method, but has greater stability and higher accuracy where near exact demographic parity is required. To cement the relationship between the regularized and the two-headed approach, we show that an unfair and optimally accurate classifier can be recovered by taking a weighted sum of a fair classifier and a classifier predicting the protected attribute. We use this to argue that the fairness approaches and our explicit formulation demonstrate *disparate treatment* and that, consequentially, they are likely to be unlawful in a wide range of scenarios under the US law.

In Chapter 4, we investigate the fairness properties of the majority vote ensemble when the individual classifiers are already fair. Can we guarantee fair binary decisions from the majority vote when the individual classifiers are fair? We answer this question in two flavors: (1) under strong conditional independence assumptions, we can guarantee fairness of the ensemble, and (2) under no independence assumptions, we cannot guarantee fairness, but we provide worst-case bounds. The first result is based on the well-known Condorcet Jury Theorem, which analyzes the accuracy of the majority vote in the case of independent voters. The worst-case fairness bounds are derived algorithmically using linear program formulations. We find that fairness constraints can help to reduce worst-case scenarios slightly, but overall, the fairness properties of individual classifiers are largely at risk in a majority vote.

### 1.5.1 Publications

This thesis is based on the following publications.

Chapter 2: Lohaus, M., Perrot, M., von Luxburg, U. (2020) Too Relaxed to Be Fair. *In International Conference of Machine Learning (ICML)*. https://proceedings.mlr.press/v119/lohaus20a.html

Chapter 3: Lohaus, M., Kleindessner, M., Kenthapadi, K., Locatello, F., Russell, C. (2022) Are Two Heads the Same as One? Identifying Treatment in Fair Neural Networks. *arXiv preprint arXiv:2204.04440*. https://arxiv.org/abs/2204.04440 (Under submission).

Chapter 4 is based on yet unpublished work.

During my PhD, I also worked on projects about ordinal data and ordinal embedding that were left out of this thesis since they are not related to my main line of work on fairness. I was the main contributor on:

Lohaus, M., Hennig, P., von Luxburg, U. (2019) Uncertainty Estimates for Ordinal Embedding. *arXiv preprint arXiv:1906.11655*. https://arxiv.org/abs/1906.11655

I was one of the main contributors on the following paper.

Chennuru Vankadara*, L., Lohaus*, M., Haghiri, S., Ul Wahab, F., von Luxburg, U. (2021) Insights into Ordinal Embedding Algorithms: A Systematic Evaluation. *arXiv preprint arXiv:1912.01666*. https://arxiv.org/abs/1912.01666 (Under submission).

I co-authored on the following paper on fairness in machine learning as well.

Zietlow, D., Lohaus, M., Balakrishnan, G., Kleindessner, M., Locatello, F., Schölkopf, B., Russell, C. (2022) Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers. *In Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2203.04913

# Chapter 2

# Too Relaxed to Be Fair

In the literature, fair binary classification is often formulated as a constrained optimization problem and solved using relaxations of the fairness constraints. These approaches work reasonably well for some applications. However, their relaxations are quite coarse and we demonstrate that **they can fail to find fair classifiers**. In particular, there is typically no guarantee on the relationship between the relaxed fairness and the true fairness of a solution: a classifier that is perfectly fair in terms of relaxed fairness can be highly unfair in terms of true fairness (see Figure 2.1 for an illustration).

We propose a new principled framework to tackle the problem of fair classification that is particularly relevant for scenarios where formal fairness guarantees are required. Our approach is based on convex relaxations and comes with theoretical guarantees that ensure that the learned classifier is fair up to sampling errors. Furthermore, we use a learning theory framework for similarity-based classifiers to exhibit sufficient conditions that guarantee the existence of a fair and accurate classifier.

## 2.1 Problem Setting

For the purpose of this chapter, we repeat some notation from Chapter 1 to slightly adjust the fairness definitions since in this chapter we have classifiers of the form $h(x) = \operatorname{sign}(f(x))$ where $f : \mathcal{X} \to \mathbb{R}$ is a real valued function. Let $\mathcal{X}$ be a feature space, $\mathcal{Y} = \{-1, 1\}$ a space of binary class labels, and $\mathcal{S} = \{-1, 1\}$ a space of binary protected attributes. In this chapter we denote the random variables $(x, s, y) \sim \mathcal{D}_{\mathcal{Z}}$, which we draw from a distribution $\mathcal{D}_{\mathcal{Z}}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, with lowercase letters. Our goal in fair classification is to obtain a function $f$ such that $h : \mathcal{X} \to \mathcal{Y}$ is fair with respect to the protected attribute while remaining accurate on the class labels.

We revisit *demographic parity* and *equality of opportunity* and define them in terms of the real valued function $f$.

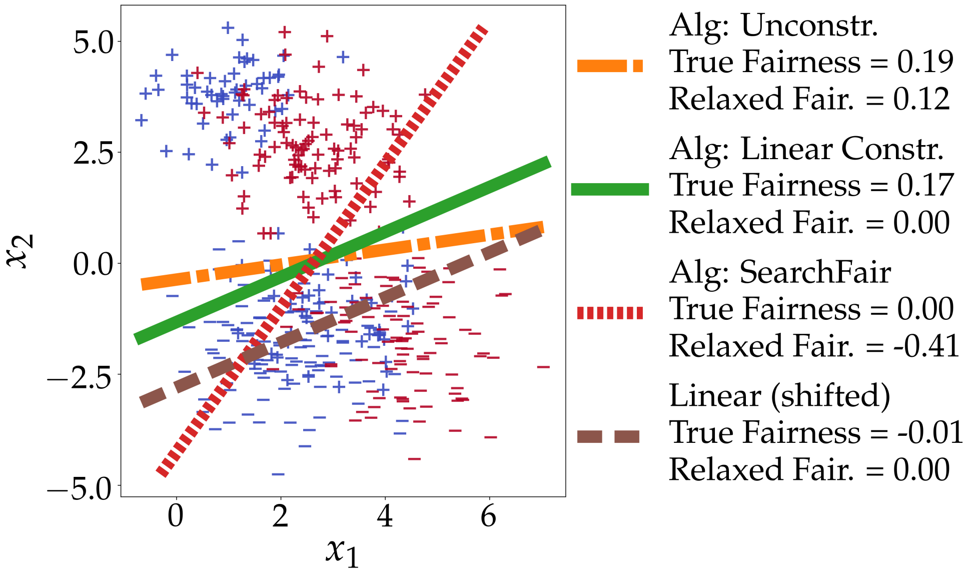


Figure 2.1: The goal is to separate the positive class (+) from the negative class (−) while remaining fair with respect to two protected groups: the blue and the red group. We evaluate the true fairness (DDP) and a linear fairness relaxation (Zafar, Section 2.2.1) of three linear classifiers learned with no fairness constraint (Unconstr., orange), a linear relaxation of the fairness constraint (Linear Constr., green), and our framework (SearchFair, red). We also plot the classifier obtained by translating Linear (Linear (shifted), brown). It has the same relaxed fairness as Linear but a different true fairness: the relaxation is oblivious to the intercept parameter. SearchFair finds the fairest classifier.

**Demographic Parity.** A classifier $f$ is fair for demographic parity when its predictions are independent of the protected attribute (Calders and Verwer, 2010; Calders et al., 2009). Formally, this can be written as

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[f(x)>0|s=1] = \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[f(x)>0|s=-1]\,.$$

In practice, enforcing exact demographic parity might be too restrictive. Instead, we consider a fairness score (Wu et al., 2019) called Difference of Demographic Parity (DDP)

$$\mathrm{DDP}(f) = \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}[\mathbb{1}(f(x)>0)|s=1] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}[\mathbb{1}(f(x)>0)|s=-1]\,, \tag{2.1}$$

where $\mathbb{1}(a)$ is the indicator function that returns 1 when $a$ is true and 0 otherwise. The DDP is positive when the favoured group is $s=1$ and negative when it is $s=-1$. Given a threshold $\tau \geq 0$, we say that a classifier $f$ is $\tau$-DDP fair if $|\mathrm{DDP}(f)| \leq \tau$. When $\tau = 0$, exact demographic parity is achieved and we say that the classifier is DDP fair.

**Equality of Opportunity.** A classifier $f$ is fair for equality of opportunity when its predictions for positively labelled examples are independent of the protected attribute (Hardt et al., 2016). Formally, it is

$$\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[f(x) > 0 | y = 1, s = 1\right] = \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[f(x) > 0 | y = 1, s = -1\right].$$

Again, instead of only considering exact equality of opportunity, we use a fairness score (Donini et al., 2018) called Difference of Equality of Opportunity (DEO):

$$\begin{aligned}\mathrm{DEO}(f) = &\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{1}(f(x) > 0) | y = 1, s = 1\right] \\ &- \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{1}(f(x) > 0) | y = 1, s = -1\right].\end{aligned} \tag{2.2}$$

This quantity is positive when the favoured group is $s = 1$ and negative when it is $s = -1$. Given a threshold $\tau \geq 0$, we say that a classifier $f$ is $\tau$-DEO fair if $|\mathrm{DEO}(f)| \leq \tau$. When $\tau = 0$, exact equality of opportunity is achieved and we say that the classifier is DEO fair.

It is worth noting that demographic parity and equality of opportunity are quite similar from a mathematical point of view. In the remainder of the chapter, we focus our exposition on DDP as results that hold for DDP can often be readily extended to DEO by conditioning on the target label.

**Learning a fair classifier.** Given a function class $\mathcal{F}$, a $\tau$-DDP fair and accurate classifier $f^*$ is given as the solution of the following problem:

$$f^* = \underset{\substack{f\in\mathcal{F} \\ |\mathrm{DDP}(f)|\leq\tau}}{\arg\min}\; L(f),$$

where $L(f) = \mathbb{E}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\ell(f(x), y)\right]$ is the true risk of $f$ for a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In practice, we only have access to a set $\widehat{\mathcal{D}}_\mathcal{Z} = \{(x_i, s_i, y_i)\}_{i=1}^n$ of $n$ examples drawn from $\mathcal{D}_\mathcal{Z}$. Hence, we consider the empirical version of this problem:

$$f^\beta = \underset{\substack{f\in\mathcal{F} \\ |\mathrm{DDP}(f)|\leq\tau}}{\arg\min}\; \widehat{L}(f) + \beta\Omega(f), \tag{2.3}$$

where $\Omega(f)$ is a convex regularization term used to prevent over-fitting, $\beta$ is a trade-off parameter, and $\widehat{L}(f) = \frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_\mathcal{Z}} \ell(f(x), y)$ is the empirical risk. The challenge in learning a fair classifier is to ensure that $|\mathrm{DDP}(f)| \leq \tau$.

## 2.2 When Fairness Relaxations Fail

To obtain a $\tau$-DDP fair classifier, most approaches consider the fully empirical version of Optimization Problem 2.3:

$$\begin{aligned}\min_{f\in\mathcal{F}}\quad &\hat{L}(f) + \beta\Omega(f) \\ \text{subject to } &|\widehat{\mathrm{DDP}}(f)| \leq \tau,\end{aligned} \tag{2.4}$$

where the empirical version of DDP is:

$$\widehat{\mathrm{DDP}}(f) = \frac{1}{n} \sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=1}} \mathbb{1}(f(x) > 0) - \frac{1}{n} \sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=-1}} \mathbb{1}(f(x) > 0).$$

The main issue with this optimization problem is the non-convexity of the constraints that makes it hard to find the optimal solution. A standard approach is then to first rewrite the DDP in an equivalent, but easier to handle form and then replace the indicator functions with a relaxation. Zafar et al. (2017a) and Donini et al. (2018) use a linear relaxation to obtain a fully convex constraint. Zafar et al. (2017b) use a convex relaxation that leads to a convex-concave constraint. Wu et al. (2019) combine a convex relaxation with a concave one to obtain a fully convex problem. Below, we show that these approaches only loosely approximate the true constraint and might lead to suboptimal solutions (see Figure 2.2). Furthermore, when theoretical guarantees accompany the method, they are either insufficient to ensure that the learned classifier is fair (Wu et al., 2019) or they make assumptions that are hard to satisfy in practice (Donini et al., 2018).

### 2.2.1 Linear Relaxations

We first study methods that use a linear relaxation of the indicator function to obtain a convex constraint in Optimization Problem 2.4. First, Zafar et al. (2017a) rewrite the DDP with $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}(s = 1)$ the proportion of individuals in group $s = 1$:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} [\mathbb{1}(f(x) > 0)|s = 1] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} [\mathbb{1}(f(x) > 0)|s = -1] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\frac{s+1}{2}\mathbb{1}(f(x) > 0)|s = 1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\frac{1-s}{2}\mathbb{1}(f(x) > 0)|s = -1\right] \\
&= \frac{1}{p_1} \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\frac{s+1}{2}\mathbb{1}(f(x) > 0)\right] - \frac{1}{1-p_1} \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\frac{1-s}{2}\mathbb{1}(f(x) > 0)\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\left(\frac{s+1}{2p_1} - \frac{1-s}{2(1-p_1)}\right)\mathbb{1}(f(x) > 0)\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\left(\frac{(s+1)(1-p_1) - (1-s)p_1}{2p_1(1-p_1)}\right)\mathbb{1}(f(x) > 0)\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\left(\frac{s+1-2p_1}{2p_1(1-p_1)}\right)\mathbb{1}(f(x) > 0)\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\frac{1}{p_1(1-p_1)}\left(\frac{s+1}{2} - p_1\right)\mathbb{1}(f(x) > 0)\right].
\end{aligned}$$

Then, they consider a linear approximation of $\mathbb{1}(f(x) > 0)$ and obtain the constraint:

$$\left| \frac{1}{n} \sum_{(x,s,y) \in \widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{1}{\hat{p}_1(1-\hat{p}_1)} \left( \frac{s+1}{2} - \hat{p}_1 \right) f(x) \right| \leq \tau, \tag{2.5}$$

where $\hat{p}_1$ is an empirical estimate of $p_1$. In their original formulation, Zafar et al. (2017a) get rid of the factor $\frac{1}{\hat{p}_1(1-\hat{p}_1)}$ by replacing the right hand side of the constraint with $c = \hat{p}_1(1-\hat{p}_1)\tau$.

Similarly, Donini et al. (2018) rewrite the DDP with $p_s = \mathbb{P}_{(x',s',y') \sim \mathcal{D}_{\mathcal{Z}}}(s' = s)$:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ \mathbb{1}(f(x) > 0) | s = 1 \right] - \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ \mathbb{1}(f(x) > 0) | s = -1 \right] \\
&= \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ s\mathbb{1}(f(x) > 0) | s = 1 \right] + \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ s\mathbb{1}(f(x) > 0) | s = -1 \right] \\
&= \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ s\mathbb{1}(f(x) > 0) | s = 1 \right] \frac{p_1}{p_1} + \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ s\mathbb{1}(f(x) > 0) | s = -1 \right] \frac{1-p_1}{1-p_1} \\
&= \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ \frac{s}{p_s} \mathbb{1}(f(x) > 0) | s = 1 \right] p_1 \qquad \text{(Law of total expectation.)} \\
&\quad + \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ \frac{s}{p_s} \mathbb{1}(f(x) > 0) | s = -1 \right] (1 - p_1) \\
&= \mathop{\mathbb{E}}_{(x,s,y) \sim \mathcal{D}_{\mathcal{Z}}} \left[ \frac{s}{p_s} \mathbb{1}(f(x) > 0) \right].
\end{aligned}$$

Then, using the same linear relaxation as Zafar et al. (2017a) of $\mathbb{1}(f(x) > 0)$ and with $\hat{p}_s$, an empirical estimate of $p_s$, they obtain the constraint[1]

$$\left| \frac{1}{n} \sum_{(x,s,y) \in \widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{s}{\hat{p}_s} f(x) \right| \leq \tau. \tag{2.6}$$

Both constraints 2.5 and 2.6 are mathematically close and only differ in terms of the multiplicative factor in front of $f(x)$ in the inner sum. Thus, they can be rewritten as

$$\left| \mathrm{LR}_{\widehat{\mathrm{DDP}}}(f) \right| = \left| \frac{1}{n} \sum_{(x,s,y) \in \widehat{\mathcal{D}}_{\mathcal{Z}}} C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right) f(x) \right| \leq \tau.$$

where $C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right)$ can be chosen to obtain any of the two constraints. In the following, we use this general formulation to show that both formulations have shortcomings that can lead to undesired behaviors.

[1] Donini et al. (2018) originally consider $\tau$-DEO fairness rather than DDP. In the constraint, instead of drawing the examples from $\mathcal{D}_{\mathcal{Z}}$, they use the conditional distribution $\mathcal{D}_{\mathcal{Z}|y=1}$. However, this does not change the intrinsic nature of the constraint, and the issues raised here remain valid.

**Linear relaxations are too loose.** In Figures 2.2a and 2.2b we illustrate the behaviors of $\widehat{\mathrm{DDP}}(f)$ and $\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f)$. In the figures, we consider linear classifiers of the form $f(x) = -x_2 + a_1 x_1 + a_0$ where $a_1$ controls the slope of the classifier and $a_0$ the intercept. The underlying data is the same as in Figure 2.1. It shows that the linear relaxation of DDP can behave completely differently compared to the true DDP. It is particularly striking to notice that the intercept does not have any influence on the constraint. This behavior can be formally verified.

Let $f$ be a predictor of the form $f(x) = g(x) + b$ where $b$ is the intercept. Then, $\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f)$ is independent of changes in $b$ since $\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} C\left(s,\widehat{\mathcal{D}}_{\mathcal{Z}}\right) = 0$ for both constraints presented above. For the formulation of Donini et al. (2018) with $C\left(s,\widehat{\mathcal{D}}_{\mathcal{Z}}\right) = \frac{s}{\hat{p}_s}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{s}{\hat{p}_s} &= \sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{s}{n_s}\\ &= \sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_1} - \sum_{(x,s=-1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_{-1}} = 0.\end{aligned}$$

Recall that $\hat{p}_s = \frac{n_s}{n}$, where $n_s$ is the number of samples with protected attribute $s$.

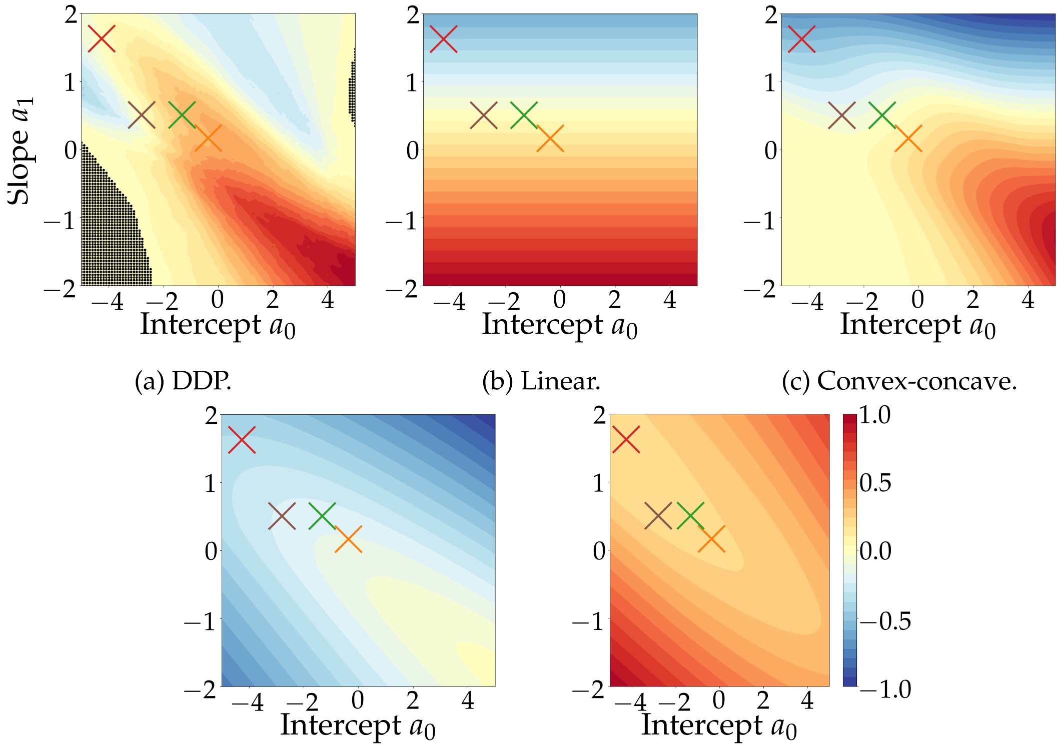
Second, we consider Zafar et al. (2017a) with $C\left(s,\widehat{\mathcal{D}}_{\mathcal{Z}}\right) = \frac{1}{\hat{p}_1(1-\hat{p}_1)}\left(\frac{s+1}{2} - \hat{p}_1\right)$.

$$\begin{aligned}\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{\hat{p}_1(1-\hat{p}_1)}\left(\frac{s+1}{2}-\hat{p}_1\right) &= n\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_1 n_{-1}}\left(\frac{s+1}{2}-\frac{n_1}{n}\right)\\ &= \frac{n}{n_1 n_{-1}}\left(\sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left(1-\frac{n_1}{n}\right) - \sum_{(x,s=-1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{n_1}{n}\right)\\ &= \frac{n}{n_1 n_{-1}}\left(\sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left(\frac{n_{-1}}{n}\right) - \frac{n_1 n_{-1}}{n}\right) = 0.\end{aligned}$$

**Theoretical guarantees for linear relaxations are not satisfactory.** Donini et al. (2018) study a sufficient condition under which the linear fairness relaxation $\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f)$ of a function $f$ is close to its true fairness, that is it holds that $\left|\widehat{\mathrm{DDP}}(f)\right| \leq \left|\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f)\right| + \hat{\Delta}$. The condition that needs to be satisfied by $f$ is

$$\frac{1}{2}\sum_{s'\in\{-1,1\}}\left|\frac{1}{2}\sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=s'}}(\mathrm{sign}(f(x)) - f(x))\right| \leq \hat{\Delta}.$$

Unfortunately, the left hand side of this condition is non-convex and thus, it is difficult to use in practice. In particular, when they learn a classifier with their linear relaxation,



(a) DDP.

(b) Linear.

(c) Convex-concave.

(d) Wu - Lower.

(e) Wu - Upper.

Figure 2.2: Consider linear classifiers for the dataset in Figure 2.1. The decision boundaries are of the form $x_2 = a_1 x_1 + a_0$ where $a_1$ controls the slope and $a_0$ the intercept. For given intercepts and slopes, we plot normalized values of (a) the DDP score (yellow is fair), (b) the linear relaxation (Section 2.2.1), (c) the convex-concave relaxation (Section 2.2.2), (d) the concave Wu lower bound, and (d) the convex Wu upper bound (Section 2.2.2). The black dotted area in (a) corresponds to trivial constant classifiers—the predicted class is the same for all points. The colored crosses correspond to the classifiers in Figure 2.1. A good relaxation should capture the true DDP reasonably well, in particular the yellow regions should match. However, none of the considered relaxations manage to achieve this.

Donini et al. (2018) do not ensure that it also has a small $\hat{\Delta}$. They only verify this condition when the learning process is over, that is when a classifier $f$ has already been produced. However, at this time, it is also possible to compute $\widehat{\text{DDP}}(f)$ directly, so the bound is not needed anymore.

If one could show that for a given function class $\mathcal{F}$, there exists a small $\hat{\Delta}$ such that the condition holds for all $f \in \mathcal{F}$, then any classifier learned from this function class would be guaranteed to be fair when $\left|\text{LR}_{\widehat{\text{DDP}}}(f)\right|$ is small. However, it is not clear whether such function class exists. Nevertheless, this argument hints that for linear relaxations of the fairness constraint, the complexity of the function class largely controls the DDP that can be achieved.

**Linear relaxations should not be combined with complex classifiers.** We demonstrate that, if the class of classifiers $\mathcal{F}$ is complex, then the linear relaxation constraint has almost no influence on the outcome of the optimization problem. In Figure 2.3, we compare the performance, in terms of empirical DDP and accuracy, of several models learned by Optimization Problem 2.4 equipped with the linear relaxation for different parameters $\beta$ (for regularization) and $\tau$ (for fairness). Intuitively, one would expect that varying $\tau$ leads to changes in the fairness level while varying $\beta$ leads to changes in accuracy. However, this is not the case: $\tau$ only has an effect on the result when $\beta$ is sufficiently large. It means that the fairness of the model is mainly controlled by the regularization parameter rather than the fairness one.

This would not be an issue if the fairness of complex classifiers was small. Unfortunately, high-complexity models have a high capacity to alter their decision boundaries. It means that to achieve both high accuracy and high fairness at the same time, they tend to alter their prediction margin for a few selected examples. While this might not affect the accuracy by a lot, the linear relaxation is sensible to this kind of changes and thus can be largely improved—which is what the optimization aims for. However, altering labels of individual points does not have a big influence on the true DDP: it remains high. This effect is reduced when one learns models of low capacity, which have less freedom to deliberately change labels of individual points. Overall, linear relaxations are mainly relevant for simple classifiers and tend to have little effect on complex ones. We outline this undesirable behavior in the experiments.

### 2.2.2 Other Relaxations

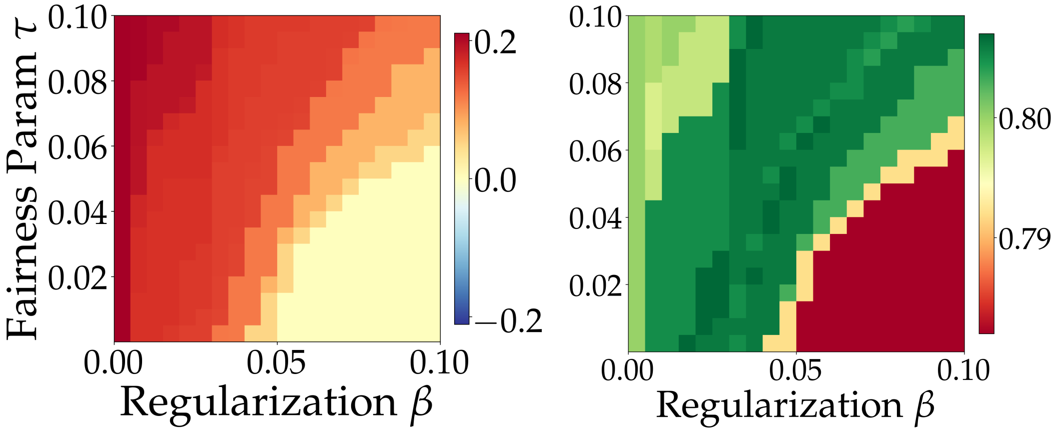
In the previous section we demonstrated that linear relaxations are not sufficient to ensure fairness of the learned classifier. We now focus on two approaches that use non-linear relaxations of the indicator function to stay close to the original fairness definition.

**Convex-concave relaxation.** In a second paper, Zafar et al. (2017b) use the same fairness formulation as Zafar et al. (2017a), but, instead of a linear relaxation of the indicator function, they use a non-linear relaxation.[2] Hence, given $\hat{p}_1$ defined as in Section 2.2.1, they obtain the constraint:

$$\left|\mathrm{CCR}_{\widehat{\mathrm{DDP}}}(f)\right| = \left|\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{\left(\frac{s+1}{2}-\hat{p}_1\right)}{\hat{p}_1(1-\hat{p}_1)}\min\left(0, f(x)\right)\right| \leq \tau.$$

In Figure 2.2c we give an illustration of $\mathrm{CCR}_{\widehat{\mathrm{DDP}}}(f)$. It more closely approximates the original $\widehat{\mathrm{DDP}}(f)$ than the linear relaxation. Nevertheless, it remains quite far from

[2] Zafar et al. (2017b) originally consider other notions of fairness than DDP, among them is the $\tau$-DEO fairness (Equation (5) in their paper). Instead of drawing the examples from $\mathcal{D}_{\mathcal{Z}}$, they consider the conditional distribution $\mathcal{D}_{\mathcal{Z}|y=1}$.



(a) DDP.

(b) Accuracy.

Figure 2.3: We consider a similarity-based classifier (Section 2.4) with rbf kernel and 1000 train and test points from the Adult dataset. Using a varying regularization parameter $\beta$ and fairness parameter $\tau$, we train several classifiers using the linear fairness relaxation (Section 2.2.1). We plot the empirical test DDP of the learned models in Figure 2.3a (red and blue are bad, yellow is good) and their accuracy in Figure 2.3b (red is bad, green is good). We can see that, if $\beta$ is small (complex model), the fairness relaxation parameter $\tau$ has no influence on the DDP score. For higher values of $\beta$ (simpler models), decreasing $\tau$ improves the DDP. Best viewed in color.

the original definition—in particular for classifiers that are not constant. Moreover, using such a convex relaxation leads to a convex-concave problem that turns out to be difficult to optimize without guarantees on the global optimality.

**Lower-upper relaxation with guarantees.** To derive their fairness constraint, Wu et al. (2019) propose to first equivalently rewrite the DDP as follows:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{1}(f(x)>0)|s=1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{1}(f(x)>0)|s=-1\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s}{p_s}\mathbb{1}(f(x)>0)\right] \qquad \text{(Formulation of Donini et al. (2018).)} \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{\mathbb{1}(s=1)}{p_1}\mathbb{1}(f(x)>0) - \frac{\mathbb{1}(s=-1)}{1-p_1}\mathbb{1}(f(x)>0)\right] \\
&= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{\mathbb{1}(s=1)}{p_1}\mathbb{1}(f(x)>0) + \frac{\mathbb{1}(s=-1)}{1-p_1}\mathbb{1}(f(x)<0) - 1\right],
\end{aligned}$$

where $p_1$ is defined as in Section 2.2.1. Replacing the indicator functions with a convex surrogate other than the linear one would lead to a convex-concave problem due to the absolute value in the constraint. Instead, Wu et al. (2019) propose to use a convex surrogate function $\kappa$ for the requirement $\mathrm{DDP}(f) < \tau$ and a concave surrogate function $\delta$ for $\mathrm{DDP}(f) > -\tau$. The corresponding relaxation is

$$\mathrm{DDP}_\kappa(f) = \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}} \left[ \frac{\mathbb{1}(s=1)}{p_1}\kappa(f(x)) + \frac{\mathbb{1}(s=-1)}{1-p_1}\kappa(-f(x)) - 1 \right],$$

and $\mathrm{DDP}_\delta(f)$ is defined analogously by simply replacing $\kappa$ with $\delta$. It leads to the following convex problem:

$$\begin{aligned} \min_{f\in\mathcal{F}} \quad & \hat{L}(f) + \beta\Omega(f) \\ \text{subject to} \quad & \widehat{\mathrm{DDP}}_\kappa(f) \leq \tau_\kappa \\ & -\widehat{\mathrm{DDP}}_\delta(f) \leq \tau_\delta. \end{aligned} \tag{2.7}$$

Individually, the relaxations are far from the original fairness constraint (as illustrated in Figures 2.2e and 2.2d) but the idea is that combining the upper bound and the lower bound will help to learn a fair classifier. However, one needs to choose $\tau_\kappa$ and $\tau_\delta$ appropriately. To address this, Wu et al. (2019) show that choosing

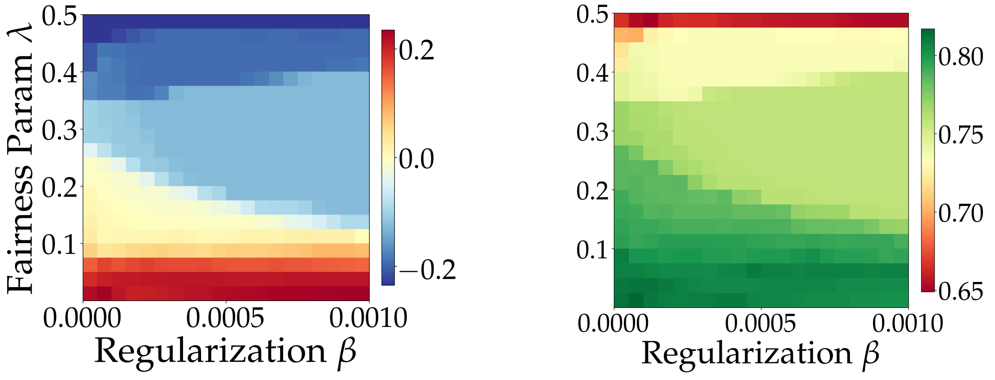
$$\begin{aligned} \tau_\kappa &= \psi_\kappa\left(\tau_{\text{upper}} - \widehat{\mathrm{DDP}}^{+}\right) + \widehat{\mathrm{DDP}}_\kappa^{-} \\ \tau_\delta &= \psi_\delta\left(\tau_{\text{lower}} + \widehat{\mathrm{DDP}}^{-}\right) + \widehat{\mathrm{DDP}}_\delta^{+}, \end{aligned}$$

guarantees that $-\tau_{\text{lower}} \leq \widehat{\mathrm{DDP}}(f) \leq \tau_{\text{upper}}$. Here $\widehat{\mathrm{DDP}}^{+}$ and $\widehat{\mathrm{DDP}}^{-}$ are the worst possible scores of $\widehat{\mathrm{DDP}}(f)$: they are attained by those functions in the given function class that advantage either group $s = -1$ or group $s = 1$ the most. The values $\widehat{\mathrm{DDP}}_\kappa^{-}$ and $\widehat{\mathrm{DDP}}_\delta^{+}$ are defined in the same way for the relaxed scores. The functions $\psi_\kappa$ and $\psi_\delta$ are invertible functions that depend on the selected surrogate.

While this solution is appealing at a first glance, it turns out that Optimization Problem 2.7 is often infeasible for meaningful values of $\tau_{\text{upper}}$ and $\tau_{\text{lower}}$ as the constraints form disjoint convex sets. To illustrate this, consider $\kappa(x) = \max\{0, 1+x\}$ and $\delta(x) = \min\{1, x\}$ as proposed by Wu et al. (2019) and the dataset used in Figure 2.1. Then, if $\tau_{\text{upper}} = \tau_{\text{lower}} \leq 1.13$, the problem is infeasible. If $\tau_{\text{lower}} = 0$ and $\tau_{\text{upper}} \leq 1.95$ the problem is also infeasible. Overall, the guarantees are often meaningless: they either make statements about the empty set (no feasible solution) or they are too loose to ensure meaningful levels of fairness.

## 2.3 SearchFair: A New Approach with Guaranteed Fairness

In the previous section, we have seen that existing approaches use relaxations of the fairness constraint that lead to tractable optimization problems but have little control over the true fairness of the learned model. For this reason, we propose a new framework that solves the problem of finding *provably fair* solutions: given a convex



(a) DDP.

(b) Accuracy.

Figure 2.4: We consider a similarity-based classifier (Section 2.4) with rbf kernel and 1000 train and test points from the Adult dataset. Using a varying regularization parameter $\beta$ and fairness parameter $\lambda$, we train several classifiers using Optimization Problem 2.8 with the same loss, convex relaxation, and regularization as SearchFair in the experiments. We plot the empirical test DDP of the learned models in (a) (red and blue are bad, yellow is good) and their accuracy in (b) (red is bad, green is good). We can see that, given a fixed regularization $\beta$, we can move from positive DDP (small $\lambda$, in red) to a negative DDP (large $\lambda$, in blue) with a region of perfect fairness in between (in yellow).

approximation of the fairness constraint, our method is guaranteed to find a classifier with a good level of fairness.

We consider the following optimization problem:

$$f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)\,, \tag{2.8}$$

where $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ is a convex approximation of the signed fairness constraint, that is we do not consider the usual absolute value. In other words, we obtain a trade-off between accuracy and fairness that is controlled by two hyperparameters $\lambda \geq 0$ and $\beta > 0$ and, given $\beta$ fixed, we can vary $\lambda$ to move from strongly preferring one group to strongly preferring the other group. Our goal is then to find a parameter setting that is in the neutral regime and does not favor any of the two groups. The main theoretical ingredient for this procedure to succeed is the next theorem, which states that the function $\lambda \mapsto \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ is continuous under reasonable assumptions on the data distribution, the candidate classifiers, and the convex relaxation.

**Theorem 1** (Continuity of $\mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_{\Lambda} = \left\{f \in \mathcal{F} : \exists \lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that the following conditions hold:*

*(i)* *Optimization Problem 2.8 is m-strongly convex in $f$,*

*(ii)* $\forall f \in \mathcal{F}$ *,* $R_{\widehat{DDP}}(f)$ *is bounded in* $[-B, B]$*,*

*(iii)* $\exists \rho$*, a metric, such that* $(\mathcal{F}_\Lambda, \rho)$ *is a metric space,*

*(iv)* $\forall x \in \mathcal{X}, g(f) : f \mapsto f(x)$ *is continuous,*

*(v)* $\forall f \in \mathcal{F}_\Lambda$*,* $f$ *is Lebesgue measurable and the set* $\{x : (x, s, y) \in \mathcal{Z}, s = 1, f(x) = 0\}$ *is a Lebesgue null set, as well as* $\{x : (x, s, y) \in \mathcal{Z}, s = -1, f(x) = 0\}$*,*

*(vi)* *the probability density functions* $f_{\mathcal{Z}|s=1}$ *and* $f_{\mathcal{Z}|s=-1}$ *are Lebesgue-measurable.*

*Then, the function* $\lambda \mapsto DDP\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ *is continuous.*

The proof of this theorem is given in the next section. The conditions (i) - (vi) are of a technical nature, but not very restrictive: Condition (i) can be satisfied by using a strongly convex regularization term, for example the squared $L_2$ norm. Condition (ii) can be satisfied by assuming that $\mathcal{X}$ is bounded. Condition (iii) is, for example, satisfied by any Hilbert Space equipped with the standard dot product. This includes, but is not restricted to, the set of linear classifiers. Condition (iv) ensures that small changes in the hypothesis, with respect to the metric associated to $\mathcal{F}$, also yield small changes in the predictions. Condition (v) ensures that the number of examples for which the predictions are zero is negligible, for example this happens when the decision boundary is sharp. Condition (vi) is satisfied by many usual distributions, for example the Gaussian distribution.

We demonstrate the continuous behavior of DDP on a real dataset in Figure 2.4. We plot the DDP score and the accuracy of classifiers learned with Optimization Problem 2.8 for varying parameters $\lambda$ and $\beta$. Given a fixed $\beta$, the results support our theoretical findings: there is a smooth transition between favouring the group $s = 1$ with small $\lambda$ and favouring the group $s = -1$ with higher $\lambda$. In between, there is always a region of perfect fairness. In the next corollary, we formally state the conditions necessary to ensure the existence of such a DDP-fair classifier.

**Corollary 1** (Existence of a DDP-fair classifier)**.** *Assume that the conditions of Theorem 1 hold and that the convex approximation* $R_{\widehat{DDP}}(f)$ *is chosen such that for Optimization Problem* (2.8) *there exist*

*(i)* $\lambda_+$ *such that* $DDP\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_+)\right) > 0$*,*

*(ii)* $\lambda_-$ *such that* $DDP\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_-)\right) < 0$*.*

*Then, there exists at least one value* $\lambda_0$ *in the interval* $[\min(\lambda_+, \lambda_-), \max(\lambda_+, \lambda_-)]$ *such that* $DDP\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$*.*

*Proof.* This corollary is a direct consequence of the intermediate value theorem and the continuity of DDP proven in Theorem 1. □

This suggests a very simple framework to learn provably fair models. First, we choose a convex fairness relaxation (e.g. the one proposed by Wu et al. (2019)). Then, we choose a lower bound $\lambda_{\text{min}}$ and an upper bound $\lambda_{\text{max}}$ that fulfill the assumptions of Corollary 1, that is $\text{sign}\left(\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\right)\right) \neq \text{sign}\left(\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\right)\right)$ (empirically, $\lambda = 0$ and 1 are good candidates). Then, we use a binary search to find a $\lambda_0$ between $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ such that $\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$. We call this procedure *SearchFair* and summarize it in Algorithm 1.

Finally, SearchFair theoretically requires to evaluate the true population fairness $\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ on the underlying distribution $\mathcal{D}_{\mathcal{Z}}$. In practice, we follow the example of existing fairness constraints (Woodworth et al., 2017) and simply approximate this quantity by its empirical counterpart $\widehat{\text{DDP}}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$. If $\text{R}_{\widehat{\text{DDP}}}(f)$ is chosen correctly then setting $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$ usually works. The number of iterations $C$ is used to control how close to 0 the fairness measure should be. Note that, instead of a number of iterations, it is also possible to choose a stopping criterion, for example when DDP falls below a threshold.

**Example of convex relaxation.** One example of a convex relaxation is to use the bounds proposed by Wu et al. (2019). When no fairness regularizer is used, we evaluate the fairness of the resulting classifier and choose an approximation accordingly. More precisely, with $\lambda = 0$ if $\text{DDP}(f(\lambda)) > 0$ we use the upper bound with hinge loss:

$$\text{R}_{\widehat{\text{DDP}}}(f) = \frac{1}{n} \sum_{(x,s,y) \in \widehat{\mathcal{D}}_{\mathcal{Z}}} \left[ \frac{\mathbb{1}(s=1)}{p_1} \max(0, 1 + f(x)) + \frac{\mathbb{1}(s=-1)}{1-p_1} \max(0, 1 - f(x)) - 1 \right].$$

If $\text{DDP}(f(\lambda)) < 0$, we use the negative lower bound with hinge loss:

$$\text{R}_{\widehat{\text{DDP}}}(f) = -\frac{1}{n} \sum_{(x,s,y) \in \widehat{\mathcal{D}}_{\mathcal{Z}}} \left[ \frac{\mathbb{1}(s = 1)}{p_1} \min(1, f(x)) - \frac{\mathbb{1}(s = -1)}{1-p_1} \min(1, -f(x)) + 1 \right].$$

With $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$ this choice often ensures that $\text{sign}(\text{DDP}(f(\lambda_{\text{min}}))) \neq \text{sign}(\text{DDP}(f(\lambda_{\text{max}})))$. We use this approach in all our experiments in this chapter.

Note that we give an example where the relaxations are in fact upper and lower bounds of the DDP score. However, we want to stress that any convex *approximation* would work as long as the conditions of Corollary 1 like $\text{sign}\left(\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\right)\right) \neq \text{sign}\left(\text{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\right)\right)$ are respected.

**Satisfying the conditions of Theorem 1.** The strong convexity of the optimization problem (condition (i)) can be ensured by choosing a strongly convex regularization term (we adopt this strategy in our experiments).

Satisfying conditions (ii) to (v) mainly depends on our choice of function class $\mathcal{F}$. For example, linear classifiers satisfy all the conditions as long as $\mathcal{X}$ is bounded

**Algorithm 1** SearchFair: A binary search framework for fairness

**Input**: A set $\widehat{\mathcal{D}}_{\mathcal{Z}} = (x_i, s_i, y_i)_{i=1}^{n}$ of $n$ labelled examples, a regularization parameter $\beta > 0$, $\lambda_{\min}$ and $\lambda_{\max}$ the lower and upper bounds for $\lambda$, a convex fairness regularizer $\mathrm{R}_{\widehat{\mathrm{DDP}}}(\cdot)$, a number of iterations $C$.
**Output**: A fair classifier.

```
if $\mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\right) > 0$ and $\mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\right) < 0$ then
  $\lambda_{+} = \lambda_{\min}$ and $\mathrm{DDP}_{+} = \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\right)$.
  $\lambda_{-} = \lambda_{\max}$ and $\mathrm{DDP}_{-} = \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\right)$.
  search_possible = True
else if $\mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\right) < 0$ and $\mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\right) > 0$ then
  $\lambda_{-} = \lambda_{\min}$ and $\mathrm{DDP}_{-} = \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\right)$.
  $\lambda_{+} = \lambda_{\max}$ and $\mathrm{DDP}_{+} = \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\right)$.
  search_possible = True
else
  search_possible = False
end if
if search_possible then
  for $c = 1, \ldots, C$ do
    $\lambda = \frac{1}{2}(\lambda_{-} + \lambda_{+})$
    $\mathrm{DDP}_{\lambda} = \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$.
    if $\mathrm{DDP}_{\lambda} > 0$ then
      $\lambda_{+} = \lambda$ and $\mathrm{DDP}_{+} = \mathrm{DDP}_{\lambda}$.
    else
      $\lambda_{-} = \lambda$ and $\mathrm{DDP}_{-} = \mathrm{DDP}_{\lambda}$.
    end if
  end for
  if $|\mathrm{DDP}_{-}| < |\mathrm{DDP}_{+}|$ then
    return $f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{-})$
  else
    return $f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{+})$
  end if
else
  Either choose new values for $\lambda_{\min}$ and $\lambda_{\max}$, or choose a new fairness regularizer
  $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$.
end if
```

(which is the case in most machine learning applications) and the classifier $f^0(x) = \mathbf{0}^T x$, where $\mathbf{0}$ is the vector of all zeros, is not part of the set of learnable functions $\mathcal{F}_{\Lambda}$

(otherwise condition (v) would be violated). To verify that $f^0 \notin \mathcal{F}_\Lambda$, it is sufficient to verify that the equation $\frac{d\hat{L}(f^0)}{df} + \lambda \frac{d\mathrm{R}_{\widehat{\mathrm{DDP}}}(f^0)}{df} + \beta \frac{d\Omega(f^0)}{df} = \mathbf{0}$ with $\beta$ fixed has no solutions for $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Note that, in practice, this is usually easy to verify and can be achieved by correctly choosing $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$. Note that the similarity-based classifiers that we use in our experiments are a particular form of linear classifiers and thus satisfy conditions (ii) to (v).

Finally, condition (vi) depends on the data distribution and should be satisfied for most non-degenerate problems.

### 2.3.1 Proof of Theorem 1

To prove Theorem 1 , that is to show the continuity of the function $\lambda \mapsto \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$, we need technical Lemmas 1 and 2. The first one shows that the function $\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)$ is continuous. The second one shows that for particular function classes, $f \mapsto \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[f(x) \leq 0]$ is a continuous function. Before proving them, we first recall the definition of a $m$-strongly convex function.

**Definition 1** ($m$-strongly convex functions)**.** *A function $f : X \mapsto \mathbb{R}$ is called $m$-strongly convex with parameter $m > 0$ if for all $x, y \in X$ and $t \in [0, 1]$*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{m}{2}t(1-t)\left\|x - y\right\|_2^2 .$$

We can now prove our two technical lemmas.

**Lemma 1** (Continuity of $\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)$)**.** *Assume that Optimization Problem 2.8 is $m$-strongly convex and that $R_{\widehat{\mathrm{DDP}}}(f)$ is bounded in the interval $[-B, B]$. Given a training set $\widehat{\mathcal{D}}_\mathcal{Z}$ and a regularization parameter $\beta > 0$, the function:*

$$\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)$$

*is continuous and there exists a constant $C = \sqrt{\frac{8B}{m}}$ such that:*

$$\left\| f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda) - f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda') \right\|_\mathcal{F} \leq C\sqrt{|\lambda - \lambda'|}.$$

*Proof.* Let $g^\lambda(f) = \hat{L}(f) + \lambda\mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)$ and $g^{\lambda'}(f) = g^\lambda(f) + \varepsilon\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ with $\varepsilon > 0$ and $\varepsilon = \lambda' - \lambda$. For the sake of readability, for the remainder of the proof, we write $f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)$ as $f(\lambda)$. Since Optimization Problem 2.8 is $m$-strongly convex, it holds that:

$$\begin{aligned}
&g^\lambda\left(tf(\lambda) + (1-t)f(\lambda')\right) + \varepsilon\mathrm{R}_{\widehat{\mathrm{DDP}}}\left(tf(\lambda) + (1-t)f(\lambda')\right) \\
&\leq tg^\lambda(f(\lambda)) + (1-t)g^\lambda\left(f(\lambda')\right) + t\varepsilon\mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) + (1-t)\varepsilon\mathrm{R}_{\widehat{\mathrm{DDP}}}\left(f(\lambda')\right) \\
&\quad - \frac{m}{2}t(1-t)\left\|f(\lambda) - f(\lambda')\right\|_\mathcal{F}^2 .
\end{aligned}$$

In particular, for $t = \frac{1}{2}$:

$$\begin{aligned}
&\frac{m}{8}\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \\
&\leq \frac{1}{2}g^{\lambda}(f(\lambda)) + \frac{1}{2}g^{\lambda}(f(\lambda')) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda')) \\
&\quad - g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) - \varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\leq \frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\quad + \frac{1}{2}g^{\lambda}(f(\lambda')) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda')) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\quad - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\quad + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\leq \frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\quad + \frac{1}{2}g^{\lambda'}(f(\lambda')) - \frac{1}{2}g^{\lambda'}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
&\quad + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right).
\end{aligned}$$

Since $f(\lambda)$ and $f(\lambda')$ respectively minimize $g^{\lambda}(f)$ and $g^{\lambda'}(f)$, it holds that

$$\begin{aligned}
\frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) &\leq 0 \\
\frac{1}{2}g^{\lambda'}(f(\lambda')) - \frac{1}{2}g^{\lambda'}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) &\leq 0
\end{aligned}$$

which, in turns, implies

$$\begin{aligned}
& \frac{m}{8}\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
\Leftrightarrow \quad & \left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{8}{2m}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{8}{2m}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\
& \hfill (\mathrm{R}_{\widehat{\mathrm{DDP}}}(f) \in [-B, B]) \\
\Leftrightarrow \quad & \left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{8B}{m}\varepsilon \qquad (\varepsilon \leq |\lambda' - \lambda|) \\
\Rightarrow \quad & \left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{8B}{m}\left|\lambda' - \lambda\right| \\
\Rightarrow \quad & \left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}} \leq \sqrt{\frac{8B}{m}}\sqrt{|\lambda' - \lambda|}.
\end{aligned}$$

Choosing $C = \sqrt{\frac{8B}{m}}$ concludes the proof. $\square$

**Lemma 2** (Continuity of $f \mapsto \mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}[f(x) \leq 0]$)**.** *Let $\mathcal{F}$ be a space of real valued functions $f : \mathcal{X} \to \mathbb{R}$. Assume that the following conditions hold:*

*(i) there exists a metric $\rho$ such that $(\mathcal{F}, \rho)$ is a metric space,*

*(ii) $\forall x \in \mathcal{X}$, the function $g(f) : f \mapsto f(x)$ is continuous,*

*(iii) $\forall f \in \mathcal{F}$, $f$ is Lebesgue measurable and the set $\{x : x \in \mathcal{X}, f(x) = 0\}$ is a Lebesgue null set,*

*(iv) the probability density functions $f_\mathcal{X}$ is Lebesgue-measurable.*

*We have that:*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{D}_\mathcal{X}}[f(x) \leq 0]$$

*is a continuous function in $f \in \mathcal{F}$.*

*Proof.* We have that:

$$\mathop{\mathbb{P}}_{x \sim \mathcal{D}_\mathcal{X}}[f(x) \leq 0] = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(f(x) \leq 0)] = \int_\mathcal{X} \mathbb{1}(f(x) \leq 0) f_\mathcal{X}(x)\, dx = \int_\mathcal{X} h(f, x)\, dx.$$

To show that this function is continuous, we apply Theorem 5.6 in Elstrodt (1996). To this extent, we need to show that all the conditions hold.

- **Condition a:** $\forall f \in \mathcal{F}, h(f, \cdot) \in \mathcal{L}^1$.
  The function $f(x) \mapsto \mathbb{1}(f(x) \leq 0)$ is Borel measurable and the function $f$ is Lebesgue measurable. By composition, the function $x \mapsto \mathbb{1}(f(x) \leq 0)$ is also Lebesgue measurable. As the product of two Lebesgue measurable functions, $h$ is also Lebesgue measurable. Furthermore, we have:

  $$\int_\mathcal{X} |h(f, x)|\, dx \leq \int_\mathcal{X} f_\mathcal{X}(x)\, dx = 1 < \infty$$

  which is the desired condition.

- **Condition b:** $h(\cdot, x)$ is continuous in $f_0 \in \mathcal{F}$ for $\mu$-almost all $x \in \mathcal{X}$.
  Since $\forall x \in \mathcal{X}, g(f) : f \mapsto f(x)$ is continuous in $f_0$, $\mathbb{1}(f(x) \leq 0)$ is also a continuous function in $f_0$ expect for the set $\{x : x \in \mathcal{X}, f(x) = 0\}$ which is a Lebesgue null set.

- **Condition c:** There exists a neighbourhood $U$ of $f_0$ and an integrable function $u : \mathcal{X} \to [0, \infty)$ such that $\forall f \in U$ we have $h(f, \cdot) \leq u$ $\mu$-a.e..
  Taking $u = f_\mathcal{X}$ satisfy the condition with $U = \mathcal{F}$.

Since all the conditions hold, we have that $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}[f(x) \leq 0]$ is continuous at $f_0$. Furthermore, given our assumptions on $\mathcal{F}$, this remains true $\forall f_0 \in \mathcal{F}$. This concludes the proof. $\square$

We are now ready to prove Theorem 1.

*Proof.* Recall that DDP is defined as follows:

$$\mathrm{DDP}(f) = \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=1}}{\mathbb{P}}\left[f(x) > 0\right] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=-1}}{\mathbb{P}}\left[f(x) > 0\right].$$

Applying Lemma 2, we have that $c : \mathcal{F}_\Lambda \to \mathbb{R}$, $c(f) = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=1}}\left[f(x) > 0\right]$ and $c' : \mathcal{F}_\Lambda \to \mathbb{R}$, $c'(f) = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=-1}}\left[f(x) > 0\right]$ are continuous functions. It implies that the function $q : \mathcal{F}_\Lambda \to \mathbb{R}$ defined as $q(f) = \mathrm{DDP}(f)$ is continuous.

Then, using Lemma 1 and recalling that the composition of two continuous functions is also continuous gives the theorem. □

We use the same proof technique to prove the continuity of DEO as stated in the next theorem. The main differences are in conditions (v) and (vi) where we only need to consider the positively labelled examples.

**Theorem 2** (Continuity of $\mathrm{DEO}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \left\{f \in \mathcal{F} : \exists\lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that the following conditions hold:*

*(i) Optimization Problem 2.8 is m-strongly convex in $f$,*

*(ii) for all $f \in \mathcal{F}$, $R_{\widehat{DDP}}(f)$ is bounded in the interval $[-B, B]$,*

*(iii) there exists a metric $\rho$ such that $(\mathcal{F}_\Lambda, \rho)$ is a metric space,*

*(iv) $\forall x \in \mathcal{X}$, the function $g(f) : f \mapsto f(x)$ is continuous,*

*(v) $\forall f \in \mathcal{F}_\Lambda$, $f$ is Lebesgue measurable and the sets $\{x : (x,s,y) \in \mathcal{Z}, y=1, s=1, f(x)=0\}$ and $\{x : (x,s,y) \in \mathcal{Z}, y = 1, s = -1, f(x) = 0\}$ are Lebesgue null sets,*

*(vi) the probability density functions $f_{\mathcal{Z}|y=1,s=1}$ and $f_{\mathcal{Z}|y=1,s=-1}$ are Lebesgue-measurable.*

*Then, the function $\lambda \mapsto \mathrm{DEO}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ is continuous.*

*Proof.* Analogous to the proof of Theorem 1. □

With these results, we can also prove the existence of a DEO-fair classifier similar to Corollary 1.

**Corollary 2** (Existence of a DEO-fair classifier). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \left\{f \in \mathcal{F} : \exists\lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that Theorem 2 holds and that there exist two hyperparameters $\lambda_+$ and $\lambda_-$ such that $\mathrm{DEO}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_+)\right) > 0$ and $\mathrm{DEO}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_-)\right) < 0$.*

*Then, there exists at least one value $\lambda_0 \in \left[\min\left(\lambda_+, \lambda_-\right), \max\left(\lambda_+, \lambda_-\right)\right]$ such that $\mathrm{DEO}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$.*

*Proof.* This corollary is a direct consequence of the intermediate value theorem and the continuity of DEO proven in Theorem 2. □

## 2.4 Towards Classifiers that are Fair and Accurate

In the last section, we presented a method that is guaranteed to find a DDP fair classifier. However, there is one important catch: we did not make any statement about the classification accuracy of this solution. Here, we take a step in this direction by proposing some sufficient conditions that ensure the existence of a classifier that is both fair and accurate. To this end, we focus on a particular set of classifiers: the family of similarity-based functions. Given a similarity function $K : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ and a set of points $S = \left\{(x'_1, s'_1, y'_1), \dots, (x'_d, s'_d, y'_d)\right\}$, we define a similarity based classifier as $f(x) = \sum_{i=1}^{d} \alpha_i K(x, x'_i)$. The goal is then to learn the weights $\alpha_i$.

A theory of learning with such functions has been developed by Balcan et al. (2008). By defining a notion of good similarities, they provide sufficient conditions that ensure the existence of an accurate similarity-based classifier. Here, we build upon this framework and we introduce a notion of good similarities for both accuracy and fairness. Hence, in Definition 2 we give sufficient conditions that ensure the existence of a classifier that is—within a guaranteed margin—fair and accurate at the same time.

**Definition 2** (Good Similarities for Fairness)**.** *A similarity function $K$ is $(\varepsilon, \gamma, \tau)$-good for convex, positive, and decreasing loss $\ell$ and $(\mu, \nu)$-fair for demographic parity if there exists a (random) indicator function $R(x, s, y)$ defining a (probabilistic) set of "reasonable points" such that, given that $\forall x \in \mathcal{X}, g(x) = \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y'K(x, x') \,|R(x', s', y')\right]$, the following conditions hold:*

*(i)* $$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}} \left[\ell\left(\frac{yg(x)}{\gamma}\right)\right] \leq \varepsilon,$$

*(ii)* $$\left| \mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{Z}|s=1}} [g(x) \geq \gamma] - \mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{Z}|s=-1}} [g(x) \geq \gamma] \right| \leq \mu,$$

*(iii)* $$\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}} [|g(x)| \geq \gamma] \geq 1 - \nu,$$

*(iv)* $$\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}} [R(x, s, y)] \geq \tau.$$

Roughly speaking, a similarity is good for classification if examples of the same class are on average closer to each other than examples of different classes up to a certain margin. Moreover, it is good for fairness if this margin is independent of group membership. Given such a similarity, we can prove the existence of a fair and accurate classifier as is summarized in the next theorem.

**Theorem 3** (Existence of a fair and accurate separator)**.** *Let $K \in [-1, 1]$ be a $(\varepsilon, \gamma, \tau)$-good and $(\mu, \nu)$-fair metric for a given convex, positive and decreasing loss $\ell$ with lipschitz* *constant $L$. For any $\varepsilon_1 > 0$ and $0 < \delta < \frac{\gamma\varepsilon_1}{2(L+\ell(0))}$, let $S = \left\{(x'_1, s'_1, y'_1), \dots, (x'_d, s'_d, y'_d)\right\}$ be a set of $d$ examples drawn from $\mathcal{D}_{\mathcal{Z}}$ with*

$$d \geq \frac{1}{\tau}\left[\frac{L^2}{\gamma^2\varepsilon_1^2} + \frac{3}{\delta} + \frac{4L}{\delta\gamma\varepsilon_1}\sqrt{\delta(1-\tau)\log(2/\delta)}\right].$$

*Let $\phi^S : \mathcal{X} \to \mathbb{R}^d$ be a mapping with $\phi^S_i(x) = K(x, x'_i)$, for all $i \in \{1, \dots, d\}$. Then, with probability at least $1 - \frac{5}{2}\delta$ over the choice of $S$, the induced distribution over $\phi^S(\mathcal{X}) \times \mathcal{S} \times \mathcal{Y}$ has a linear separator $\alpha$ such that*

$$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha, \phi^S(x)\right\rangle}{\gamma}\right)\right] \leq \varepsilon + \varepsilon_1,$$

*and, with $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}[s = 1]$,*

$$|DDP(\alpha)| \leq \mu + (\nu + 2\delta)\max\left(\frac{1}{p_1}, \frac{1}{1-p_1}\right).$$

*Proof.* Let $S = \left\{(x'_1, s'_1, y'_1), \dots, (x'_d, s'_d, y'_d)\right\}$ be a sample of size $d$ drawn from $\mathcal{D}_{\mathcal{Z}}$ and let $\phi^S : \mathcal{X} \to \mathbb{R}^d$ be a mapping defined as $\phi^S_i(x) = K(x, x'_i)$, for all $i \in \{1, \dots, d\}$. Recall that $|K(x, x)| \leq 1$ for all $x$. It implies that $\left\|\phi^S\right\|_\infty \leq 1$. Furthermore, let $\alpha \in \mathbb{R}^d$ be defined as $\alpha_i = \frac{y'_i R\left(x'_i, s'_i, y'_i\right)}{d_1}$ with $d_1 = \sum_i R(x'_i, s'_i, y'_i)$ which ensures that $\|\alpha\|_1 = 1$.

The proof is in two parts. First, we show the bound on the target criterion, that is, given $d$ chosen as in the theorem, we show that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^d_{\mathcal{Z}}}\left[\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha, \phi^S(x)\right\rangle}{\gamma}\right)\right] \leq \varepsilon + \varepsilon_1\right] \geq 1 - \delta.$$

Second, we show a bound on the true DDP, that is, given $d$ chosen as in the theorem, we show that

$$|\mathrm{DDP}(\alpha)| \leq \mu + \nu\max\left(\frac{1}{p_1}, \frac{1}{1-p_1}\right)$$

where $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}[s = 1]$.

**Bound on the target criterion.** For any example $(x, s, y) \sim \mathcal{D}_{\mathcal{Z}}$, we have

$$y\left\langle\alpha, \phi^S(x)\right\rangle = \frac{\sum_{i=1}^d yy'_i R(x'_i, s'_i, y'_i)\,K(x, x'_i)}{d_1}$$

which is an empirical average of $d_1$ terms with $R(x'_i, s'_i, y'_i) = 1$ and

$$-1 \leq yy'_i R(x'_i, s'_i, y'_i)\,K(x, x'_i) \leq 1.$$

Using Hoeffding's inequality, we can show that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-t\right]\leq \exp\left(-\frac{t^{2}d_{1}}{2}\right)$$

which implies that, with probability at least $1-\frac{\delta^2}{4}$, we have

$$y\left\langle \alpha,\phi^{S}(x)\right\rangle \geq \mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}.$$

This inequality holds for any $(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}$ and thus we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\leq\frac{\delta^{2}}{4}$$

$$\Rightarrow\mathbb{E}\left[\mathbb{P}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathbb{E}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\right]\leq\frac{\delta^{2}}{4}$$

$$\Rightarrow\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathbb{E}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\right]\leq\frac{\delta^{2}}{4}.$$

Then, applying Markov's inequality, we obtain that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathbb{E}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\geq\delta\right]\leq\frac{\delta}{4},$$

which implies

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[\mathbb{P}\left[y\left\langle \alpha,\phi^{S}(x)\right\rangle \leq \mathbb{E}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\leq\delta\right]\geq 1-\frac{\delta}{4}.$$

In other words, with a probability at least $1-\frac{\delta}{4}$ at most $\delta$ fraction of points violate

$$y\left\langle \alpha,\phi^{S}(x)\right\rangle \geq \mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}. \tag{2.9}$$

Therefore, let $g(x) = \mathbb{E}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[y'K(x,x')\,|R(x',s',y')\right]$, with a probability at least $1-\frac{\delta}{4}$ for at least $1-\delta$ fraction of points, which do not violate (2.9), we have, for our decreasing loss $\ell$ (for example the hinge loss, $\ell(w) = \max(0, 1-w)$):

$$\begin{aligned}\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right) &\leq \ell\left(\frac{yg(x)-\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}}{\gamma}\right)\\ &\leq \ell\left(\frac{yg(x)}{\gamma}\right)+L\left|\frac{1}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\right|\\ &\leq \ell\left(\frac{yg(x)}{\gamma}\right)+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}.\end{aligned}$$

For at most a $\delta$ fraction of points violating (2.9), we use a bound on the worst case loss derived from its lipschitzness.

$$\begin{aligned}\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right) &\leq L\left|\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right|+\ell(0)\\ &\leq L\max_x\frac{\left|\left\langle\alpha,\phi^S(x)\right\rangle\right|}{\gamma}+\ell(0) && \text{(Cauchy-Schwarz Inequality.)}\\ &\leq L\max_x\frac{\|\alpha\|_1\left\|\phi^S(x)\right\|_\infty}{\gamma}+\ell(0)\\ &\leq \ell(0)+\frac{L}{\gamma} && (\gamma\leq 1.)\\ &\leq \frac{L+\ell(0)}{\gamma}.\end{aligned}$$

Altogether, we obtain with a probability of at least $1-\frac{\delta}{4}$ over $S$ that

$$\begin{aligned}&\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right)\right]\\ &\leq \mathbb{E}\left[\frac{L+\ell(0)}{\gamma}\mathbb{1}(x\textit{ violates }(2.9))+\left(\ell\left(\frac{yg(x)}{\gamma}\right)+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\right)\mathbb{1}(x\textit{ satisfies }(2.9))\right]\\ &\leq \frac{(L+\ell(0))\,\delta}{\gamma}+\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{yg(x)}{\gamma}\right)\right]+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}} && \text{(def. of good similarity.)}\\ &\leq \frac{(L+\ell(0))\,\delta}{\gamma}+\varepsilon+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}} && \left(\delta<\tfrac{\gamma\varepsilon_1}{2(L+\ell(0))}.\right)\\ &\leq \frac{\varepsilon_1}{2}+\varepsilon+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}.\end{aligned}$$

Furthermore, the number $d_1$ of reasonable landmarks follows a binomial distribution $B(d,p)$ with $p \geq \tau$. With our choice of $d$, we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^d}\left[\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\leq\frac{\varepsilon_1}{2}\right]\geq 1-\frac{\delta}{4}.$$

Using the union bound, we obtain with a probability of at least $1-\frac{\delta}{2}$ over $S$ that

$$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right)\right]\leq\varepsilon+\varepsilon_1.$$

**Bound on the fairness criterion** For any example $(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}$, we have

$$\left\langle\alpha,\phi^S(x)\right\rangle=\frac{\sum_{i=1}^d y_i'R(x_i',s_i',y_i')\,K(x,x_i')}{d_1},$$

which is an empirical average of $d_1$ terms with $R(x_i',s_i',y_i')=1$ and

$$-1\leq y_i'R(x_i',s_i',y_i')\,K(x,x_i')\leq 1.$$

Let $g(x)=\mathbb{E}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[y'K(x,x')\,|R(x',s',y')\right]$. Using the same kind of argument than in the first part of the proof, that is applying Hoeffding's inequality followed by Markov's inequality, we can show that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^d}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle-g(x)\right|\geq\sqrt{\frac{2\log\left(\frac{8}{\delta^2}\right)}{d_1}}\right]\leq\delta\right]\geq 1-\frac{\delta}{4}.$$

Furthermore, notice that the number $d_1$ of reasonable landmarks follows a binomial distribution $B(d,p)$ with $p\geq\tau$. With our choice of $d$, with probability at least $1-\frac{\delta}{4}$ over the choice of $S$, it implies that

$$\sqrt{\frac{2\log\left(\frac{8}{\delta^2}\right)}{d_1}}\leq\gamma.$$

As a consequence, we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^d}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle-g(x)\right|\geq\gamma\right]\leq\delta\right]\geq 1-\frac{\delta}{2}. \tag{2.10}$$

To derive a bound on $|\mathrm{DDP}(\alpha)|$, we first derive bounds on $\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle| s=1\right]$ and $\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle| s=-1\right]$. Notice that:

$$\begin{aligned}
&\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle| s=1\right]\\
&\geq \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq\gamma \cap \left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| \leq \gamma \middle| s=1\right]\\
&\geq 1-\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)<\gamma \cup \left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma \middle| s=1\right]\\
&\hfill \text{(Union's bound.)}\\
&\geq 1-\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)<\gamma | s=1\right]-\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma \middle| s=1\right]\\
&\hfill \left(\mathbb{P}\left[A|B\right]\leq\tfrac{\mathbb{P}[A]}{\mathbb{P}[B]}\right)\\
&\geq \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq\gamma | s=1\right]-\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma \middle| s=1\right]\\
&\geq \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq\gamma | s=1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1},
\end{aligned}$$

where $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[s=1\right]$. With a symmetric argument, we have that

$$\begin{aligned}
&\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[\left\langle\alpha,\phi^S(x)\right\rangle < 0\middle| s=1\right] \geq\\
&\quad\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\leq-\gamma | s=1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1}.
\end{aligned}$$

With $\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left\langle\alpha,\phi^S(x)\right\rangle < 0\middle| s=1\right] = 1-\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle| s=1\right]$ and $1-\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[g(x)\leq-\gamma | s=1\right] = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[g(x)\geq-\gamma | s=1\right]$, we have

$$\begin{aligned}
&\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle| s=1\right]\\
&\quad\leq \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq-\gamma | s=1\right]+\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1}.
\end{aligned}$$

Furthermore, we have that

$$\begin{aligned}
\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq-\gamma | s=1\right] &\leq \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[-\gamma\leq g(x)\leq\gamma \cup g(x)\geq\gamma | s=1\right]\\
&\hfill\text{(Using the union bound and by definition of a good similarity.)}\\
&\leq \frac{\nu}{p_1}+\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}\left[g(x)\geq\gamma | s=1\right].
\end{aligned}$$

This implies that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle \alpha, \phi^S(x)\right\rangle \geq 0 \middle| s=1\right] \leq \frac{\nu}{p_1} + \\ \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma \middle| s=1\right] + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1}.$$

In a similar fashion, we have that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle \alpha, \phi^S(x)\right\rangle \geq 0 \middle| s=-1\right] \geq \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma \middle| s=-1\right] \\ - \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}$$

and

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle \alpha, \phi^S(x)\right\rangle \geq 0 \middle| s=-1\right] \leq \frac{\nu}{1-p_1} + \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma \middle| s=-1\right] \\ + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}.$$

These inequalities imply an upper bound on $\mathrm{DDP}(\alpha)$,

$$\begin{aligned}\mathrm{DDP}(\alpha) &= \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle \alpha, \phi^S(x)\right\rangle \geq 0 \middle| s=1\right] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle \alpha, \phi^S(x)\right\rangle \geq 0 \middle| s=-1\right] \\ &\leq \frac{\nu}{p_1} + \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma \middle| s=1\right] + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} \\ &\quad - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma \middle| s=-1\right] + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1} \\ &\qquad \text{(By definition of a good similarity.)} \\ &\leq \frac{\nu}{p_1} + \mu + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} \\ &\quad + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}\end{aligned}$$

and, similarly these inequalities imply a lower bound on $\mathrm{DDP}(\alpha)$,

$$\mathrm{DDP}(\alpha) \geq -\frac{\nu}{1-p_1} - \mu - \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} \\ - \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle \alpha, \phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}.$$

Then, using Inequality 2.10 and the union bound, we obtain that

$$\underset{S \sim \mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}} \left[ \mathrm{DDP}(\alpha) \leq \frac{\nu}{p_1} + \mu + \frac{\delta}{p_1} + \frac{\delta}{1-p_1} \right] \geq 1 - \delta$$

In a similar fashion, we also obtain that

$$\underset{S \sim \mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}} \left[ \mathrm{DDP}(\alpha) \geq -\frac{\nu}{1-p_1} - \mu - \frac{\delta}{p_1} - \frac{\delta}{1-p_1} \right] \geq 1 - \delta$$

We can combine both inequalities with the union bound to obtain

$$\underset{S \sim \mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}} \left[ |\mathrm{DDP}(\alpha)| \leq \mu + (\nu + 2\delta) \max\left( \frac{1}{p_1}, \frac{1}{1-p_1} \right) \right] \geq 1 - 2\delta$$

Using the union one last time to combine the fairness bound and the target criterion bound gives the theorem. □

## 2.5 Experiments

In this section, we empirically evaluate SearchFair by comparing it to 5 baselines on 6 real-world datasets. In all the experiments, SearchFair either reliably finds the fairest classifier and is comparable to a very recent non-convex optimization approach.

**Datasets.** We consider 6 different datasets: CelebA (Liu et al., 2015), Adult (Kohavi and Becker, 1996), Dutch (Zliobaite et al., 2011), COMPAS (Larson et al., 2016), Communities and Crime (Redmond and Baveja, 2002), and German Credit (Dua and Graff, 2017). We present results for CelebA, Adult, and COMPAS in this chapter. In Appendix A, we give detailed descriptions of the other datasets, how we pre-process the data, and the sizes of the train and test splits. Note that we remove the protected attribute $s$ from the set of features $x$ so that it is not needed at decision time.

We pre-process the datasets by normalizing and centering continuous variables. For categorical values, we use a one-hot encoding. We select a fixed number of randomly selected points for training, and use the rest of the points for testing.

**CelebA.** The CelebA dataset (Liu et al., 2015) contains $202,599$ images of celebrity faces from the web. In addition to the image data, there exist 40 binary attribute labels describing the content of the images, such as ‘Black Hair’, ‘Bald’, and ‘Eyeglasses’. We use 38 of those descriptions as features, the sex as the protected attribute, and the attribute ‘Smiling’ as the class label. We use $10,000$ randomly selected points for training.

**Adult.** The Adult dataset (Kohavi and Becker, 1996) contains data from the U.S. 1994 Census database. There are $48,842$ instances with 14 features, among others age and education, including the two protected attributes sex and race. We apply the preprocessing of Wu et al. (2019): we consider sex with values male and female as the protected attribute and use 9 features for training, dropping FNLWGT, EDUCATION, CAPITAL-GAIN, CAPITAL-LOSS. The goal is to predict the income: $y = 1$ if it is more than fifty thousand U.S. Dollars, $y = -1$ otherwise. We use $10,000$ randomly selected points for training.

**Dutch.** The Dutch dataset (Zliobaite et al., 2011) contains data from the 2001 Netherlands Census and consists of $60,420$ data points which are characterized by 12 features. We use gender as the protected attribute and predict *low income* or *high income* as it is determined by occupation. Hence, we learn with the remaining 10 features. We use $10,000$ randomly selected points for training.

**Baselines.** We compare SearchFair to 5 baselines. For 3 of them, we use Optimization Problem 2.4 with hinge loss and a squared $\ell_2$ norm as the regularization term. As a function class $\mathcal{F}$, we use similarity-based classifiers presented in Section 2.4 with either the linear or the rbf kernel and with 70% (at most 1000) of the training examples as reasonable points. As a fairness constraint, we use either the linear relaxation of Zafar et al. (2017a) (Zafar), the linear relaxation of Donini et al. (2018) (Donini), or no constraint at all (Unconst). The fourth baseline is a recent method for non-convex constrained optimization by Cotter et al. (2019) (Cotter). Our last baseline is the constant classifier (Constant) that always predicts the same label but has perfect fairness.

**SearchFair.** [3] For SearchFair we also use the hinge loss, a squared $\ell_2$ norm as the regularization term (it is strongly convex), and similarity-based classifiers. As a convex approximation of the fairness constraint, we use the bounds with hinge loss proposed by Wu et al. (2019) (see Section 2.3for details).

**Metrics.** Our main goal is to learn fair classifiers. Hence, our main evaluation metrics are the empirical DDP and DEO scores on the test set (lower is better). As a secondary metric (in case of ties in the fairness scores), we consider the classification performance of the models and we report the errors on the test set (lower is better). All the experiments are repeated 10 times and we report the mean and standard deviation for all the metrics.

**Hyperparameters.** Zafar, Donini and Cotter use a fairness parameter, that we call $\tau$, to control the fairness level. Since our goal is to learn classifiers that are fair, we set $\tau = 0$ such that perfect fairness is required. For SearchFair, there is no fairness parameter since $\lambda_0$ is automatically searched for between a lower bound $\lambda_{\text{min}}$ and an

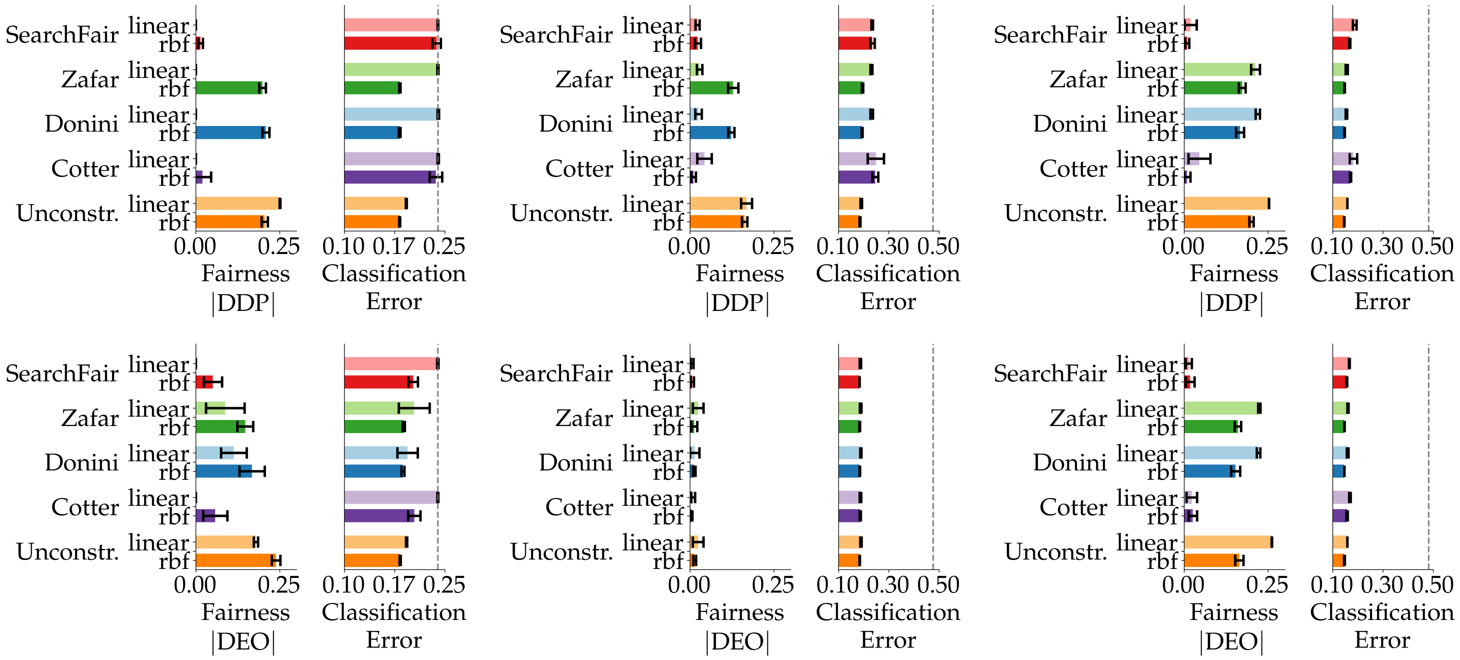
[3] The code is freely available online: `github.com/mlohaus/SearchFair`.

upper bound $\lambda_{\max}$. We set $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$ as these values usually lead to classifiers with fairness scores of opposite sign (as needed). We use 10 iterations in the binary search.

We use 5-fold cross validation to choose other hyperparameters. For Cotter, only the width of the rbf kernel has to be tuned since we use the framework of the original paper with no regularization term. For all remaining methods we need to choose the regularization parameter $\beta$ and the width of the rbf kernel. These values are respectively chosen in the sets $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and $\{10^{-\lceil \log d \rceil -1}, 10^{-\lceil \log d \rceil}, d^{-1}, 10^{-\lceil \log d \rceil +1}, 10^{-\lceil \log d \rceil +2}\}$, with $d$ the number of features. We select the set of parameters that lead to the most accurate classifier on average over the 5 folds. Indeed, the fairness level is automatically taken care of by the methods.

**Results.** We present the results for 3 out of 6 datasets in Figure 2.5. The other results are deferred to Appendix A as they follow the same trend. We make two main observations. First, SearchFair always obtains fairness values that are very close to zero. It learns the fairest classifiers out of all the methods and is only matched by Cotter, the non-convex approach. This sometimes comes with a small increase in terms of classification error. For example, in order to achieve perfect DDP fairness on the Adult dataset, SearchFair, and all the other fair methods, yield classifiers close to the trivial constant one. Second, the complexity of the model greatly influences the performances of the linear relaxations. For example, using the complex rbf kernel almost always results in an increase in the fairness score of Zafar and Donini. This is particularly striking for Adult and Dutch where the linear kernel yields reasonable fairness scores. Note that this trend is not always respected. For example, on CelebA, using an rbf kernel improves the fairness score compared to the linear kernel. However, neither of them obtain reasonable fairness levels in the first place.

**Discussion on hyperparameter selection.** Apart from the hyperparameter selection method used in our experiments, one can think of other cross validation procedures. For example, Donini et al. (2018) proposed NVP, a cross validation method where one selects the set of hyperparameters that gives the fairest classifier while obtaining an average accuracy above a given threshold. Similarly, one could select the set of hyperparameters that yields the most accurate classifier under a given fairness threshold. In Appendix A, we report results that empirically show that these more complex procedures tend to improve the fairness of the baselines (SearchFair remains competitive on all the datasets). Unfortunately, they also blur the dividing line between hyperparameters that control the fairness of the model and the ones that control its complexity. In other words, it becomes unclear whether fairness is achieved thanks to the relaxation or thanks to the choice of hyperparameters (we already evoked this issue in Figure 2.3). We believe that it is better to have a method that is guaranteed to find a fair classifier for any given family of models and does not rely on a complex cross validation procedure.



(a) Adult.

(b) Dutch.

(c) CelebA.

Figure 2.5: We report the average and standard deviation of classification error and absolute fairness scores DDP and DEO (closer to 0 is better) over 10 repetitions. The constant classifier is perfectly fair as it always predicts the same label. Its classification error is shown by the grey dashed vertical line. (a) To obtain good fairness on Adult, all DDP fair methods learn the constant classifier. Only SearchFair and Cotter reliably find fair classifiers for both kernels. (b) On Dutch, SearchFair obtains the lowest DDP with a slight loss in accuracy. Cotter performs comparably for both kernels, whereas the other methods only do well with a linear kernel but fail to learn fair classifiers with the rbf kernel. (c) For CelebA, SearchFair and Cotter are the only methods that obtain a low DDP and DEO with only a slight loss in accuracy. The other methods only provide little to no improvement.

## 2.6 Conclusion

In this chapter, we have shown that existing approaches to learn fair and accurate classifiers have many shortcomings. They use loose relaxations of the fairness constraint and guarantees that relate the relaxed fairness to the true fairness of the solutions are either missing or not sufficient. We empirically demonstrated how these approaches can produce undesirable models. If “fair machine learning” is supposed to be employed in real applications in society, we need algorithms that actually find fair solutions, and ideally come with guarantees. We made a first step in this direction by proposing SearchFair, an approach that uses convex relaxations to learn a classifier that is guaranteed to be fair.

# Chapter 3

# Disparate Treatment in Neural Networks

Autonomous systems that make substantive decisions about people must often conform to relevant anti-discrimination legislation. Within the US legal system, two of the most common tests of anti-discrimination legislation are referred to as disparate treatment and disparate impact (Chapter 1, King and Hemenway (2020)). Importantly, it has been argued (Barocas and Selbst, 2016) that there are a large range of scenarios where disparate treatment is unlawful even when performed as a remedy to disparate impact. This is in marked departure from the EU and the UK where considerably more latitude exists when rectifying indirect discrimination (analogous to disparate impact) (Wachter et al., 2021).

Consequentially, disparate treatment doctrine prevents a wide range of actions intended to address sustained inequality (Bent, 2019). Of particular relevance to our work is a 1991 amendment to Title VII (Statute, 1991). This amendment explicitly prohibits the "use [of] different cutoff scores for . . . employment related tests on the basis of race, color, religion, sex, or national origin", even if done for reasons of affirmative action. In this chapter, we examine the relationship between existing methods for enforcing demographic parity, and methods for demographic parity that alter the cutoff score on the basis of inferred race or gender, raising fundamental questions about the legality of existing approaches for enforcing demographic parity.

In particular, we examine the behavior of deep neural networks trained to enforce demographic parity, either by using a regularizer (i.e., an additional loss term) or by preprocessing (Kamiran and Calders, 2012). There are two plausible routes for how these networks might exhibit demographic parity. Such models could either *(i)* learn an internal representation that is unpredictive with respect to the protected attribute, or *(ii)* they could learn to distinguish between groups and tune a separate classifier for each group in a way that happens to result in a demographically fair outcome.

In the context of US law, it is vital to understand which of these cases occurs in practice. If the learned algorithms treat people differently on the basis of their race or gender, this may correspond to *disparate treatment*. We find that networks trained

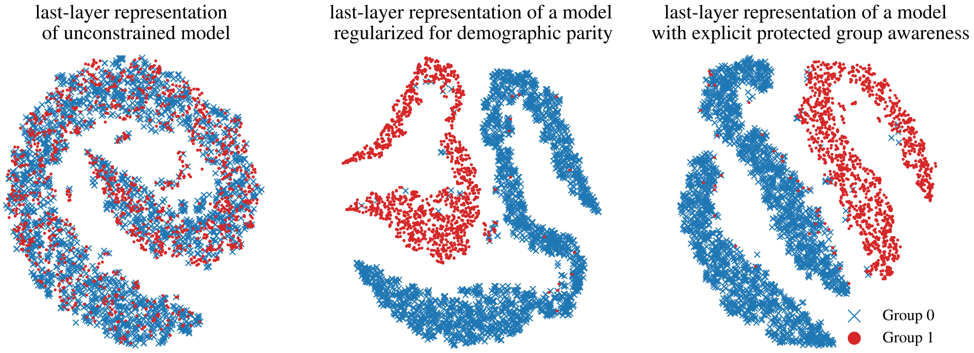


Figure 3.1: **Feature representations of unconstrained (left), fairness-regularized (center), and group-aware (right) ResNet50 models.** The plots show tSNE (van der Maaten and Hinton, 2008) embeddings of the last-layer representations of CelebA celebrity images. Each plot shows a classifier trained to identify if people are smiling, with red points corresponding to individuals labeled male, and blue female. The additional use of fairness constraints at training causes a mixed manifold of men and women (left) to separate into largely disjoint sets (center). Similar behavior is observed when training a two-headed model (right) to predict both gender and 'smiling'.

to satisfy demographic parity fall into this second case. Moreover, we show that the more strongly demographic parity is enforced, the more predictive the internal representation is of the protected attribute; see Figure 3.1.

Building on the observation that the protected attribute is implicitly learned, we train a neural network with a second classification head that explicitly predicts the protected attribute, and use the second head response to directly reduce demographic disparity. Compared to regularized approaches, our training strategy provides better interpretability and improved stability when demographic parity is strongly enforced.

Formally, we consider an ex-post method of enforcing demographic parity over a standard classifier $f$ trained without consideration of fairness. Our method makes a positive decision if the inequality

$$f(x) \geq a_1 g(x) + a_2 \tag{3.1}$$

is observed, where $x$ is a datapoint, such as a photograph for a curriculum vitae, $g$ is a predictor of the protected attribute, and $a_1, a_2 \in \mathbb{R}$ are tunable parameters that control the accuracy-fairness trade-off. Note that this classifier is constructed to explicitly treat people differently based on their perceived protected attribute (as represented by the response $g(x)$), and hence, demonstrates disparate treatment (cf. Section 3.3 and Figure 3.2).

We compare this classifier $f$ to a fair classifier $r$, which is trained with a fairness regularizer or preprocessing to approximately satisfy demographic parity, and show that $f$ and $r$ are tightly related. We conclude that $r$ also demonstrates disparate treatment. In summary, we make the following contributions in this chapter.

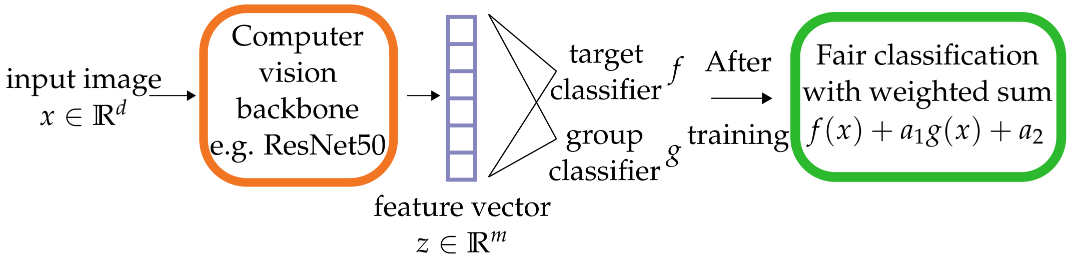


Figure 3.2: **Our simple two-headed model for demographic parity**, which exhibits explicit disparate treatment. We add two heads to a pretrained backbone: the target classifier $f$, and the group classifier $g$, which predicts the protected attribute. We jointly train both heads and the backbone. After training, we estimate coefficients $a_1$ and $a_2$ such that the weighted sum $f(x) + a_1 g(x) + a_2$ is accurate and fair.

1. The internal state of $r$ is strongly predictive of the protected attribute, and as the emphasis on fairness increases, so does the predictive accuracy for the protected attribute (Section 3.2).

2. The formulation of eq. (3.1) is an effective approach for creating demographic–parity–fair classifiers. Namely, we show that on computer vision datasets, the performance of eq. (3.1) closely tracks the optimal accuracy-fairness trade-offs of Lipton et al. (2018) without requiring explicit access to the protected attribute (Section 3.3).

3. Decisions of the fair model $r$, which makes a positive decision if $r(x) \geq 0$, are well-approximated by $f(x) + a_1 g(x) + a_2$ and demonstrate the same accuracy-fairness trade-off. Similarly, decisions made by an unconstrained classifier are closely approximated by $r(x) - b_1 g(x) - b_2$ (Section 3.4.1).[1]

4. Using this close relationship between fair model and the unconstrained model in eq. (3.1) we are able to identify individuals who have been systematically disadvantaged by decisions made by the fair classifier, and who would have received a positive decision, if their apparent race or gender were different. This allows us to conclude that the demographic–parity–fair method also exhibits disparate treatment (Section 3.4.2). Legal implications are discussed in Section 3.5.

## 3.1 Preliminaries

### 3.1.1 Background on Fair Classification

We define a binary classifier as a function $h : \mathcal{X} \rightarrow \{0,1\}$ which, given a datapoint $x$ from a feature space $\mathcal{X}$, aims to accurately predict the datapoint's ground-truth label $y \in \mathcal{Y} = \{0,1\}$. We consider thresholded classifiers of the form $h(x) =$

[1] As we are interested in mimicking discrete yes/no *decisions* rather than continuous *classifier responses*, the two statements are not equivalent.

$\mathbb{1}(f(x) > 0)$ where $f$ is a continuous function $f : \mathcal{X} \to \mathbb{R}$. We require that $h$ be as accurate as possible, while remaining fair with respect to a protected attribute $s \in \mathcal{S} = \{0, 1\}$[2]. We assume a joint distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ from which we can draw $(X, Y, S) \sim \mathbb{P}$. We focus on the well-known notion of demographic parity (aka statistical parity; (Dwork et al., 2012; Feldman et al., 2015; Kamiran and Calders, 2012)) defined as:

**Definition 3** (Demographic Parity)**.** *A binary classifier $h : \mathcal{X} \to \{0, 1\}$ exhibits* demographic parity *under a distribution $\mathbb{P}$ if its prediction $h(X)$ is independent of the protected attribute $S$, that is*

$$\mathbb{P}(h(X) = 1|S = 0) = \mathbb{P}(h(X) = 1|S = 1).$$

*We measure the violation of demographic parity by the* demographic disparity *(DDP), given by*

$$\text{DDP} = \text{DDP}(h; \mathbb{P}) := \mathbb{P}(h(X) = 1|S = 1) - \mathbb{P}(h(X) = 1|S = 0). \tag{3.2}$$

**Regularized Fair Classification.** One common approach for learning a fair classifier is to add a regularizer to the standard training objective (Bendekgey and Sudderth, 2021; Beutel et al., 2019; Kleindessner et al., 2021; Lohaus et al., 2020; Manisha and Gujar, 2020; Risser et al., 2021; Wick et al., 2019, e.g.,). Such regularizers are used to enforce numerous fairness definitions and typically impose a continuous relaxation of a discrete fairness measure such as the DDP (3.2). The regularizer used by (Wick et al., 2019) is a sigmoid-based relaxation of the squared value of (3.2), evaluated on the training dataset $\{(x_i, s_i, y_i)\}_{i=1}^n$, which is an i.i.d. sample from $\mathbb{P}$:

$$\widehat{\mathcal{R}}_{\text{DP}}(f) := \left( \frac{1}{|\{i : s_i = 1\}|} \sum_{i \in [n]: s_i = 1} \sigma(f(x_i)) - \frac{1}{|\{i : s_i = 0\}|} \sum_{i \in [n]: s_i = 0} \sigma(f(x_i)) \right)^2. \tag{3.3}$$

The fairness regularizer trades-off fairness against the accuracy of the classifier via a hyperparameter $\lambda$. In the main body of the paper, we only report results for the regularizer $\widehat{\mathcal{R}}_{\text{DP}}(f)$.

We also consider a similar regularizer (Manisha and Gujar, 2020), denoted by $\widehat{\mathcal{R}}_{\text{DP}}^{\text{abs}}$, where the squaring function is replaced by the absolute value, that is

$$\widehat{\mathcal{R}}_{\text{DP}}^{\text{abs}}(f) := \left| \frac{1}{|\{s : s_i = 1\}|} \sum_{i \in [n]: s_i = 1} \sigma(f(x_i)) - \frac{1}{|\{s : s_i = 0\}|} \sum_{i \in [n]: s_i = 0} \sigma(f(x_i)) \right|. \tag{3.4}$$

We follow (Bendekgey and Sudderth, 2021; Wick et al., 2019) in applying the regularizer (3.3) to the sigmoid output of the networks. This differs from approaches such

[2] In this chapter, we stick to the annotations available in the CelebA or FairFace dataset (compared to the other chapters). These are externally assigned binary labels, but in principle, our method in Section 3.3 works for multiple / non-binary protected groups.

as (Donini et al., 2018; Zafar et al., 2017a), which enforce fairness constraints on logits representing real-valued margin distances and have recently been criticized for being easy to satisfy without requiring the classifier to make fair decisions (Lohaus et al., 2020).

**Preprocessing: Massaging the Dataset.** We also examine a preprocessing method (Kamiran and Calders, 2012), which alters target labels prior to training. *Massaging* 'promotes' negative points from the disadvantaged group to the positive label if they have a high positive class probability as calculated by a unconstrained classifier and 'demotes' positive points from the advantaged group to the negative label if they have a low positive class probability. The number of points whose label is flipped is controlled by a parameter $\lambda \in [0, M]$, where $\lambda = M$ results in the same fraction of positive points for both groups.

### 3.1.2 Implicit Disparate Treatment

Lipton et al. (2018) examined the popular claim that machine learning models that do not use protected information at test time cannot exhibit disparate treatment (Donini et al., 2018; Goh et al., 2016; Harned and Wallach, 2019; Manisha and Gujar, 2020; Wu et al., 2019; Zafar et al., 2017a,b). Lipton et al. (2018) recommended caution and observe that if the protected attribute $s$ is a deterministic function $s = g(x)$ of the non-protected features $x$, any sufficiently powerful ML model can learn a function $f(x,s) = \tilde{f}(x)$ with $\tilde{f}(x) = f(x, g(x))$. They argue that even though the protected attribute is not provided at test time, such a model would constitute a case of disparate treatment since it makes decisions based on the *implicitly reconstructed* protected attribute. However, beyond a synthetic experiment in which a classifier discriminates based on hair length (as a proxy of gender), they do not study whether–and how–implicit disparate treatment happens in practice. We complement their work by providing strong evidence that deep neural networks suffer from disparate treatment even when not explicitly using the protected attribute at test time.

Lipton et al. (2018) proposed a postprocessing method that requires protected attributes at test time to apply per-group decision thresholds on the classifier response. These thresholds are greedily chosen. Lipton et al. (2018) proved that their approach finds optimal thresholds under a given fairness constraint. We compare to them in Section 3.3.

### 3.1.3 Computer Vision and Anti-discrimination Law

We evaluate on standard computer vision datasets. There are two reasons for this: first, large organizations are increasingly aware of the ramifications of releasing data and reluctant to release data that might reveal gender or racial biases. Computer vision remains one of the few areas of modern machine learning where large high-dimensional datasets continue to be released alongside race- or gender-based annotations. As such, the use of these datasets is perhaps more representative of fair machine learning as it can be practiced within industry, rather than the use of low-dimensional historic datasets that are common to the field. The second reason concerns the recent shift in the way computer vision is used. US anti-discrimination law is primarily concerned with decisions made in the contexts of employment, education, and housing. However, in the past year, we have seen the rise of online proctoring systems that attempt to detect cheating by using computer vision systems to determine where a person is looking. Automatically identifying an individual as cheating can have long-term impact on continued access to education. While the systems are proprietary and little information about them is publicly available, some are known to exhibit racial biases (Feathers, 2021). AI systems are also being used to determine if workers are smiling sufficiently (Vincent, 2021), and it is likely a matter of time until similar technology is used to evaluate the job performance of customer-facing workers. As such, it is vital that we understand how such systems interact with an individual's rights and with anti-discrimination legislation.

### 3.1.4 Datasets and Technical Details

**CelebA dataset.** The CelebA dataset (Liu et al., 2015) contains $202,599$ images of celebrity faces with 40 binary annotations, such as WEARING_GLASSES, SMILING or MALE. We use the Aligned&Cropped subset and its standard split into train, test, and validation data. We center-crop the images and resize them to $224 \times 224$. During training, we randomly crop and flip images horizontally. We use (Ramaswamy et al., 2021) as reference for choosing target and protected labels.

**FairFace dataset.** The FairFace dataset (Karkkainen and Joo, 2021) contains $108,501$ images collected from the YFCC-100M Flickr dataset and are annotated with GENDER, RACE, and AGE. We binarize the attribute RACE into WHITE and the union of all other groups. From AGE, we build several binary attributes: BELOW_20, BELOW_30, and BELOW_40. In our experiments, we use the provided validation data with 1.25 padding as our test data, and from the provided train data, we prepared our own random and balanced validation split. We center-crop the images and resize them to $224 \times 224$. During training, we crop randomly, and randomly flip the images horizontally (p=0.5).

**Models.** Given a fixed target attribute and protected attribute, we train all parameters of a pretrained ResNet50 (He et al., 2016) or MobileNetV3-Small (Howard et al., 2019) backbone provided by PyTorch with the binary cross entropy loss. MobileNetV3-Small contains 2.8M parameters and is more resource friendly than the much bigger ResNet50. Hence, for some experiments we only used MobileNetV3-Small to save computation time. The dimension $m$ of the last-layer representation $z \in \mathbb{R}^m$ is $m = 2048$ for the ResNet50 and $m = 1024$ for MobileNetV3-Small.

We train all models, including our two-headed approach, with the Adam Optimizer (Kingma and Ba, 2017) (learning rate is $10^{-4}$ on CelebA and $10^{-5}$ on FairFace, batchsize is 64) for a total of 20 epochs and select the model with the highest average precision achieved on the validation set. In addition, we employ a learning rate scheduler that reduces the learning rate by a factor of 10 if there is no progress on the validation loss for more than 8 epochs. To have meaningful regularizer losses for each batch, we use stratified batches, such that the prevalence of the protected attribute is roughly the same as the overall prevalence. For the classification loss, we use binary cross entropy loss with a sigmoid activation.

If we train the models with one of our two fairness regularizers, the range for the fairness parameter $\lambda$ is $[0, 0.1, 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 30]$. For the *Massaging* preprocessing method, the range is $[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$.

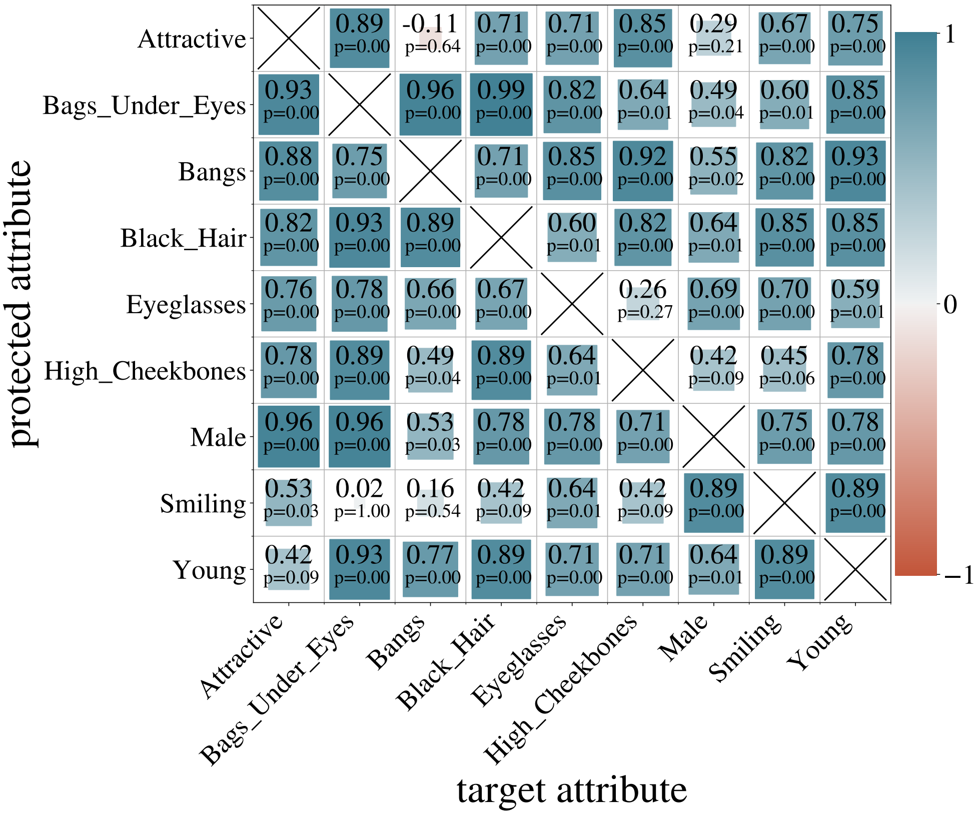
## 3.2 Protected Attribute Awareness in Fair Networks

We develop tools and statistical tests that allow us to understand if deep networks enforce demographic parity by either *(i)* learning an internal representation that is unpredictive with respect to the protected attribute, or *(ii)* by learning a representation that separates groups, thus allowing each group to be treated differently in a way that results in a fair outcome. For both regularized and preprocessing approaches, awareness of the protected attribute increases as fairness is more strongly enforced. To measure how well a network separates the protected groups, we examine how accurately a linear classifier can recover the protected attribute from the responses of the last layer of a neural network trained to enforce demographic parity. As we increase the fairness parameter $\lambda$, the accuracy of predicting the protected attribute from the last layer is approaching the performance of a model that is explicitly trained to do so. In this section, we present results for the fairness regularizer defined in (3.3). Results for the other regularizer and for the preprocessing method *Massaging* (Kamiran and Calders, 2012) are presented in Appendix B.1.

**Experimental Setup** We choose a target and a protected attribute from 9 binary attributes of the CelebA dataset. For each distinct pair of target and protected attribute, we train 12 models — corresponding to fairness parameter $\lambda \in \{0, 0.1, 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 30\}$. For every model, we train a linear classifier using logistic regression to predict the protected attribute from the model's frozen last-layer representation; we refer to such a classifier as protected attribute classifier or group classifier.

**Evaluation** For each pair of target and protected attribute, we evaluate if increasing $\lambda$ increases the accuracy of a linear classifier trained on the last layer. We test for a monotonic relationship using the Kendall-tau correlation $\tau$ (Kendall, 1945).

We perform this test on 12 datapoints consisting of two values: the fairness parameter $\lambda$ and the accuracy of the protected attribute classifier that predicts the protected attribute from the last-layer representation. Additionally, we compute a two-sided p-value for the null hypothesis of independence between $\lambda$ and the accuracy. Since the regularized approach can collapse to a trivial near-constant classifier when $\lambda$ is too



| | Attractive | Bags_Under_Eyes | Bangs | Black_Hair | Eyeglasses | High_Cheekbones | Male | Smiling | Young |
|---|---|---|---|---|---|---|---|---|---|
| Attractive | × | 0.89<br>p=0.00 | -0.11<br>p=0.64 | 0.71<br>p=0.00 | 0.71<br>p=0.00 | 0.85<br>p=0.00 | 0.29<br>p=0.21 | 0.67<br>p=0.00 | 0.75<br>p=0.00 |
| Bags_Under_Eyes | 0.93<br>p=0.00 | × | 0.96<br>p=0.00 | 0.99<br>p=0.00 | 0.82<br>p=0.00 | 0.64<br>p=0.01 | 0.49<br>p=0.04 | 0.60<br>p=0.01 | 0.85<br>p=0.00 |
| Bangs | 0.88<br>p=0.00 | 0.75<br>p=0.00 | × | 0.71<br>p=0.00 | 0.85<br>p=0.00 | 0.92<br>p=0.00 | 0.55<br>p=0.02 | 0.82<br>p=0.00 | 0.93<br>p=0.00 |
| Black_Hair | 0.82<br>p=0.00 | 0.93<br>p=0.00 | 0.89<br>p=0.00 | × | 0.60<br>p=0.01 | 0.82<br>p=0.00 | 0.64<br>p=0.01 | 0.85<br>p=0.00 | 0.85<br>p=0.00 |
| Eyeglasses | 0.76<br>p=0.00 | 0.78<br>p=0.00 | 0.66<br>p=0.00 | 0.67<br>p=0.00 | × | 0.26<br>p=0.27 | 0.69<br>p=0.00 | 0.70<br>p=0.00 | 0.59<br>p=0.01 |
| High_Cheekbones | 0.78<br>p=0.00 | 0.89<br>p=0.00 | 0.49<br>p=0.04 | 0.89<br>p=0.00 | 0.64<br>p=0.01 | × | 0.42<br>p=0.09 | 0.45<br>p=0.06 | 0.78<br>p=0.00 |
| Male | 0.96<br>p=0.00 | 0.96<br>p=0.00 | 0.53<br>p=0.03 | 0.78<br>p=0.00 | 0.78<br>p=0.00 | 0.71<br>p=0.00 | × | 0.75<br>p=0.00 | 0.78<br>p=0.00 |
| Smiling | 0.53<br>p=0.03 | 0.02<br>p=1.00 | 0.16<br>p=0.54 | 0.42<br>p=0.09 | 0.64<br>p=0.01 | 0.42<br>p=0.09 | 0.89<br>p=0.00 | × | 0.89<br>p=0.00 |
| Young | 0.42<br>p=0.09 | 0.93<br>p=0.00 | 0.77<br>p=0.00 | 0.89<br>p=0.00 | 0.71<br>p=0.00 | 0.71<br>p=0.00 | 0.64<br>p=0.01 | 0.89<br>p=0.00 | × |

Figure 3.3: **Kendall-tau rank correlations between fairness parameter $\lambda$ and accuracy of protected attribute classifier.** For each pair of protected attribute (row) and target (column), we train 12 regularized ResNet50 models while varying the fairness parameters $\lambda$. Given their fixed last-layer representations, we train linear classifiers to predict the protected attribute. From the resulting 12 pairs of $\lambda$ and protected attribute accuracy, we test for a monotonic relationship between $\lambda$ and group accuracy by computing the Kendall-tau rank correlation. The color and size of a square correspond to the value and magnitude of the correlation coefficient. **For almost all tasks, the accuracy of predicting the protected attribute increases as the fairness parameter increases.**

large, we discard models with accuracy in the lowest quartile between the accuracy of a constant classifier and the accuracy of an unconstrained classifier.

**Results.** Figure 3.3 summarizes our results. For 71 out of 72 experiments the Kendall-tau correlation shows a positive correlation between $\lambda$ and the protected attribute accuracy. In 62 out of the 72 experiments, we can reject the null hypothesis of independence at a significance level of $p < 0.05$. The correlation in those experiments shows a very strong monotonic relationship with a correlation higher than 0.49. For

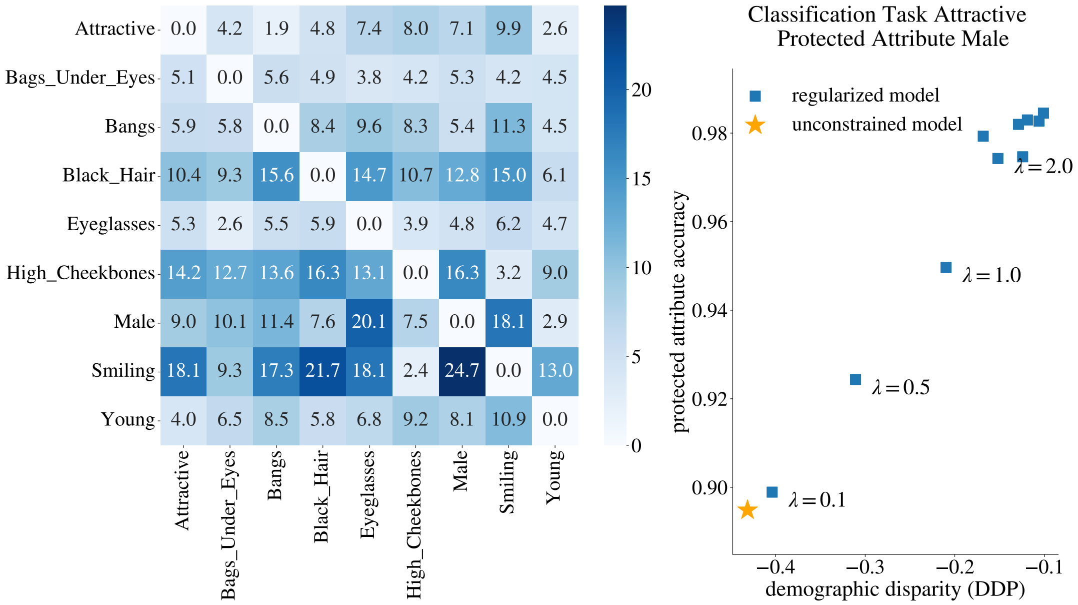


Figure 3.4: (Right) **Increase in protected attribute awareness.** For target ATTRACTIVE and protected attribute MALE, we show how the linear separability of gender increases together with the regularization strength. For all experiments, the network is first trained with fairness constraints, and the group classifier is subsequently trained on the frozen last-layer representation to predict the protected attribute. (Left) **Maximum increase of protected attribute accuracy.** Compared to the unconstrained model, we show the highest difference to the second head accuracy of fair models.

nine experiments, we observe lower but still positive values of the correlation coefficient $\tau$ and higher p-values. Only when the protected attribute is ATTRACTIVE and the target is BANGS, we find a negative correlation, but with an insignificant p-value of $p = 0.64$. The right panel of Figure 3.4 depicts the behavior of one specific experiment. For this experiment, the protected attribute is MALE and the target is ATTRACTIVE. As we increase the fairness parameter $\lambda$, we see a monotone increase in the group classifier accuracy (up to 8%). The maximum difference in group accuracy between the unconstrained model and the fairer models is shown in the left panel.

**Conclusion. Increasing the fairness regularization of neural networks makes it easier to recover the protected attribute.** We often find a strong monotonic relationship between the fairness parameter and the ability to recover the protected attribute using information in the last layer. Moreover, this relationship was always statistically significant when MALE was the protected attribute. As Lipton et al. (2018) observed that disparate treatment can occur if a system is able to infer the protected attribute (cf. Section 3.1.2), this increase in accuracy is a cause for concern. However, it does not guarantee that knowledge of the protected attribute is used and that disparate treatment is occurring. Sections 3.3 and 3.4 build on this initial analysis and show that it occurs in practice.

## 3.3 A new fairness method: Explicit Group Awareness to Enforce Demographic Parity

To understand if and how existing approaches use knowledge of the protected attribute, we contrast their behavior with that of a novel approach that explicitly infers the protected attribute and rescores individuals with it. As well as analyzing other existing approaches, the approach is valuable for its ability to efficiently find accurate classifiers with a particular demographic disparity, and it's reliability in generating classifiers with extremely low demographic disparity.

**Model Construction.** We consider a standard network backbone (e.g., ResNet 50), with two heads. The first head $f : \mathcal{X} \to \mathbb{R}$ outputs a score for predicting the target attribute and the second head $g : \mathcal{X} \to \mathbb{R}$ a score for the protected attribute (see Figure 3.2 for a sketch). We present the esults with a ResNet50 in this chapter, and include further results with the MobileNetV3-Small architecture in Appendix B.2.

The first head $f$ is trained to minimize binary cross entropy, while $g$ minimizes the mean squared error. This results in second-head outputs that are close to the binary labels $\{0, 1\}$ rather than a calibrated score. We minimize the objective:

$$\widehat{\mathrm{L}}(f, g) = \frac{1}{n}\left(\sum_{i=1}^{n} \widehat{\mathrm{L}}_{\mathrm{BCE}}(\sigma(f(x_i)), y_i) + (g(x_i) - s_i)^2\right).$$

We train the two heads directly, rather than using a regularizer or preprocessing to achieve fairness. After training, we combine the two heads $f$ and $g$ to create a new fair scoring function $F$ that maximizes accuracy on the original classification task while enforcing fairness. The function $F$ takes the form $F(x) = f(x) + a_1 g(x) + a_2$ for coefficients $a_1, a_2 \in \mathbb{R}$. Note that at test time $F$ can be compressed into a single head, that is if $f(x) = w_f \cdot z(x) + b_f$ and $g(x) = w_g \cdot z(x) + b_g$, with $z(x)$ being the last-layer representation, then $F(x) = (w_f + a_1 w_g) \cdot z(x) + (b_f + a_1 b_g + a_2)$.

To find coefficients $a_1$ and $a_2$ such that the predictions of the thresholding rule $\mathbb{1}(F(x) > 0)$ are fair and maximally accurate, we perform a grid search on validation data. The grid search procedure of our two-headed approach chooses all combinations of $a_1$ and $a_2$ from a grid of 200 equidistant points between $-15$ and 15. Going through all combinations, we choose the most accurate model that matches a given demographic disparity. We continue to search in the interval of the grid points which are closest to the current solution by forming another grid of 200 equidistant points in this interval. We continue this recursion 4 times. We compare our approach to Lipton et al. (2018) who's similar approach provided optimal per-group thresholds but requires explicit knowledge of the protected attributes at test time.

Table 3.1: **Accuracy under strict fairness constraints.** We show the effectiveness of various approaches at substantially decreasing demographic disparity. A cross indicates failure to reach the required fairness. For small reductions in disparity, all methods have a similar accuracy-fairness trade-off. However, the regularized approach repeatedly failed to find sufficiently fair solutions. For greater reductions, it does not reliably find a sufficiently fair model in seven out of eight cases. **Our two-headed approach always finds a sufficiently fair solution** and is comparable to Lipton's approach, which, unlike ours, requires the protected attribute at test time.

| | Attractive | Bags_Under_Eyes | Bangs | Black_Hair | Eyeglasses | High_Cheekbones | Smiling | Young |
|---|---|---|---|---|---|---|---|---|
| 50% disparity reduction | | | | | | | | |
| Lipton | 0.7955 | 0.8452 | 0.9469 | 0.8959 | 0.9680 | 0.8641 | 0.9240 | 0.8770 |
| Our Approach | 0.8021 | 0.8419 | 0.9463 | 0.8905 | 0.9775 | 0.8615 | 0.9227 | 0.8756 |
| Massaging | 0.7992 | 0.8337 | 0.9481 | 0.9002 | 0.9718 | 0.8546 | 0.9207 | 0.8679 |
| Regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ | 0.8021 | 0.8274 | ✗ | 0.9023 | ✗ | 0.8350 | ✗ | 0.8701 |
| 80% disparity reduction | | | | | | | | |
| Lipton | 0.7719 | 0.8344 | 0.9336 | 0.8959 | 0.9632 | 0.8456 | 0.9137 | 0.8617 |
| Our Approach | 0.7698 | 0.8332 | 0.9301 | 0.8881 | 0.9647 | 0.8436 | 0.9118 | 0.8592 |
| Massaging | 0.7674 | 0.8185 | ✗ | 0.8989 | ✗ | ✗ | ✗ | 0.8573 |
| Regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ | ✗ | 0.8274 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |

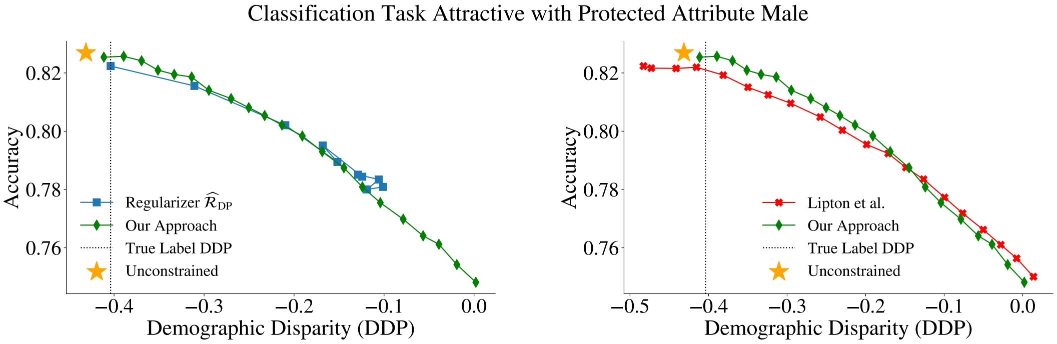


Figure 3.5: **Comparison of different fairness approaches.** We compare our group aware model to a fairness-regularized model (*left plot*) and the approach of Lipton et al. (2018) (*right plot*) on when predicting the target ATTRACTIVE with respect to the protected attribute MALE. For all methods, we observe the typical trade-off: as the model becomes fairer (DDP is closer to 0), the target accuracy for ATTRACTIVE decreases. All methods obtain similar accuracy for a particular DDP value. However, the regularizer approach is unable to achieve near perfect fairness and saturates around a DDP value of $-0.1$. Note that Lipton et al. (2018) requires the protected attribute at test time, while we infer the protected attribute. More datapoints are shown in the scatter plot for Lipton et al. (2018), and our approach because they can be generated by varying thresholds without retraining. In contrast, the regularized approach, requires a full-retraining for every choice of fairness parameter.

If the group classifier $g$ is perfect, our approach and Lipton's coincide. If $g$ does not predict the protected attribute well, the classifier that our procedure yields can perform only worse than the one obtained from the approach of (Lipton et al., 2018) (at best, the two classifiers are the same; cf. Theorem 4 in their paper). In our computer vision setting, however, the accuracy of predicting the protected attribute is typically very high (e.g., on MALE from CelebA, we achieve an accuracy of around 98%), and, as we show, the performance of our approach is very close to that of (Lipton et al., 2018).

**Experimental Setup.** We train the regularizer approach with the same range of 12 parameter values as in Section 3.2. Since the computational cost of training 12 models is much higher than training one model of our two head approach or one unconstrained model for Lipton, we compute solutions with our grid search or Lipton's greedy search for 20 equidistant fairness constraints between perfect fairness and the fairness of the unconstrained classifier.

**Results.** Figure 3.5 compares the accuracy-fairness trade-off of our approach, the regularizer approach and Lipton et al. (2018) for a range of fairness parameters. All three methods offer similar accuracy for a particular choice of demographic disparity.

However, for the regularized model, it is difficult to control this trade-off or to find extremely fair solutions. In contrast, our approach and (Lipton et al., 2018) allow easy selection of a model with a particular demographic disparity. While the performance of our approach and (Lipton et al., 2018) is similar, we do not require the protected attribute at test time.

Table 3.1 reports the accuracy obtained under strict fairness constraints. In the first block of rows, we require the minimum fairness improvement compared to the unconstrained classifier to be 50%, that is we want, in absolute value, the DDP to be at most half of the DDP of the unconstrained classifier. For all methods, we choose the most accurate model among the models that are fair enough. In the second block, we want the DDP to be at most 20% of the DDP of the unconstrained model. With respect to the accuracy-fairness trade-off all methods perform comparably. However, if we require a substantial reduction of unfairness, the regularizer approach and preprocessing often fail to find a valid solution. In contrast, Lipton et al. (2018) and our approach, always find sufficiently fair solutions due to their direct search for per-group thresholds.

**Conclusion.** Our method has several advantages compared to standard approaches. (i) The latter requires training a new model for every fairness parameter $\lambda$, which might make tuning $\lambda$ very expensive until a desirable level of fairness is reached. Our approach on the other hand requires a single explicit model and the output scores of the two heads. (ii) We make the influence of the group classifier in the final decision explicit and transparent. Due to the simple weighted sum, we can determine the different group-wise decision thresholds. In summary, **our approach reliably finds high accuracy solutions for a given demographic disparity** without requiring the protected attribute at test time.

## 3.4 Disparate Treatment in Fair Networks

Here we examine the tight relationship between our explicit approach and the behavior of fair networks. Given a fair neural network, we aim to recover the corresponding unconstrained model using only the fair network and the protected attribute classifier from our explicit approach presented in Section 3.3. Similarly, we reconstruct the fair network with our group-aware method by building a weighted sum of the target classifier and the protected attribute classifier. These reconstructed decisions allow us to identify individuals treated differently based on inferred group membership and demonstrate disparate treatment.[3]

### 3.4.1 Fair Networks Behave like the Explicit Approach

We recover the predictions of the fair model $r_\lambda$ by using the target task head $f$ and group classifier $g$. We use logistic regression to find parameters $a_1, a_2 \in \mathbb{R}$ such that $\mathbb{1}(f(x) + a_1 g(x) + a_2 > 0)$ accurately replicates $\mathbb{1}(r_\lambda(x) > 0)$. Given fixed $f$ and

[3] In this chapter, we report results for the regularized approach (3.3). Appendix B.3 reports preprocessing results.

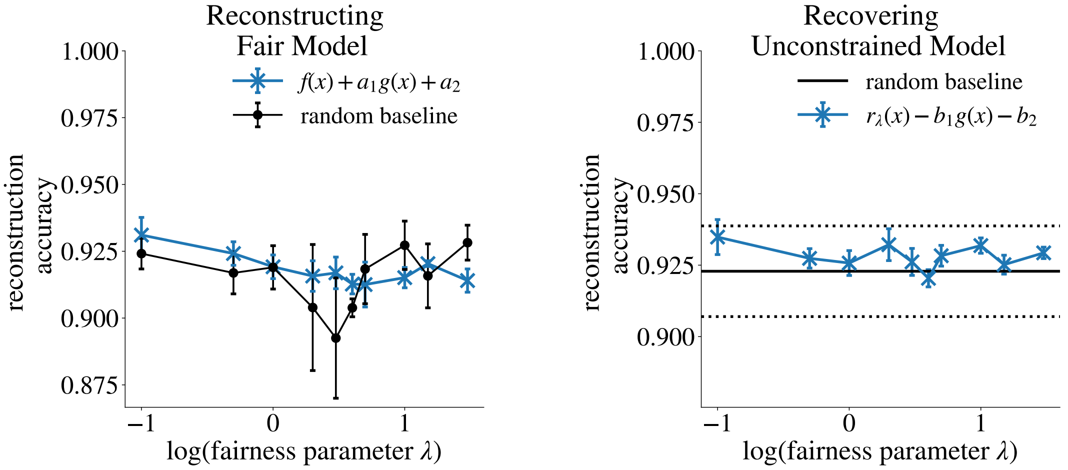


Figure 3.6: *Left:* **Reconstructing the fair classifier.** For a range of parameters $\lambda$ (x-axis) we find $a_1$ and $a_2$ so that $f + a_1 g + a_2$ mimics the predictions of a regularized classifier $r_\lambda$. For the entire range of fairness parameters, the predictions of the regularized model are closely recovered by the two-headed approach. *Right:* **Recovering the unconstrained classifier.** We find parameters $b_1$ and $b_2$ such that $r_\lambda - b_1 g - b_2$ recovers the predictions of an unconstrained classifier $h$. From the fair model $r_\lambda$ and the protected attribute classifier $g$ we can replicate an unconstrained classifier's predictions, as accurately as another unconstrained model. In this example we predict ATTRACTIVE, the protected attribute is MALE, and we use a ResNet50. **Random Baseline.** We also measure the disagreement among retrained models initialized with different random seeds. For the left figure, we retrain the fair model for each $\lambda$, on the right, we retrain the unconstrained model.

$g$, we repeat this process using five different initial random seeds of $r_\lambda$. Next, we show that it is possible to recover the predictions of the unconstrained model $h$ with a weighted sum of the fair classifier $r_\lambda$ and the group classifier $g$. We run logistic regression to find $b_1, b_2 \in \mathbb{R}$ such that $\mathbb{1}(r_\lambda(x) - b_1 g(x) - b_2 > 0)$ recovers the prediction $y = \mathbb{1}(h(x) > 0)$.

Coefficients $a_1$ and $a_2$, or $b_1$ and $b_2$ are found using validation data. We learn new coefficients for every $\lambda$ and random seed; repeating this second experiment for five unconstrained models trained with different random seeds.

For both experiments, a substantial challenge lies in the random behavior of deep learning classifiers. Training a deep network is a nonconvex problem, and the solution found is highly dependent on its initial seed. To take this instability into account, we provide baselines that measure how decisions vary when retraining networks. We retrain regularized networks for every choice of $\lambda$ in the first experiment, and retrain the unregularized network in the second.

**Results.** In Figure 3.6 we show ResNet50 models trained to predict ATTRACTIVE; the protected attribute is MALE (see Appendix B.3 for other tasks). The left panel evaluates how accurately our explicit approach recovers the predictions of the regularized model $r_\lambda$. We find that most of the error in recovering predictions can be attributed to classifier instability, and that retraining a classifier from scratch with a new random seed gives similar disagreement to using our reconstructed classifier. In the right panel of Figure 3.6, we plot the reconstruction accuracy of $r_\lambda - b_1 g - b_2$ with respect to the unconstrained classifier. By simply adding the group classifier response $g(x)$ to $r_\lambda(x)$, we obtain the predictions of the unconstrained classifier. **Compared to the baseline, we recover the unconstrained classifier responses for all $r_\lambda$ with similar fidelity to simply retraining the target classifier from scratch.**

### 3.4.2 Identifying Disparate Treatment in Deep Networks

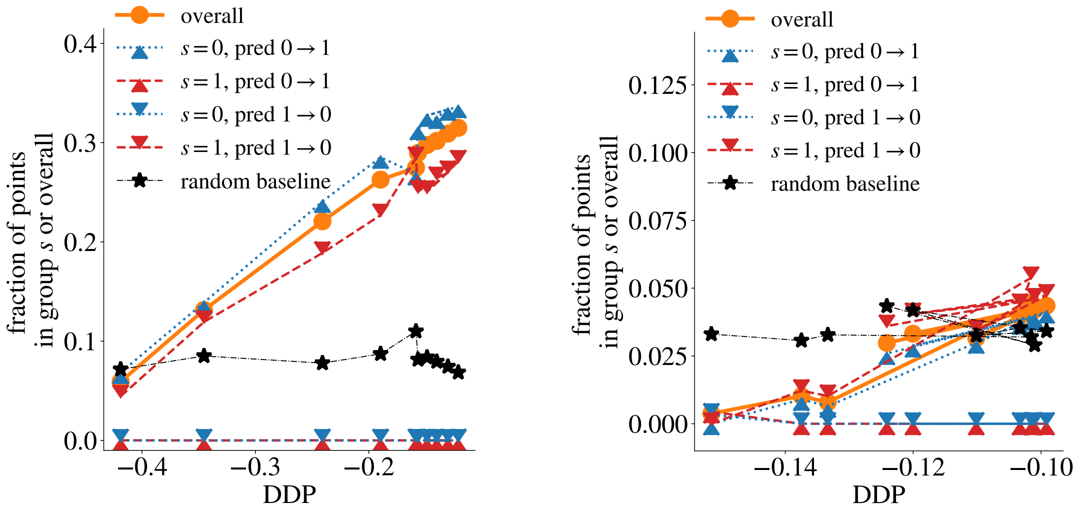
We can now quantify the disparate treatment of a neural network. By exploiting our explicit estimation of the protected attribute, we can ask the counterfactual question: how would the decision have changed if the individual had belonged to the other group?

**Experimental Setup.** As described in the previous subsection, we find the closest weighted sum $f + a_1 g + a_2$ of the two heads that best replicates the decisions of a given model $r_\lambda$. Then, for every individual $x$ in the test set, we replace the group classifier response $g(x)$ with the median output of the group that $x$ does *not* belong to. We evaluate how many times the prediction changes when the second head output is replaced by this counterfactual.

**Results.** Figures 3.7 and 3.8 show the proportion of individuals for whom their prediction changes. For the fairest models in the left panel of Figure 3.7, up to 30% of all individuals receive a different outcome when their second head output $g(x)$ is replaced by the median output of the other group. This is substantially more than the number of changed predictions, which we obtain when retraining with a different random seed (roughly 7% of the points).

As expected, the proportion of changed predictions linearly increases with model fairness (governed by the parameter $\lambda$). Similar behavior occurs for both the regularized approach (Figure 3.7, left) and preprocessing (Figure 3.8, right).

While the behavior of our two-headed system is difficult to distinguish from that of a retrained fair classifier, the disagreement between retrained classifiers means that we can not point to an individual and conclude that they received a different decision because of their protected attribute. Nonetheless, in scenarios where changing the protected attribute alters a much greater proportion of decisions than the proportion of decisions where the classifiers disagree (see Figure 3.7 left, in contrast to center) we can conclude that it is likely particular individuals suffered disparate treatment.

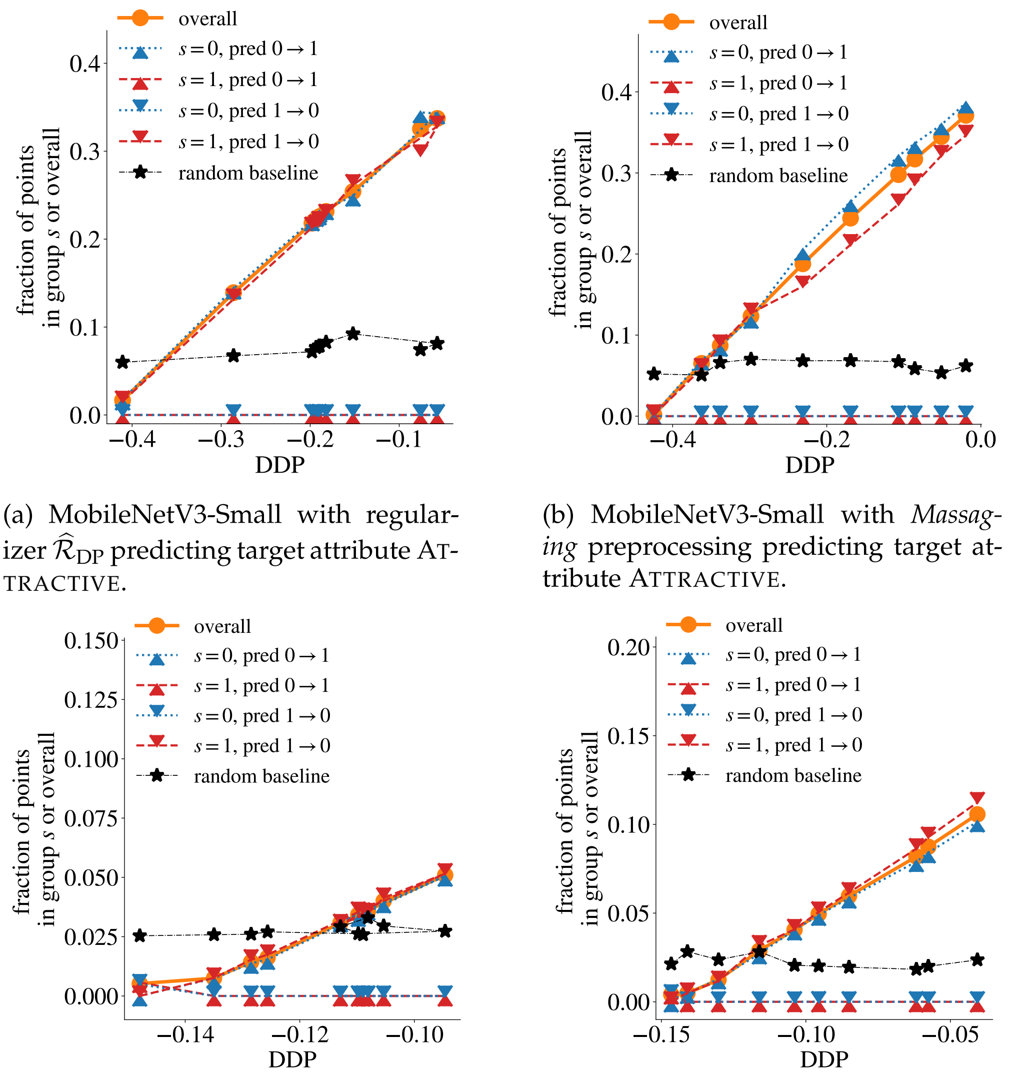


(a) ResNet50 with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ predicting target attribute ATTRACTIVE.

(b) ResNet50 with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ predicting target attribute SMILING.

Figure 3.7: **Uncovering disparate treatment—proportion of changed predictions under counterfactual group classifier.** How many individuals are treated differently based on their protected attribute? Using our two-headed approach, we replace the group classifier output $g(x)$ of an individual $x$ from group $s \in \{0,1\}$ with the median output $\bar{g}_{1-s}$ of the other group $1-s$. We plot the proportion of all points where the label changes (orange curves), and the proportion of points in each protected group for which the prediction either changed from 0 to 1 or changed from 1 to 0 (red and blue curves with markers pointing up or down). *Left:* As the fairness parameter increases and fairness of the regularized model improves (DDP closer to 0 is fairer), the proportion of changed predictions increases. For the fairest model, around 30% of points would obtain a different outcome if their perceived gender changed. Moreover, when the positive label is a benefit, only the disadvantaged group benefits and only the advantaged group is harmed from a changed protected attribute. *Right:* We also observe cases where there is a only small change in demographic disparity and no substantial proportion of points are treated differently based on the protected attribute.

**Conclusion.** When fair networks show the same behavior as our explicit awareness model, we can analyze the influence of group membership. Using our explicit approach, we can evaluate how fair networks systematically treat individuals differently on the basis of their protected attribute.



(a) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ predicting target attribute Attractive.

(b) MobileNetV3-Small with *Massaging* preprocessing predicting target attribute Attractive.

(c) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ predicting target attribute Smiling.

(d) MobileNetV3-Small with *Massaging* preprocessing predicting target attribute Smiling.

Figure 3.8: **Uncovering disparate treatment—proportion of changed predictions under counterfactual group classifier.** We perform our analysis described in Section 3.4.2 on CelebA with target attributes Attractive and Smiling and protected attribute Male. For both the regularizer and the preprocessing (Kamiran and Calders, 2012), up to 35% of all points would receive a different outcome if their inferred attribute changed for the task Attractive. As expected, only in one group negative predictions change into positive predictions; at the same time only for the other group positive prediction change to negative predictions.

## 3.5 Legal Implications of our Analysis

The analysis set out in this section is restricted to areas where the doctrine of disparate treatment is relevant. This includes areas where decisions are made concerning an individual's access to: education, employment, and housing. We start by noting that by design our two-headed approach exhibits disparate treatment. It assigns individuals into different racial or gender-based groups and uses this to alter cutoff scores in a way that explicitly violates Title VII (Statute, 1991) as outlined in the introduction. What is less straightforward is the relationship of the methods that we have shown to have the same systematic behavior as our new approach.

**Overview** Our argument can be decomposed into three parts. We address each point in detail below:

1. Disparate treatment may occur even if the treatment is based on inferred attributes (such as race or sex) rather than explicitly provided attributes.

2. Our explicit formulation (Section 3.3) exhibits the behavior used to *indirectly* identify disparate treatment.

3. Other fairness approaches considered exhibit the same behavior as our explicit approach, and as the relevant tests for disparate treatment are based on systematic behavior, these approaches also exhibit disparate treatment.

Finally, we explicitly identify the individuals likely disadvantaged by enforcing fairness, and consider circumstances where the use of such systems may be acceptable, and look at existing legal arguments.

**1. Implicit Disparate Treatment** Multiple machine learning papers (Agarwal et al., 2018; Donini et al., 2018; Perrone et al., 2021; Zafar et al., 2017a) have asserted that machine learning systems that do not explicitly take into account knowledge of the protected attribute at prediction time cannot be performing disparate treatment. From a legal perspective, this is an oversimplification. Many of the legal tests for disparate treatment involve a demonstration of intent to treat protected groups differently (King and Hemenway, 2020), and it is irrelevant if knowledge of the groups is given as data from a trusted party or inferred from a photograph or other data. As an example of case law supporting this, (Hellman, 2020) gives Hunt v. Cromartie (Supreme Court, 1999) where the plaintiffs demonstrated that location was used as a proxy for race in an instance of disparate treatment. The situation considered here is even more extreme than that of Hunt v. Cromartie. As we use photographic data as input, to deny that disparate treatment can occur here is the same as denying that disparate treatment can occur on the basis of an individual's appearance.

**2. The Disparate Treatment of Our Approach** Systems which explicitly alter scoring on the basis of race or gender are widely acknowledged as being examples of disparate treatment, with (Hellman, 2020) writing that there is such wide agreement that it is not worth discussing. While creating a system that makes explicit use of a protected attribute when making decisions demonstrates intent, it is not the only way to do so. In particular, as it is difficult to explicitly demonstrate intent when someone is either unable or unwilling to explain honestly why they made decisions, the courts recognize indirect evidence of the form: ". . . evidence, whether or not rigorously statistical, that employees similarly situated to the plaintiff other than in the characteristic (pregnancy, sex, race, or whatever) on which an employer is forbidden to base a difference in treatment received systematically better treatment" (Troupe v. May Dept. Stores, 1994). By design, our approach explicitly treats 'similarly situated' individuals, i.e. those who receive a similar score $f(x)$ from a classifier trained without consideration of their protected attribute, differently by changing their score and adding the term $a_1 g(x) + a_2$ which explicitly depends on their inferred protected attribute.

**3. The Disparate Treatment of Other Approaches** As the implicit argument of the previous section relies on the systematic behavior of a decision-making system, it can be directly applied to systems trained to satisfy a fairness constraint. In such cases, we only know that the system enforces the constraint, but not necessarily how. However, as we are only concerned with the system behaviour, i.e. the decisions made, if the system is closely mimicked by our approach, the same argument applies, and we can identify those individuals with similar scores $f(x)$ who probably[4] receive different decisions by virtue of their race or gender.

**Identifying Disadvantaged Individuals** We can therefore identify individuals that are likely to have received an unfavorable decision by virtue of their inferred protected attribute. As we can recover a close approximation of the decisions of the fair model of the form $r(x) \approx f(x) - a_1 g(x) - a_2$, individuals who initially receive a score $f(x)$ in the region $f(x) \in [a_2, a_1 + a_2)$ are likely to receive different decisions by virtue of their inferred race or gender $g(x)$[5].

**Potential Mitigation** One possible defense is to reject the relevance of the classifier $f$ trained on ground-truth data without fairness considerations, and to claim that individuals with similar $f(x)$ scores are not in fact "similarly situated" (Troupe v. May Dept. Stores, 1994). It is always possible to generate some classifier[6] $f$ from an existing classifier $r$, by subtracting any arbitrary terms of the form $a_1 g(x) + a_2$ from $r(x)$. As

[4] As the reconstruction is not exact, we cannot be certain, however, note that the evidence provided does not even need to be *rigorously statistical* (Troupe v. May Dept. Stores, 1994).

[5] This relies on $g(x)$ being close to an indicator function with 94% of function responses being either 0 or 1 with a tolerance of ±0.1.

[6] Here, we consider $f$ an arbitrary classifier, and not necessarily the unconstrained classifier trained on the data.

such, the existence of $f$ is insufficient to conclude that the existing classifier $r$ exhibits disparate treatment, and we also need to know that individuals with similar scores $f(x)$ are "similarly situated". As such, where this unaware classifier $f$ was trained on data that does not correspond to a direct measure of performance, and is known to be systematically biased (Barocas and Selbst, 2016; Wachter et al., 2021), there is little reason to believe that individuals with similar scores are also "similarly situated". This argument was proposed outside of algorithmic fairness by Selmi (2013) who argued that a stronger defense could have been mounted in (Supreme Court, 2009) by challenging the predictive value of the test.

**Existing Legal Arguments** A full summary of the existing debate is out of scope for what is primarily a machine learning paper, and as such we touch upon three papers to indicate the range of opinions. Kroll et al. (2017) argued that protected attributes should not be used as part of the decision-making system, but that the use of ML fairness methods that use protected attributes at training time was less likely to be considered an instance of disparate treatment than an ex post correction that explicitly alters the score of individuals with a particular race or gender. In particular, Kroll et al. (2017) considered (Troupe v. May Dept. Stores, 1994) and similar judgements and concluded that compared to ex-post measures[7] "incorporating nondiscrimination in the initial design of algorithms is the safest path that decision makers can take." This argument hinges upon an explicit lack of understanding of the behavior of ML systems, i.e., without knowing the details of how an opaque fair system behaves it is not possible to say that it exhibits disparate treatment. Hellman (2020) argued that as disparate treatment depends upon intent, in certain circumstances, the use of protected attributes can be acceptable at decision time. Bent (2019) offered two main arguments: First, that as disparate treatment hinges upon intent, the combined use of racial data (even if only at training time) with any form of fairness constraints shows an expressed intent to treat different races differently, and should also be considered disparate treatment[8] [9].

In response to Bent (2019), we note that: *(i)* bias-preserving fairness metrics (Wachter et al., 2020) (i.e., the majority of existing fairness metrics) ensure that classifiers are sufficiently accurate for all groups, by matching the distribution of errors over different groups. *(ii)* Kroll et al. (2017) also argued that in (Supreme Court, 2009)

[7] Ex-post methods adjust an already generated scoring system by choosing different thresholds for members of different groups.

[8] Bent (2019) argues that this should hold for any form of fairness constraint, not just the forms of demographic parity we consider.

[9] The second argument Bent (2019) considers is the difference between running an algorithm with and without race-based fairness constraints. Bent argues that any individual receiving a change in decision should be considered evidence of a difference in treatment. This argument is problematic due to the non-deterministic behavior of deep learning algorithms. As shown in Figures 3.6 and 3.7, simply rerunning the same algorithm with a different seed can result in significant changes in labels assigned to the test set, and what is needed is evidence of a systematic change in treatment over and above the expected intrinsic variability. Inherently, such evidence cannot come from considering a single individual, but must occur at the population level.

the court's lack of concern regarding the use of race to determine that the scoring mechanism was equally effective for all racial groups indicated that this was a legitimate use of racial data. Putting these two arguments together, it seems likely that enforcing many forms of fairness need not be a form of disparate treatment, and that it depends instead on specific facts of implementation and particularly if it is possible to identify individuals who are systematically disadvantaged by the fact of their race under a fair system.

Our position lies midway between Kroll et al. (2017) and the first argument of Bent (2019), and is simply that a blanket decision if algorithmic fairness violates disparate treatment is inappropriate and depends upon the facts of the particular ML system considered and the data it is deployed on. Our position is that the improved understanding from using the techniques set out in this chapter is sufficient to determine that some ML fairness systems also exhibit disparate treatment (see Section 3.2 and Figure 3.1); and, more importantly, the decisions made by a fair regularized classifier are indistinguishable from those training a classifier designed to exhibit disparate treatment (see Section 3.3).

**Conclusion.** While it may not be possible to determine a priori if a particular fairness definition gives rise to disparate treatment, our decomposition of existing fair classifiers into an unconstrained classifier $f(x)$, and a second head that rescores the response using inferred race or gender, strongly aligns with the legal definition of disparate treatment. As such, we believe that this decomposition will be of value in determining the legality of deploying fair systems in practice. From a practical perspective, this disparate treatment is most strongly observed when the unconstrained classifier $f$ has high demographic disparity, and demographic parity is strongly enforced. This makes it unlikely for any of the considered fairness methods to be appropriate tools to enforce equity without legal reform.

## 3.6 Other Related Work

**Exploiting Disparate Treatment** Oneto et al. (2019) propose to infer the protected attribute from non-protected features and to use the inferred attribute to learn "group specific" models, as a way to enforce equalized odds. While their approach is similar to our two-headed approach presented in Section 3.3, their interpretation is quite the opposite from ours: they consider their approach as a means of overcoming disparate treatment while we argue that such an approach should not be treated differently than an approach that explicitly uses protected information. Other papers proposed the use of protected information for learning group-dependent models, when doing so is legally acceptable and the protected information is available to improve performance and / or fairness (Dwork et al., 2018; Klare et al., 2012; Ustun et al., 2019).

**Bias and Bias Amplification in Deep Neural Networks** Numerous papers have found deep networks discriminate based on protected groups (Albiero et al., 2020; Balakrishnan et al., 2020; Buolamwini and Gebru, 2018; Feldman and Peake, 2021; Klare et al., 2012) and may even amplify bias present in the training data (Burns et al., 2018; Jia et al., 2020; Prates et al., 2020; Wang and Russakovsky, 2021; Wang et al., 2019b; Zhao et al., 2017). Deep models have also been found to "overlearn", that is they learn representations encoding concepts that are not part of the learning objective; e.g., encoding race when trained to predict gender (Serna et al., 2020; Song and Shmatikov, 2020). (Song and Shmatikov, 2020) argued that overlearning is problematic from a privacy perspective as it reveals sensitive information. However, they did not consider if it allows models to disparately treat different groups. To detect unintended classifier bias, (Balakrishnan et al., 2020; Denton et al., 2019) synthesized counterfactual images by changing latent factors of a generative model, corresponding to attributes such as race, and seeing how performance alters. This is related to our approach, however, they do not examine how fair models alter this behavior, and our decomposition of models into two heads allows us to reason counterfactually without generating images.

**Bias Mitigation Methods** In the last few years, a plethora of fairness notions, that is definitions of fairness-concerning *bias*, along with methods for mitigating such bias have been proposed, both in supervised and unsupervised learning. The methods in supervised learning are usually categorized into three groups: preprocessing methods, in-processing methods, and postprocessing methods (see (Caton and Haas, 2020; Mehrabi et al., 2021) for survey papers). In this chapter we study methods from each of the three groups (cf. Section 3.1): the regularizer approach belongs to the group of in-processing methods (and so does our strategy proposed in Section 3.3), the massaging method of (Kamiran and Calders, 2012) is a preprocessing method, and the strategy of (Lipton et al., 2018) is a postprocessing method. While the earlier papers on fair ML mainly considered tabular data, more recently, bias mitigation has also been studied in the context of deep learning (Du et al., 2020; Ramaswamy et al., 2021; Wang et al., 2019a,b, 2020). Fazelpour and Lipton (2020) discuss a broader human-centered view going beyond altering algorithms with parity metrics.

**Fair Representation Learning** This line of work proposes techniques to learn data representations such that an ML model trained on top of such a representation is fair, without enforcing the latter model to be fair (Adel et al., 2017; Alvi et al., 2019; Beutel et al., 2017; Edwards and Storkey, 2016; Feng et al., 2019; Jia et al., 2018; Louizos et al., 2016; Madras et al., 2018; Moyer et al., 2018; Raff and Sylvester, 2018; Xie et al., 2017; Zemel et al., 2013; Zhao and Gordon, 2019). These techniques come in various flavors, and depending on the concrete design they are considered either as preprocessing or in-processing methods. When aiming for demographic-parity-fair classifiers, they try to learn representations that do not contain any information about the protected attribute, or are independent of the attribute. Hence, for these techniques one should not observe the phenomenon of attribute awareness we found for networks regularized to satisfy demographic parity or trained on massaged datasets, at least not when examining the final representation layer. However, it is an interesting question for future work whether methods for fair representation learning suffer from attribute awareness in earlier network layers.

## 3.7 Conclusion

This chapter makes two contributions to the literature. *First*, our two-headed approach for enforcing fairness offers a more efficient and reliable way of enforcing a chosen degree of demographic parity. Unlike regularization-based approaches that require multiple training runs with different regularization parameters to find a desired trade-off, our approach allows for joint training of both heads once, and then a search for the desired trade-off by tuning weights using precomputed network responses on a validation set. As such, it is possible to efficiently find a family of classifiers of varying fairness and accuracy for little more compute than simply training an unfair classifier in the first place.

*Second*, we have shown how existing methods for enforcing demographic parity in deep networks learn a latent representation that is more predictive of the protected attribute. We have shown a close coupling between the behavior of these approaches to the explicit two-headed model. This tight coupling allows us to identify individuals who are likely to be systematically favored or disfavored by virtue of their protected attribute and to conclude that existing methods for enforcing fairness also enforce disparate treatment.

In hindsight, our findings are perhaps unsurprising. The existing methods considered can be seen as altering the labels[10] assigned to individuals in the training set, on the basis of their protected attribute, and as the trained network generalizes from training to test data it brings this behavior with it. Still, it requires the comparison to our two-headed approach to confirm these intuitions and to demonstrate that disparate treatment of this kind not only happens on training data, but also on test data. We wish to emphasize that we have provided tools for identifying disparate treatment, and shown that the use of some fairness methods on some datasets exhibit disparate treatment. Our findings do not imply that *all* methods for enforcing demographic parity suffer from disparate treatment, merely that some can, and that caution should be used when deploying such methods.

While our analysis presents several challenges to deploying fair ML systems in the US, it is consistent with other rulings on discrimination in US law. In general, the US requires that considerations of equity and affirmative action are satisfied by shaping an entire process to be more inclusive, and not simply by imposing race or gender-based quotas on outcomes (Fazelpour and Lipton, 2020; Joshi, 2018). However, the opaque nature of ML makes it extremely challenging to define fair algorithms without formulating the definitions in terms of outcomes. For this reason, we believe that legal reform is needed to explicitly allow the use of fair ML techniques as a tool to reduce disparate impact and increase equity.

[10] Regularized approaches do so by enforcing a soft quota, while preprocessing explicitly relabels individuals.

# Chapter 4

# The Fairness of Crowds: If the Few are Fair, are the Many?

Recent work about fairness in algorithmic decision-making has typically considered a single machine learning model. In practice, however, multiple models are combined in ensembles in order to boost the performance on a given task such as classification. The main idea of combining multiple models in an ensemble is to compensate the errors of a single model by the other models, and to leverage the *Wisdom of Crowds* to obtain more accurate and robust ensemble predictions. A popular, yet simple, approach for aggregating the predictions of various models into one prediction is the majority vote (Grofman et al., 1983), which is known to improve prediction accuracy under certain conditions. From a fairness perspective, it remains unclear whether the majority vote improves fairness properties when every single model is already fair. In this chapter, we analyze if there is a *Fairness of Crowds*: If we have a set of fair models, is the majority vote ensemble fair? And if not, how much fairness do we lose?

## 4.1 Introduction

Ensemble learning is a state-of-the-art approach for solving a wide range of machine learning tasks and, consequentially, it is often the leading approach in popular machine learning competitions (Sagi and Rokach, 2018). The ensemble learning paradigm is counting on the *Wisdom of Crowds* (Surowiecki, 2005): if we ask many individuals for their opinion on a matter, we can expect better decisions from the crowd than from any individual. This intuition has led to several approaches, such as Random Forests (Breiman, 2001), Boosting (Schapire, 1990, 1999), Bagging (Breiman, 1996), XGBoost (Chen and Guestrin, 2016), and many more (Friedman, 2001; Sagi and Rokach, 2018; Wolpert, 1992).

These ensemble approaches are built by training multiple machine learning models, called base learners, such as decision trees, neural networks, or logistic regression classifiers. The base learners can be trained in parallel, for example deep decision trees for Random Forests, or in succession—simple decision stumps for Adaboost. There are various ways to aggregate the predictions of the base learners into one ensemble prediction like weighting them, or learning a meta-classifier on the predictions (see Sagi and Rokach (2018) for an overview). In this chapter, we assume that we are given a set of trained base learners and we focus on majority voting as the aggregation scheme, a very old and popular strategy for decision making.

A first theoretical justification for the majority vote has been given in Condorcet's Jury Theorem, presented in 1785 by the Marquis de Condorcet, Marie Jean Antoine Nicalos de Caritat (Shapley and Grofman, 1984). Condorcet considers a jury of voters that are supposed to make a "correct" binary decision with a majority vote. If the decision of a single jury member is correct with a probability greater than 0.5, the probability of a correct decision by the jury can be increased by simply adding new members to the jury. Then, a correct decision from the jury is more likely than a correct decision from any single member. This result is obtained by a crucial assumption–Condorcet assumes that the jury members vote independently of each other. More recent work that is dedicated to understanding the majority vote has relaxed this assumption and has considered different dependencies between the classifiers (Battiti and Colla, 1994; Kuncheva et al., 2003; Lam and Suen, 1997; Matan, 1996).

In this work, we are concerned with the fairness properties of the majority vote. In particular, we want to understand how fair the majority vote is when we are given a set of fair base learners that have been trained with the multitude of fairness approaches for binary classification (Agarwal et al., 2018; Cotter et al., 2019; Donini et al., 2018; Goh et al., 2016; Hardt et al., 2016; Iosifidis and Ntoutsi, 2019; Lohaus et al., 2020; Manisha and Gujar, 2020; Perrone et al., 2021; Wick et al., 2019; Zafar et al., 2017a). Do the fairness properties of the individual classifiers transfer to the majority vote ensemble? In this preliminary work on this question, we make the following contributions.

1. Inspired by Condorcet's Jury Theorem, we prove for the group fairness notions demographic parity and equality of opportunity that a majority vote ensemble of fair classifiers retains the fairness property under strong independence assumptions (Section 4.5).

2. In Section 4.6, we drop all independence assumptions on the base learners. We present an algorithmic framework to construct the worst-possible configuration of base learners and thus, determine how unfair the majority vote ensemble can become when the individual classifiers are fair.

3. In Section 4.7, we evaluate fairness bounds on demographic parity and equality of opportunity. As a result, we show that in the worst case the ensemble can exhibit considerable unfairness, but the *worst-case guarantees* can be improved. In some cases, a reasonably accurate ensemble cannot be fair at all.

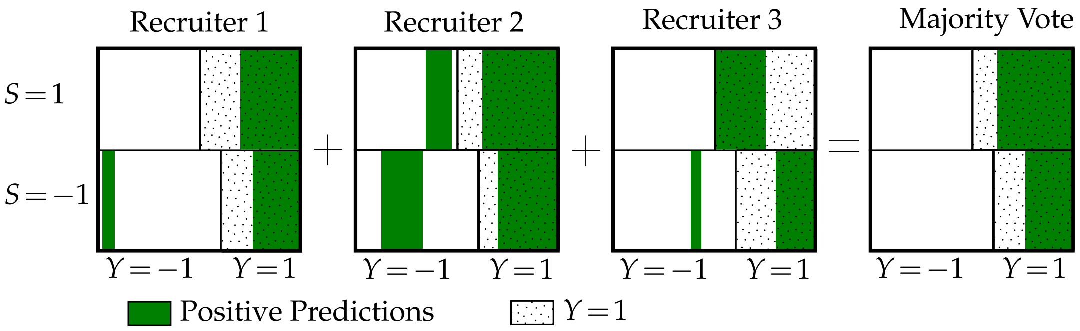


Figure 4.1: **Majority vote of three job recruiters.** We visualize the voting behavior of three recruiters on applicants from Group $S = 1$ (upper half of the box), applicants from Group $S = -1$ (lower half), qualified applicants $Y = 1$ (right quadrants), and unqualified applicants $Y = 1$ (left quadrants). Green areas correspond to the fraction of applicants that a recruiter accepts (positive prediction), for example Recruiter 1 accepts 30% from Group $-1$. For the majority vote it is important where the boxes are colored in green because the majority vote accepts a candidate only if at least two recruiters accept. In the plot this means that an area in the majority vote is only green if this area is green in at least two of the three boxes. **Each recruiter is accepting applicants from both groups at the same rate–the size of the green areas is the same in the upper and lower half. However, the majority vote accepts Group 1 at a higher rate than Group $-1$.**

## 4.2 Motivational Example

Imagine three recruiters interviewing potential employees for a company. Their task is to suggest qualified applicants and reject unqualified ones. Currently, each recruiter works on their own. In the past, the assessment of Recruiter 1 about an applicant‘s qualification was correct in 79% of all cases, with an overall acceptance rate of 30% (see Figure 4.1 for a visualization). Recruiter 2 and 3 were correct in 72.5% and 75% of all cases, with an acceptance rate of 50% and 25%, respectively. More importantly, all recruiters have not shown any bias with respect to two protected groups: there is Group 1 and Group $-1$. Applicants from both groups have had the same chance to be accepted.

In order to improve the quality of the interviewing process[1], the company requests that the three recruiters work together and make their decision about an applicant with a majority vote: at least two recruiters need to agree on suggesting the candidate to the company.

And indeed, the freshly founded jury correctly judges an applicant’s qualification 85.75% of the time. However, even though each individual recruiter is fair, the jury is not: an applicant from Group 1 is now 13.5% more likely to be accepted than an

[1] The overall accuracy of the previous approach is 75.5%, if we assume that each recruiter is equally likely to interview an applicant.

applicant from Group $-1$. Moreover, a qualified applicant from Group 1 is 15% more likely to be accepted than a qualified applicant from Group $-1$. **The majority vote ensemble of the three recruiters is more accurate, but unfair.**

In Figure 4.1, we visualize the votes of each recruiter to show why the their fairness properties are lost in the majority vote. The recruiters mostly vote for the same qualified applicants on Group $-1$. Recruiter 1 and 3 have more contrary opinions about qualified candidates on Group 1. Thus, on both groups the recruiters do not vote independently of each other. In Section 4.5, we investigate the ensemble fairness when the recruiters make decisions independently of each other on every protected group.

Even though we observed in Figure 4.1 that the majority vote can be unfair, we do not know how unfair it could be in the worst case. Can we construct other voting patterns (color the boxes differently) with the same accuracies and acceptance rates for each recruiter, but with worse ensemble fairness? In Section 4.6, we provide algorithmic bounds on the worst-case fairness when no assumptions are made on the independence of the individual classifiers.

## 4.3 Related Work

There is few work that focuses on the specific fairness properties of model aggregation. Most related to us is the work by Grgić-Hlača et al. (2017). They consider the fairness of random ensembles, which classify a data point by randomly picking a classifier from a diverse set of classifiers. This is a common scenario in decision-making, for instance in our example above one recruiter was randomly assigned to an applicant before the company decided to use the majority vote. Grgić-Hlača et al. (2017) show for demographic parity, equality of opportunity, and equal odds that the whole ensemble is fair when each single classifier is fair. This cannot be guaranteed for sufficiency notions like predictive parity. These findings are not easily transferable to different aggregation schemes of base learners. In this work, we address similar questions for the setting of a majority vote aggregation instead of a random selection.

Dwork and Ilvento (2018) analyze fairness under the composition of classifiers in a setting, where each classifier solves a different task (predicting gender or predicting age), but they compete for the same unit of benefit such as an ad placement. For this scenario, Dwork and Ilvento (2018) show that the composition (the placement of the final ad) need not be fair, even though each classifier is fair.

The technical tools that we use to analyze the fairness of ensembles are inspired by Matan (1996) and Narasimhamurthy (2003) that both analyze best and worst cases of the majority vote accuracy when given the performance of binary classifiers. Matan (1996) prove exact bounds as a function of the individual performances for the more general ‘k-out-of-n’ majority vote. Even though Narasimhamurthy (2003) do not provide explicit formulations of the worst and best case, they provide an algorithmic framework by formulating the lower and upper bounds as linear programs. In our work, we extend this framework to compute fairness bounds for the majority vote.

Within the related PAC-Bayesian framework, fairness guarantees have been analyzed by Oneto et al. (2019).

While many approaches for fairness-aware methods have been proposed in recent years, few past works have considered the development of fair ensemble methods, such as fair random forests (Grari et al., 2020; Raff et al., 2018), fair boosting classifiers (Bhaskaruni et al., 2019; Iosifidis and Ntoutsi, 2019), or more general fair ensemble frameworks for arbitrary model classes (Alves et al., 2020; Iosifidis et al., 2019).

Grari et al. (2020) train a gradient tree boosting classifier while minimizing the ability of a neural network to predict the protected attribute from the classifiers' outputs. The weights that are assigned during gradient tree boosting are reduced if it tells us much about the protected attribute. Raff et al. (2018) introduce fair random forests by introducing a new GINI measure which discourages the selection of features that correlate with the target label or a protected attribute.

Iosifidis and Ntoutsi (2019) have developed an in-processing fairness approach by altering the weight updates in the classical Adaboost. Only empirically, their new method *Adafair* can improve the fairness with respect to equal odds. Similarly, Bhaskaruni et al. (2019) propose an Adaboost variant where the weighting considers only fairness, which is measured for each data point in a local k–neighborhood. In experiments, this comes at a high cost of accuracy.

## 4.4 Setup and Notation for Ensembles

In this section, we introduce the ensemble setting and recall important notation and fairness notions that we presented in Chapter 1. Let $\mathcal{X}$ be a feature space, $\mathcal{Y} = \{-1, 1\}$ the binary label space, and $\mathcal{S} = \{-1, 1\}$ the binary protected attribute. We consider an ensemble classifier $H : \mathcal{X} \to \mathcal{Y}$, also called strong classifier, and base learners $h_j : \mathcal{X} \to \mathcal{Y}$ with $j \in [T]$, where $[T]$ denotes the set $\{1, \ldots, T\}$. Assume that we can draw examples $(X, S, Y) \sim \mathcal{D}_\mathcal{Z}$ from a distribution $\mathcal{D}_\mathcal{Z}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$. In order to build a strong classifier, we fuse the base learner's output for an input point $x$ with the majority vote:

$$H(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{T} h_i(x) \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{4.1}$$

Therefore, if at least $\lceil T/2 \rceil$ base learners output the positive label, the overall prediction is positive.

**Recap: Fair Binary Classification.** We focus on the statistical group fairness notions *demographic parity* and *equal opportunity*. If the predictions of a given classifier $h$ are independent of the protected attribute, the classifier $h : \mathcal{X} \to \mathcal{Y}$ is fair with respect to *demographic parity*. More formally, it holds that

$$\mathop{\mathbb{P}}_{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}} [h(X) = 1 | S = 1] = \mathop{\mathbb{P}}_{(X,S,Y)\sim\mathcal{D}_\mathcal{Z}} [h(X) = 1 | S = -1] \, .$$

We measure the violation of demographic parity with the *difference of demographic parity*:

$$\mathrm{DDP}(h) = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=1] - \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|S=-1]\,.$$

A classifier $h$ is fair with respect to *equal opportunity* if its predictions of positive examples are independent of the protected attribute, that is if

$$\underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|Y=1, S=1] = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|Y=1, S=-1]\,.$$

Again, we measure the fairness violation with the *difference of equal opportunity*:

$$\mathrm{DEO}(h) = \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|Y=1, S=1] - \underset{(X,S,Y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[h(X)=1|Y=1, S=-1]\,.$$

## 4.5 Ensembles of Conditionally Independent Fair Classifiers are Fair

The Condorcet Jury Theorem in its original form (Grofman et al., 1983) assumes that every member of the jury makes a correct decision with probability $p$ and that the number of jury members is odd. Most importantly, the jury members are assumed to vote independently. Condorcet shows that the majority vote is more accurate when more members are added to the jury (the ensemble). The results have been generalized to any number of jury members or heterogeneous performance (Lam and Suen, 1997; Shapley and Grofman, 1984).

In the following, we say that the classifier outputs are independent (Kuncheva, 2014, p. 113), if for any subset of classifiers $A \subset [T]$ with $A = \{i_1, \ldots, i_K\}$ the joint probability can be factorized as

$$\mathbb{P}\left(h_{i_1}(X) = \hat{y}_{i_1}, \ldots, h_{i_K}(X) = \hat{y}_{i_K}\right) = \mathbb{P}\left(h_{i_1}(X) = \hat{y}_{i_1}\right) \cdot \ldots \cdot \mathbb{P}\left(h_{i_K}(X) = \hat{y}_{i_K}\right). \qquad (4.2)$$

Under these assumptions, we can directly compute the ensemble accuracy. With the original assumptions of the Jury Theorem—$\mathbb{P}\left(h_1(X) = Y\right) = \ldots = \mathbb{P}\left(h_T(X) = Y\right) = p$ and $T$ odd— we can compute

$$\mathbb{P}\left(H(X) = Y\right) = \sum_{k=\frac{T+1}{2}}^{T} \binom{T}{k} p^k (1-p)^{T-k}. \qquad (4.3)$$

Condorcet's Jury Theorem further states that, if $p > 0.5$, we have $\mathbb{P}\left(H(X) = Y\right) \to 1$ as $T \to \infty$.

We analyze the majority vote with respect to its fairness by making similarly strong assumptions on the independence of the base learners. In the following proposition, we consider a set of demographic–parity–fair base learners. We need the assumption of Condorcet **and joint independence conditioned on** $S$.

**Proposition 1.** *Given a set of classifiers $h_1, ..., h_T$. Assume that the classifier outputs are independent in the sense of* (4.2) *and jointly independent conditioned on $S$. If every classifier is fair with respect to demographic parity, then the majority vote ensemble $H$ is demographic-parity-fair.*

*Proof.* Similar to the proof of Condorcet's Jury Theorem, we can express the positive rate of the majority vote on group $s$ in terms of the positive rates of each base learner using the joint independence conditional on $S$:

$$\begin{aligned}
\mathbb{P}\left(H(X) = 1 \mid S = s\right) = \; & \mathbb{P}\left(\sum_{i=1}^{T} h_i(X) \geq 0 \mid S = s\right) \\
= \; & \sum_{k=\lceil T/2 \rceil}^{T} \mathbb{P}\left(\sum_{i=1}^{T} h_i(X) = k \mid S = s\right) \\
= \; & \sum_{k=\lceil T/2 \rceil}^{T} \sum_{\substack{K \subseteq [T] \\ k=|K|}} \mathbb{P}\left(h_i(X)=1 \; \forall i \in K, h_j(X)=-1 \; \forall j \in \overline{K} \mid S=s\right) \\
= \; & \sum_{k=\lceil T/2 \rceil}^{T} \sum_{\substack{K \subseteq [T] \\ k=|K|}} \prod_{i \in K} \mathbb{P}\left(h_i(X)=1 \mid S=s\right) \prod_{j \in \overline{K}} 1 - \mathbb{P}\left(h_j(X)=1 \mid S=s\right).
\end{aligned}$$

Since the base learners are fair with respect to demographic parity, we have

$$\mathbb{P}\left(h_i(X) = 1 \mid S = s\right) = \mathbb{P}\left(h_i(X) = 1\right)$$

for all $i \in [T]$. Using the joint independence of the base learners, we can 'reverse' the above steps and imply that $\mathbb{P}\left(H(X) = 1 \mid S = s\right) = \mathbb{P}\left(H(X) = 1\right)$, and thus, the strong classifier $H$ is demographic–parity–fair. □

We prove the same result for equality of opportunity by conditioning on $Y = 1$ since we only care about the positive rate on the positive class to fulfill equality of opportunity.

**Proposition 2.** *Given a set of classifiers $h_1, ..., h_T$. Assume that the classifier outputs are independent in the sense of* (4.2) *conditioned on $Y = 1$ and jointly independent conditioned on $S$ and $Y = 1$. If every classifier is fair with respect to equality of opportunity, then the majority vote ensemble $H$ is also fair with respect to equality of opportunity.*

The conclusions of Propositions 1 and 2 follow almost directly from the strong assumptions we make. The base learners (a) predict the positive label independently of the protected attribute because they are fair, (b) act independently of each other, and (c) act independently of each other when the protected attribute is known. Taken together these strong conditions are sufficient. It is not sufficient if we only assume that

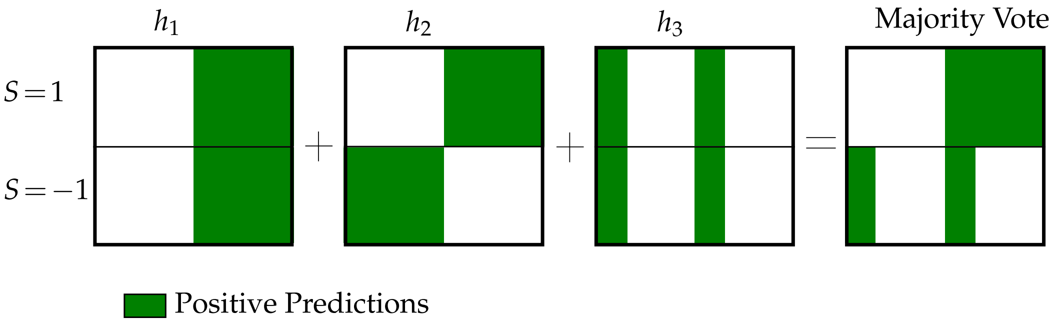


Figure 4.2: **Majority vote of three independent and fair classifiers.** Similar to Figure 4.1, we have three fair classifiers with positive rates $0.5, 0.5, 0.3$. We can confirm that the classifier outputs are independent, for example in the middle of the top half we have a voting pattern with $\mathbb{P}(h_1 = 1, h_2 = 1, h_3 = 1) = 0.075$, which also equals $0.5^2\dot{0}.3$. However, the majority vote is unfair with a DDP $= 0.5 - 0.3 = 0.2$. With respect to Proposition 1, we see that the conditional independence assumption is not fulfilled since $h_1$ and $h_2$ are not independent on group $S = 1$.

the base learners are independent (without conditioning on $S$). Since we cannot imply $\mathbb{P}\left(h_i(X) = \hat{y}_i, h_j(X) = \hat{y}_j \mid S = s\right) = \mathbb{P}\left(h_i(X) = \hat{y}_i \mid S = s\right) \mathbb{P}\left(h_j(X) = \hat{y}_j \mid S = s\right)$ if $\mathbb{P}\left(h_i(X) = \hat{y}_i, h_j(X) = \hat{y}_j\right) = \mathbb{P}\left(h_i(X) = \hat{y}_i\right) \mathbb{P}\left(h_j(X) = \hat{y}_j\right)$ for two classifiers $h_i, h_j$, the base learners could be dependent once the protected attribute is known and we can use this to construct an unfair majority vote. In Figure 4.2, we construct three independent and fair classifiers similar to Example 4.2. The majority vote is unfair, since the classifiers are not independent on each group. Similarly, we can construct an example where the majority vote is fair, but the base learners are neither independent nor conditionally independent.

At the end of this section, a remark is in order about the fact that the strong independence assumptions are not realistic. We cannot expect the base learners to give independent outputs in practice. Typically, the classifiers are trained on the same dataset instead of collecting new random samples from the data distribution. Although bagging attempts to create independent classifiers by training each base learner on a bootstrap sample of the train data, this does not guarantee that the classifiers give independent outputs. With the goal of a good ensemble performance, independence is not even always desirable since certain patterns of dependencies result in more accurate ensembles (Kuncheva et al., 2003). One line of work investigates several measures of *diversity* of ensembles and the question how diverse ensembles need to be for a good overall performance (Krogh and Vedelsby, 1994; Kuncheva and Whitaker, 2003; Tang et al., 2006).

Due to the unrealistic independence assumptions that we identified as sufficient conditions for fairness, we will drop any independence assumption in the next section. In two example, we have seen that majority votes can be unfair, but we do not know yet how unfair the majority vote can be in the worst case.

## 4.6 Worst-case Fairness Bounds

In Figures 4.1 and 4.2 we have seen that the ensemble fairness can deteriorate, even though fair classifiers are combined, since they lack the strong independence assumptions from the previous section. In this section we make **no independence assumptions** at all and we investigate what the worst fairness violation in a majority vote could be if the fair base learners depend on each other arbitrarily. We provide algorithmic worst-case fairness bounds as the solutions to different linear programs.

Depending on what we know from each base learner, we can consider different cases. In Example 4.2, we know the acceptance rates, the accuracy, and the fact that each recruiter is fair. In order to understand the notation, we will start off with an easier case and then move on to more general notation to address the example.

We will first familiarize the reader with the main idea and the notation needed.

**Notation.** The majority vote ensemble has a finite number of possible voting patterns $2^T$. We fix the order of the base learners $h_1, \dots, h_T$ according to the indices $1, \dots, T$. Using this order we first encode a specific voting pattern in a binary number, for example $\mathbf{110}_2$ corresponds to $h_1(X) = \mathbf{1}, h_2(X) = \mathbf{1}$, and $h_3(X) = \mathbf{0}$ reading the binary number from left to right. Second, we enumerate all possible voting patterns of $T$ classifiers with an index in $\{0, \dots 2^T - 1\}$. A given index $i \in \{0, \dots 2^T - 1\}$ uniquely determines the voting pattern since we convert the index $i$ into the corresponding voting pattern with the operator $\text{bin}(i, T)$, which converts an integer into a binary number with $T$ bits (Narasimhamurthy, 2003).

The algorithmic bounds are achieved by "moving around" the probability mass of all possible voting patterns—this corresponds to moving around the green areas in Figure 4.1. We introduce a variable $\boldsymbol{x} = [x_0, ..., x_{2^T-1}]^\intercal$, where $x_i \in [0, 1]$ corresponds to the probability of the voting pattern $\text{bin}(i, T)$. By the law of total probability the probabilities of all voting patterns sum to one:

$$\begin{aligned} 1 &= \sum_{\hat{y}_1, ..., \hat{y}_T \in \{-1,1\}} \mathbb{P}\left(h_1(X) = \hat{y}_1, ..., h_T(X) = \hat{y}_T\right) \\ &= \sum_{i=0}^{2^T-1} x_i = \mathbf{1}^\intercal \boldsymbol{x}. \end{aligned}$$

Since the vector $\boldsymbol{x}$ includes the probabilities of all possible voting patterns, we can use $\boldsymbol{x}$ to evaluate the probability of meaningful voting patterns, for instance voting patterns where the majority of base learners votes positive. The probability of a positive prediction by the majority vote ensemble is then given by the sum of the probabilities of all these voting patterns. Using our binary encoding, we define a constant vector $\boldsymbol{c}^{\text{maj}} \in \{0, 1\}^{2^T}$, where for all $i \in \{0, \dots 2^T - 1\}$ we have

$$(\boldsymbol{c}_{\text{maj}})_i = \begin{cases} 1 & \text{number of 1's in bin}(i, T) \text{ is at least } \lceil T/2 \rceil, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

With $\boldsymbol{c}_{\text{maj}}$ we have $\mathbb{P}\left(H(X) = 1\right) = \boldsymbol{c}_{\text{maj}}^{\intercal}\boldsymbol{x}$. Similarly, we express the positive rate of a single base learner by summing up all voting patterns, where the base learner votes positive. For every base learner $h_j$ with $j \in [T]$, we define a constant vector $\boldsymbol{a}_j \in \{0,1\}^{2^T}$ such that

$$(\boldsymbol{a}_j)_i = \begin{cases} 1 & j\text{'th digit in } \text{bin}(i,T) \text{ is } 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

Then, we can compute the positive rate of $h_j$ with $\mathbb{P}\left(h_j(X) = 1\right) = \boldsymbol{a}_j^{\intercal}\boldsymbol{x}$.

**Minimal and maximal positive rate.** Let us use the established notation in a first example to find the minimal and maximal positive rate of the majority vote ensemble $H$ when we are given the positive rates $p_i := \mathbb{P}\left(h_i = 1\right)$ of the base learners for all $i \in [T]$.[2] With the voting patterns $A = [\boldsymbol{a_1}, ..., \boldsymbol{a_T}]^{\intercal}$ and the positive rates $\boldsymbol{p} = [p_1, ..., p_T]$, we need to constrain $\boldsymbol{x}$ such that $A\boldsymbol{x} = \boldsymbol{p}$. We maximize or minimize the positive rate $\boldsymbol{c}_{\text{maj}}^{\intercal}\boldsymbol{x}$ with a linear program which reads

$$\begin{aligned} \min/\max \quad & \boldsymbol{c}_{\text{maj}}^{\intercal}\boldsymbol{x} \\ & 0 \leq \boldsymbol{x} \leq 1 \\ & \boldsymbol{1}^{\intercal}\boldsymbol{x} = 1 \\ & A\boldsymbol{x} = \boldsymbol{p}. \end{aligned} \tag{4.6}$$

In Figure 4.3, left panel, we assume that $p = p_i$ for all $i \in [T]$ and we plot the minimal and maximal positive rate as a function of $p$. As $p$ approaches one, that is when all classifiers constantly predict 1, the minimal and maximal positive rate of the majority vote go to one as well. On the other hand, as $p$ approaches zero, the positive rate of the majority vote goes to zero. The largest gap between minimal and maximal positive rate is achieved when $p$ is around 0.5 and $T$ is sufficiently large.

In the following two cases, we extend the established notation to answer the original research question: What is the maximal and minimal unfairness of the majority vote classifier $H$ given certain pieces of information about the base learners?

### 4.6.1 Case 1: Given Fair Base Learners and their Positive Rates.

In the first case, we are given the positive rates of demographic–parity–fair classifiers. Due to demographic parity, the positive rate of each base learner is the same for each group; we have $\mathbb{P}\left(h_i = 1 \mid S = s\right) = \mathbb{P}\left(h_i = 1\right) = p_i$ for all base learners $h_i$. Since we do not know the accuracy of the base learners (for simplicity), we can optimize the positive rates on each group separately. Hence, the minimally (maximally) achievable

[2] Note that this setting is equivalent to minimizing/maximizing the accuracy of the strong classifier when we encode a 'positive' prediction as a correct prediction and the positive rate of a base learner as the accuracy of the base learner (Narasimhamurthy, 2003).

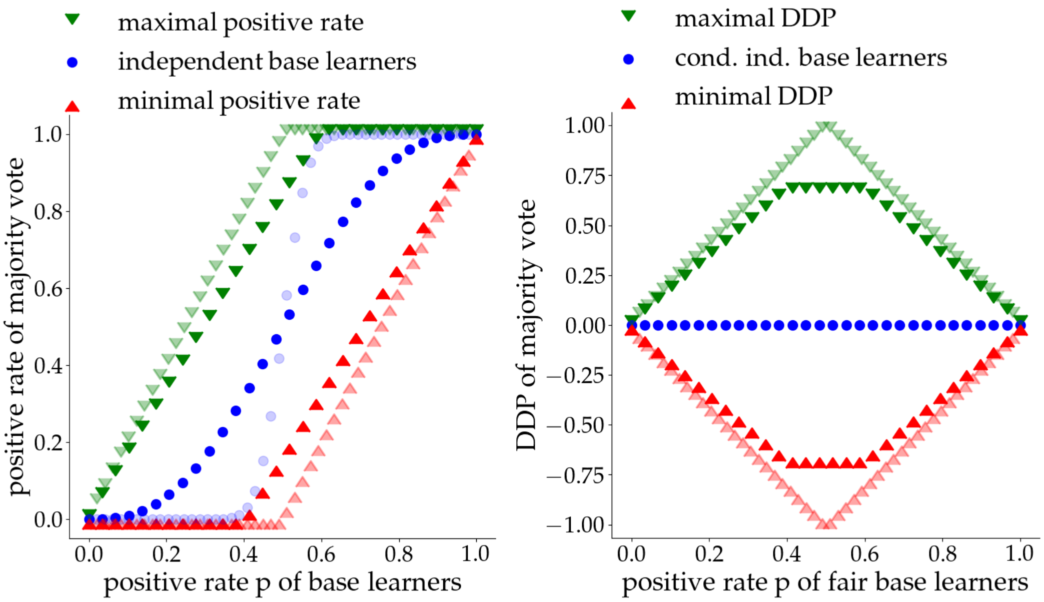


Figure 4.3: (Left) **Bounds on the positive rate.** Using the LP 4.6 we plot the minimal and maximal positive rate of the majority vote given 5 homogeneous base learners (transparent: 101 base learners) with the same positive rate $p$. We also compute the majority vote positive rate with Equation 4.3 when the base learners are independent. According to Condorcet's Jury Theorem adding more base learners increases the overall positive rate if $p > 0.5$ (the transparent blue curve has a steeper slope). (Right) **Bounds on DDP.** Given 5 (or 101 in transparent) classifiers that are **fair with respect to demographic parity**, we can use the bounds on the positive rate from the left plot and compute the extreme DDP values. In other words, for a fixed $p$, the vertical distance between the green and red curve in the left plot is the largest DDP that the majority vote can achieve.

positive rate of the ensemble is the same for both groups and we need to minimize and maximize (4.6) only once.

We denote the minimal and maximal solution of (4.6) with $\pi^{\min}$ and $\pi^{\max}$, respectively. Following the arguments above, note that $\pi^{\min}$ denotes the minimal positive rate that the majority vote can achieve on both group $s = 1$ and group $-1$, and overall. Importantly, the majority vote can assume the minimal positive rate on one group and the maximal positive rate on the other.

We can now maximize (minimize) the fairness measure $\mathrm{DDP}(H)$ of the ensemble by maximizing the positive rate on one group and minimizing it on the other group; recall $\mathrm{DDP}(H) = \mathbb{P}(H(X) = 1 \mid S = 1) - \mathbb{P}(H(X) = 1 \mid S = -1)$. Using the argument above that $\pi^{\min}$ and $\pi^{\max}$ are the extreme values for both groups, we can bound the ensemble $\mathrm{DDP}(H)$:

$$\pi^{\min} - \pi^{\max} \leq \mathrm{DDP}(H) \leq \pi^{\max} - \pi^{\min}.$$

In Figure 4.3, we plot the above upper and lower bound for the setting where each base learner has the same positive rate $p = p_i$ for all $i \in [T]$. As $p$ approaches 1 the bounds approach 0 (fairness!) from above and below—the majority vote is necessarily fair when all base learners only predict the positive label. Similarly, the majority is fair if the base learners never predict the positive label at all. The largest fairness violations can be achieved when $p$ is around 0.5.

### 4.6.2 Case 2: Given Fair Base Learners and their Positive Rates and Accuracy.

We can now extend our notation to address the setting of Example 4.2, where we are given the positive rates and the accuracy of each fair base learner. The accuracy information requires us to formulate a more general version of LP (4.6), since we cannot arbitrarily distribute positive predictions anymore. We need to consider the true labels and the fact that changing predictions in one group, changes the accuracy overall. To this end, we extend our notation and introduce more variables.

- We split the probabilities $\boldsymbol{x} \in [0,1]^{2^T}$ of all voting patterns into 4 different cases depending on the class label and the protected group. The vector $\boldsymbol{\alpha}^{(s)} \in [0,1]^{2^T}$ corresponds to the joint probabilities of all voting patterns, the event $Y = 1$ and $S = s$, that is $\boldsymbol{\alpha}_i^{(s)} = \mathbb{P}(\text{voting pattern is bin}(i,T), Y = 1, S = s)$. Similarly, $\boldsymbol{\beta}^{(s)} \in [0,1]^{2^T}$ represents the joint probabilities of all voting patterns, the event $Y = 0$ and $S = s$. Respecting the law of total probability, we have $\boldsymbol{\alpha}^{(1)} + \boldsymbol{\alpha}^{(-1)} + \boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(-1)} = \boldsymbol{x}$.

- The vector $\boldsymbol{p} = [p_1, ..., p_T]$ contains the positive rates $p_j = \mathbb{P}(h_j(X) = 1)$. If the base learners are demographic–parity–fair these rates are also the group-wise positive rates.

- The vector $\boldsymbol{q} = [q_1, ..., q_T]$ contains the accuracies $q_j = \mathbb{P}(h_j(X) = Y)$ for all $j \in [T]$.

- We define $B = 1 - A$ to indicate all voting patterns (indexed by the columns) where a base learner (indexed by row) predicts the negative label.

We are only interested in a worst-case voting behavior of the base learners, so we need to make sure that the underlying statistics of the dataset are not changed as a means to to fulfill the accuracy constraints. In our more general LP, we fix the group prevalences $w_s = \mathbb{P}[S = s]$ and the joint probabilities $\mathbb{P}(Y = 1, S = s)$ in order to determine the group sizes and the ratio of a positive label in each group.

Finally, we obtain the following optimization problem.

$$\begin{aligned}
&\max/\min \boldsymbol{c}_{\text{maj}}^{\mathsf{T}} \left( \frac{1}{w_1} \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\beta}^{(1)} \right) - \frac{1}{w_{-1}} \left( \boldsymbol{\alpha}^{(-1)} + \boldsymbol{\beta}^{(-1)} \right) \right) && (4.7)\\
&\quad 0 \le \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(-1)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(-1)} \le 1 && \text{(need probabilities)}\\
&\quad \mathbf{1}^{\mathsf{T}} \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\alpha}^{(-1)} + \boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(-1)} \right) = 1 &&\\
&\quad \mathbf{1}^{\mathsf{T}} \boldsymbol{\alpha}^{(1)} = \mathbb{P}\left( Y = 1, S = 1 \right) && \text{(dataset constraints)}\\
&\quad \mathbf{1}^{\mathsf{T}} \boldsymbol{\alpha}^{(-1)} = \mathbb{P}\left( Y = 1, S = -1 \right) &&\\
&\quad \mathbf{1}^{\mathsf{T}} \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\beta}^{(1)} \right) = \mathbb{P}\left( S = 1 \right) &&\\
&\quad A \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\alpha}^{(-1)} + \boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(-1)} \right) = \boldsymbol{p} && \text{(positive rates)}\\
&\quad A \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\alpha}^{(-1)} \right) + B \left( \boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(-1)} \right) = \boldsymbol{q} && \text{(accuracies)}\\
&\quad A \left( \frac{1}{w_1} \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\beta}^{(1)} \right) - \frac{1}{w_{-1}} \left( \boldsymbol{\alpha}^{(-1)} + \boldsymbol{\beta}^{(-1)} \right) \right) = \mathbf{0}. && \text{(fairness)}
\end{aligned}$$

The above optimization problem outputs the extreme DDP values for all possible ensembles. Typically, we are interested in ensembles that perform at least as good as the best base learner. We can compute the DDP bounds for an ensemble with fixed accuracy $a \in [0, 1]$ by adding the following constraint to the LP:

$$\boldsymbol{c}_{\text{maj}}^{\mathsf{T}} \left( \boldsymbol{\alpha}^{(1)} + \boldsymbol{\alpha}^{(-1)} \right) + \left( 1 - \boldsymbol{c}_{\text{maj}}^{\mathsf{T}} \right) \left( \boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(-1)} \right) = a. \qquad \text{(ensemble accuracy)}$$

In general, the above LP formulation provides great flexibility. Depending on which information we assume certain constraints can be removed or new constraints can be added. For example, we might be given only the accuracy of each base learner and the fact that they are fair, but we drop the constraint that fixes the positive rates. Instead of demographic parity as the optimization objective we can optimize any kind of fairness measure that can be expressed linearly in terms of the probability vectors, for example equal odds and equal accuracy. The flexibility of this approach is due to the 'brute force' decomposition into all possible voting patterns. However, this comes at a computational cost since the number of constraints grows exponentially with the number of classifiers $T$.

## 4.7 Experiments

In this section, we explore and summarize the main insights that optimization problem 4.7 provides. To this end, we fix the ensemble accuracy to a value $a \in [0, 1]$. We minimize and maximize optimization problem 4.7 for every fixed ensemble accuracy $a$ in order to determine the minimal DDP (DEO) and the maximal DDP (DEO). Then, we can plot the fairness bounds as a function of the ensemble accuracy. Recall that a

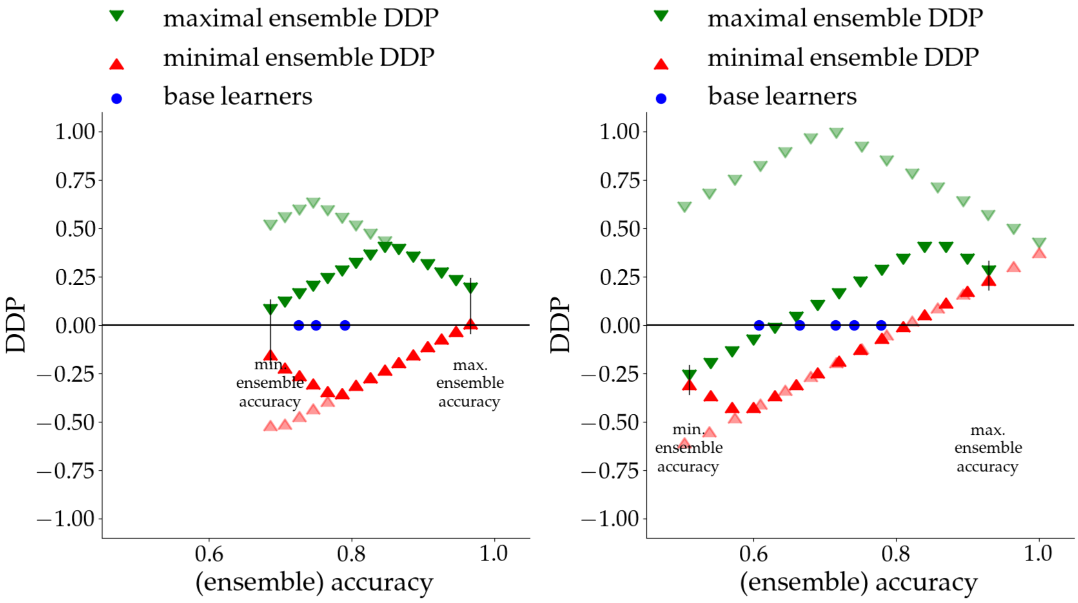


Figure 4.4: **Bounds on ensemble DDP.** (Left) Given the accuracy and positive rates of three fair base learners from Example 4.2, we can compute the minimal and maximal DDP for any fixed ensemble accuracy. The ensemble accuracy is bounded by the left and right vertical lines as determined by a variation of LP (4.7). In this example, demographic–parity–fair majority vote ensembles are possible for any ensemble accuracy. As the ensemble accuracy increases the lower bound approaches zero, while the upper bound is around 0.25. Hence, **the ensemble can at best be fair or it disadvantages group $S = -1$.** (Right) For different dataset statistics, we sample a set of five fair classifiers. If we want the ensemble to be more accurate than the best base learner, **the majority vote cannot be fair** since the lower bound moves above DDP $= 0$. The transparent lines correspond to the fairness bounds, when we do not assume fair base learners. The green upper bound, and thus the advantage for group $S = 1$, is significantly reduced. **Fair base learners can provide better worst-case fairness bounds.**

positive DDP signifies that group $S = 1$ is advantaged and a negative DDP that group $S = -1$ is advantaged. When the DDP is zero, the ensemble is fair.

**Dataset Statistics.** The linear program is controlled by constant dataset statistics. We fix the group prevalence $w_1 = \mathbb{P}(S = 1)$ and the group–wise positive rates $p_{1,1} = \mathbb{P}(Y = 1 \mid S = 1)$ and $p_{1,-1} = \mathbb{P}(Y = 1 \mid S = -1)$. In the motivational example, we have assumed $p_{1,1} = 0.5$, $p_{1,-1} = 0.4$, and $w_1 = 0.5$, and we use three classifiers (the recruiters) with positive rates $\boldsymbol{p} = [0.3, 0.5, 0.25]$ and accuracies $\boldsymbol{q} = [0.79, 0.725, 0.75]$. We provide bounds on the DDP and DEO of the example and an additional set of dataset statistics.

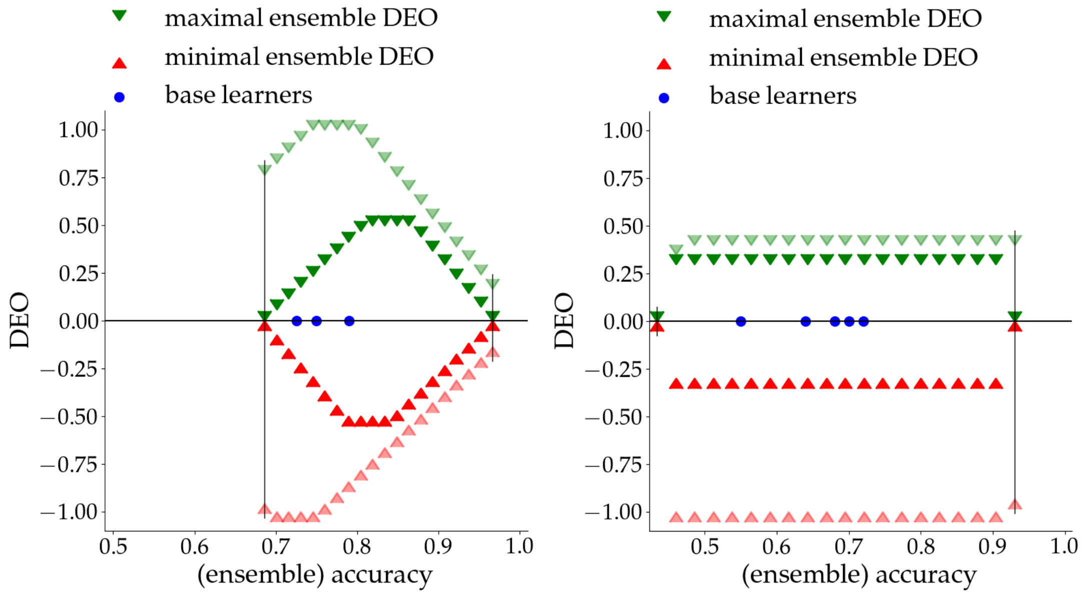


Figure 4.5: **Bounds on ensemble DEO.** (Left) Given the accuracies and positive rates from Example 4.2, we compute bounds on the DEO. When the base learners are fair (non-transparent) we can guarantee a better fairness compared to when they are not fair (transparent). In particular, **the ensemble with maximally feasible accuracy is guaranteed to be fair** when it is composed of fair base learners. (Right) In a second example, we are given five base learners with different accuracies and positive rates. The worst-case DEO is almost constant over all possible ensemble accuracies. **The lower bound can be improved significantly if the base learners are fair**, this is we can guarantee a smaller disadvantage of group $S = 1$ by more than 50%. Again, the maximally accurate ensemble is fair, when the base learners are fair.

**Bounds on Demographic Parity.** The left panel of Figure 4.4 depicts the DDP bounds of the Motivational Example 4.2. All points between the upper bound, the lower bound, the minimal ensemble accuracy on the left, and the maximal ensemble on the right, are possible outcomes of the majority vote. In this case, for any ensemble accuracy a DDP–fair majority vote is possible, but, more importantly, the worst-case bounds are such that **the majority vote can exhibit distinct unfairness** (a DDP of 0.25 means that the positive rate of group $S = 1$ is larger by 25%!).

As we increase the ensemble accuracy, the bounds on the DDP improve. Finally, maximal accuracy can only be achieved when **the majority vote is either fair or unfair to group** $S = -1$. This is due to the unbalanced dataset and the fairness notion, which does not necessarily allow the perfect classifier—in this case, the perfect classifier has $DDP = 0.1$. Consequently, the higher the expected ensemble accuracy is, the more will the data statistics govern the bound (this can be seen more strongly in the right panel).

In the right panel, we fix the dataset statistics to $p_{1,1} = 0.5$, $p_{1,-1} = 0.1$, and $w_1 = 0.5$. We compute the bounds for five randomly chosen base learners with $\boldsymbol{q} = [0.78, 0.66, 0.71, 0.61, 0.74]$ and $\boldsymbol{p} = [0.44, 0.62, 0.53, 0.63, 0.52]$. In this example, **any ensemble with higher accuracy than the best base learner has to be unfair** since the lower bound on the DDP moves above zero when the ensemble accuracy is larger than the best base learner's accuracy.

In both plots, we also compute fairness bounds when we remove the fairness constraint in (4.7) while every input and every constraint remains the same. Now,the base learners can be arbitrarily unfair. By removing this constraint, the worst-case bounds (necessarily) become worse in both panels (see the transparent bounds). In other words, if we know that the base learners are fair, **we can guarantee fairer ensembles** in absolute terms. In the right panel, we reduce in particular the possible advantage of group $S = 1$. We want to emphasize at this point that we can only improve *the guarantees*, but we cannot imply from these bounds that a majority vote of fair classifiers is in general more fair than a majority vote of unfair classifiers.

**Bounds on Equality of Opportunity.** In Figure 4.5 we plot bounds on equality of opportunity assuming both fair and unfair base learners. To this end, we change the objective function of optimization problem 4.7 accordingly. In the left panel, we consider our motivational example. First, we observe that a higher accuracy ensemble has better fairness guarantees. In this case, we even have that the **ensemble with maximal accuracy is fair** because the upper and lower bound coincide at DEO $= 0$. Second, we again observe that the knowledge about fair base learners improves the worst-case bounds compared to base learners with arbitrary fairness (in transparent). We even lose the perfect fairness guarantee for the ensemble with maximal accuracy.

In the right panel, we have $p_{1,1} = 0.1$, $p_{1,-1} = 0.1$, and $w_1 = 0.75$, with $\boldsymbol{q} = [0.55, 0.68, 0.70, 0.72, 0.64]$ and $\boldsymbol{p} = [0.35, 0.36, 0.38, 0.2, 0.3]$. In this example, the fairness bounds are constant, but the worst-case lower bound on the DEO can be reduced by more than 50% if the base learners are fair. If the fair base learners attain the most accurate ensemble, it is guaranteed to be fair.

## 4.8 Conclusion

In this chapter, we provided worst-case bounds of popular fairness measures for majority vote ensembles. In particular, we investigated the worst-case fairness guarantees of the majority vote, when it is composed of fair base learners. We found that (1) fair base learners can improve the worst-case fairness bounds, but that (2) a fair and accurate majority vote is sometimes impossible to achieve. Even though the worst-case fairness bounds can provide a certain fairness guarantee, they are often too large to guarantee a (nearly) fair majority vote. Without any further assumptions on the dependence between the base learners, we did not observe a fairness of crowds for the setting, where each individual learner is fair.

This work is a preliminary analysis of the fairness properties in ensembles and it opens up interesting future research questions. In order to analyze the properties of the majority vote, Kuncheva et al. (2003) have identified specific dependencies between the base learners, called the ’pattern of success’ and the ’pattern of failure’, under which the majority vote improves or deteriorates accuracy. Since we found that the majority vote does not attain the fairness properties of its base learners, it might require a *pattern of fairness*. Based on the result in Proposition 1, a pattern of fairness could state that any deviation from independence, like the pattern of success, must occur in both protected groups to the same degree. However, actually incorporating such a dependence during the parallel training of base learners is challenging.

Fair ensembles might also be achieved by a different aggregation scheme instead of the majority vote. Our fairness bounds showed that any fairness considerations during the training of the base learner can be in vain due the cancellation effects. As a consequence, no effort should be put into training fair base learners, but instead, in post-processing with fairness constraints. Incorporating fairness during training might only be useful for ensembles where classifiers are trained dependently, such as in Adaboost.

# Chapter 5

# Bibliography

T. Adel, I. Valera, Z. Ghahramani, and A. Weller. One-network adversarial fairness. In *AAAI Conference on Artificial Intelligence*, 2017.

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

AGG. Allgemeines Gleichbehandlungsgesetz vom 14. August 2006 (BGBl. I S. 1897), das zuletzt durch Artikel 8 des Gesetzes vom 3. April 2013 (BGBl. I S. 610) geändert worden ist. 2006. URL https://www.gesetze-im-internet.de/agg/index.html.

V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACV)*, 2020.

A. Altman. Discrimination. In *The Stanford Encyclopedia of Philosophy*. Winter 2020 edition, 2020.

G. Alves, V. Bhargava, M. Couceiro, and A. Napoli. Making ml models fairer through explanations: the case of limeout. In *International Conference on Analysis of Images, Social Networks and Texts (AIST)*, 2020.

M. Alvi, A. Zisserman, and C. Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision (ECCV) - Workshop*, 2019.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias - propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

P. S. Arcidiacono. Expert report by Peter S. Arcidiacono, Students for Fair Admissions. 2019. https://tinyurl.com/2p9cw7pp.

G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona. Towards causal benchmarking of bias in face analysis algorithms. In *European Conference on Computer Vision (ECCV)*, 2020.

M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *Conference on Learning Theory (COLT)*, 2008.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671–732, 2016.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. 2019. http://www.fairmlbook.org.

R. Battiti and A. M. Colla. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 1994.

Y. Bechavod and K. Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

H. Bendekgey and E. B. Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

J. R. Bent. Is algorithmic affirmative action legal? *Geo. LJ*, 108:803–853, 2019.

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H.-h. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

D. Bhaskaruni, H. Hu, and C. Lan. Improving prediction fairness via model ensemble. In *IEEE Conference on Tools with Artificial Intelligence (ICTAI)*, 2019.

D. Biddle. Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing. 2005.

R. Binns. Fairness in machine learning: Lessons from political philosophy. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2018.

L. Breiman. Bagging predictors. *Machine learning*, 1996.

L. Breiman. Random forests. *Machine learning*, 2001.

BT Drucksache 19/23700. Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. 2020. https://dserver.bundestag.de/btd/19/237/1923700.pdf.

BT Drucksache 19/2978. Einsetzung einer Enquete-Kommission “Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale”. 2018. https://dserver.bundestag.de/btd/19/029/1902978.pdf.

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

K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, 2018.

T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. In *International Conference on Data Mining and Knowledge Discovery (DMKD)*, 2010.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops (ICDM)*, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2016.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A. Cotter, H. Jiang, and K. Sridharan. Two-player games for efficient non-convex constrained optimization. In *International Conference on Algorithmic Learning Theory (ALT)*, 2019.

E. Denton, B. Hutchinson, M. Mitchell, T. Gebru, and A. Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*, 2019.

W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. In *IEEE Intelligent Systems*, 2020.

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

C. Dwork and C. Ilvento. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

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

H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016.

J. Elstrodt. *Maß-und Integrationstheorie*. Springer, 1996.

Equal Employment Opportunity Commission (EEO). Uniform guidelines on employee selection procedures, 29 cfr part 1607. 1978. https://www.uniformguidelines.com/uniform-guidelines.html.

V. Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., 2018.

S. Fazelpour and Z. C. Lipton. Algorithmic fairness from a non-ideal perspective. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

T. Feathers. Proctorio is using racist algorithms to detect faces, 2021. https://www.vice.com/en/article/g5gxg3/proctorio-is-using-racist-algorithms-to-detect-faces.

M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

T. Feldman and A. Peake. End-to-end bias mitigation: Removing gender bias in deep learning. *arXiv preprint arXiv:2104.02532*, 2021.

R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341*, 2019.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001.

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Science and Engineering (DSE)*, 2020.

N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller. On fairness, diversity and randomness in algorithmic decision making. *arXiv preprint arXiv:1706.10208*, 2017.

N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI Conference on Artificial Intelligence*, 2018.

B. Grofman, G. Owen, and S. L. Feld. Thirteen theorems in search of the truth. *Theory and decision*, 1983.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Z. Harned and H. Wallach. Stretching human laws to apply to machines: The dangers of a" colorblind" computer. *Fla. St. UL Rev.*, 47:617–648, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

D. Hellman. Measuring algorithmic fairness. *Va. L. Rev.*, 106:811–866, 2020.

A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019.

V. Iosifidis and E. Ntoutsi. Adafair: Cumulative fairness adaptive boosting. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

V. Iosifidis, B. Fetahu, and E. Ntoutsi. Fae: A fairness-aware ensemble framework. In *IEEE International Conference on Big Data (IEEE Big Data)*, 2019.

S. Jia, T. Lansdall-Welfare, and N. Cristianini. Right for the right reason: Training agnostic networks. In *International Symposium on Intelligent Data Analysis (IDA)*, 2018.

S. Jia, T. Meng, J. Zhao, and K.-W. Chang. Mitigating gender bias amplification in distribution by posterior regularization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Y. Joshi. Racial indirection. *UC Davis L. Rev.*, 52:2495–2568, 2018.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems (KAIS)*, 2012.

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

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2012.

K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

M. Kearns and A. Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Inc., 2019.

M. Kendall. The treatment of ties in ranking problem. *Biometrika*, 1945.

J. S. Kim, J. Chen, and A. Talwalkar. FACT: A diagnostic for group fairness trade-offs. In *In International Conference on Machine Learning (ICML)*, 2020.

A. King and A. Hemenway. Blurred lines: Disparate impact and disparate treatment challenges to subjective decisions-the case of reductions in force. *Wm. & Mary Bus. L. Rev.*, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 2012.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

M. Kleindessner, S. Samadi, M. B. Zafar, K. Kenthapadi, and C. Russell. Pairwise fairness for ordinal regression. *arXiv preprint arXiv:2105.03153*, 2021.

R. Kohavi and B. Becker. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid, 1996.

A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1994.

J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165:633–706, 2017.

L. I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 2003.

L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 2003.

L. Lam and S. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 1997.

J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

R. Lidell and M. O'Flaherty. Handbook on european non-discrimination law – 2018 edition. 2018. https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition.

Z. C. Lipton, A. Chouldechova, and J. McAuley. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *In International Conference on Machine Learning (ICML)*, 2018.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

M. Lohaus, P. Hennig, and U. von Luxburg. Uncertainty estimates for ordinal embeddings. *arXiv preprint arXiv:1906.11655*, 2019. URL http://arxiv.org/abs/1906.11655.

M. Lohaus, M. Perrot, and U. von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning (ICML)*, 2020.

M. Lohaus, M. Kleindessner, K. Kenthapadi, F. Locatello, and C. Russell. Are two heads the same as one? identifying disparate treatment in fair neural networks, 2022. URL https://arxiv.org/abs/2204.04440.

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

D. Madras, E. Creager, T. Pitassi, and R. S. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*, 2018.

P. Manisha and S. Gujar. Fnnc: Achieving fairness through neural networks. In *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, 2020.

O. Matan. On voting ensembles of classifiers. In *AAAI workshop on integrating multiple learned models*, 1996.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 2021.

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

D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. Invariant representations without adversarial training. *arXiv preprint arXiv:1805.09458*, 2018.

A. M. Narasimhamurthy. A framework for the analysis of majority voting. In *Scandinavian Conference on Image Analysis*, 2003.

S. U. Noble. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, 2018.

C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

L. Oneto, M. Donini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2019.

V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau. Fair bayesian optimization. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

M. O. R. Prates, P. H. C. Avelar, and L. Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 2020.

E. Raff and J. Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.

E. Raff, J. Sylvester, and S. Mills. Fair forests: Regularized tree induction to minimize model bias. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

J. Rawls. *A theory of justice*. Harvard University Press, 1996.

J. Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

M. A. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 2002.

L. Risser, A. G. Sanz, Q. Vincenot, and J.-M. Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *arXiv preprint arXiv:1908.05783*, 2021.

J. E. Roemer. *Theories of distributive justice*. Harvard University Press, 1996.

J. E. Roemer. *Equality of opportunity*. Harvard University Press, 1998.

O. Sagi and L. Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018.

M. J. Sandel. *Justice: What's the Right Thing to Do?* Farrar, Straus and Giroux, 2009.

R. E. Schapire. The strength of weak learnability. *Machine learning*, 1990.

R. E. Schapire. A brief introduction to boosting. In *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, 1999.

M. Selmi. Indirect discrimination and the anti-discrimination mandate. *Philosophical foundations of discrimination law*, pages 250–268, 2013.

I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*, 2020.

L. Shapley and B. Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 1984.

C. Song and V. Shmatikov. Overlearning reveals sensitive attributes. In *International Conference on Learning Representations (ICLR)*, 2020.

F. Statute. Amendment of Civil Rights Act, Pub. L. No. 102-166, sec. 106, 105 Stat. 1071, 1075 (codified at 42 U.S.C. sec 2000e-2(l)). 1991.

Supreme Court. Hunt v. Cromartie. 526 U.S. 541, 1999.

Supreme Court. Ricci v. DeStefano. 557 U.S. 557, 2009.

J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

E. K. Tang, P. N. Suganthan, and X. Yao. An analysis of diversity measures. *Machine learning*, 2006.

Troupe v. May Dept. Stores. *(7th Cir.)*, pages 20 F.3d 734, 736, 1994.

B. Ustun, Y. Liu, and D. Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning (ICML)*, 2019.

L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008.

L. C. Vankadara, M. Lohaus, S. Haghiri, F. U. Wahab, and U. von Luxburg. Insights into ordinal embedding algorithms: A systematic evaluation. *arXiv preprint arXiv:1912.01666*, 2021. URL http://arxiv.org/abs/1912.01666.

J. Vincent. Canon put ai cameras in its chinese offices that only let smiling workers inside, 2021. https://www.theverge.com/2021/6/17/22538160/ai-camera-smile-recognition-office-workers-china-canon.

S. Wachter, B. Mittelstadt, and C. Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123: 735–790, 2020.

S. Wachter, B. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 2021.

A. Wang and O. Russakovsky. Directional bias amplification. In *International Conference on Machine Learning (ICML)*, 2021.

M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019a.

T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019b.

Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

M. Wick, S. Panda, and J.-B. Tristan. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

D. H. Wolpert. Stacked generalization. *Neural networks*, 1992.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, 2017.

Y. Wu, L. Zhang, and X. Wu. On convexity and bounds of fairness-aware classification. In *International World Wide Web Conference (WWW)*, 2019.

Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. *arXiv preprint arXiv:1705.11122*, 2017.

H. P. Young. *Equity: In Theory and Practice*. Princeton University Press, 1995.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017a.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web (WWW)*, 2017b.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, 2013.

H. Zhao and G. J. Gordon. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

D. Zietlow, M. Lohaus, G. Balakrishnan, M. Kleindessner, F. Locatello, B. Schölkopf, and C. Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *IEEE International Conference on Data Mining (ICDM)*, 2011.

# Appendix A

# Too Relaxed To Be Fair

In this appendix, we present detailed results from the experiments in Chapter 2 of all 6 datasets in Figures A.1– A.6. The overall trend of the results is the same as described in the chapter and we discuss the results of each dataset in the corresponding figure. Lastly, we present the results of the experiments under two different hyperparameters selection methods in Figures A.7 and A.8. For all results, each experiment has been repeated 10 times and we report the average and standard deviation of classification error and absolute fairness scores DDP and DEO. Furthermore, we report the value of the Donini linear relaxation (Section 2.2.1) on the test set to show the discrepancy between the true and relaxed fairness (this metric was omitted in Chapter 2).

**General setup.** For all the methods except Cotter, as the set of functions $\mathcal{F}$, we use the similarity-based classifiers that were presented in Chapter 2. As similarities, we consider both the linear and the rbf kernel. As reasonable points, we use a random subset of 70% (at most 1000) of the training examples. As regularization term we use a squared $\ell_2$ norm (which is a strongly convex function). The loss function in the empirical risk is the hinge loss, that is

$$\ell(f(x), y) = \max(0, 1 - f(x)\,y).$$

For the linear version of Cotter et al. (2019), we use the approach suggested in their example on the Adult dataset. We use a single-layer neural network where the input size is the number of features. The parameters are then learned using the RATEMINIMIZATIONPROBLEM provided by the package TENSORFLOW-CONSTRAINED-OPTIMIZATION. In order to use more complex classifiers based on the rbf kernel, we precompute the kernel matrix between the training points and the reasonable points. Then, the input size of the single-layer neural network is set to the number of reasonable points. For both linear and complex classifiers, no further regularization is used. However, to obtain reasonable and stable results, the number of epochs has to be carefully chosen. We use between 1000 and 5000 epochs depending on the dataset, and for the minibatch size we use the default of 200 points.

**Compas–Figure A.4.** The Compas dataset (Larson et al., 2016) contains 7214 points with 53 features, such as name, age, degree of crime, and number of prior crimes. We use the same pre-processing as Zafar et al. (2017b) and, in particular, we select the same 5 features. The goal is to predict if a defendant has been arrested again within two years of the decision. The protected attribute is race. It has been changed to a binary attribute with the values 'White' and 'NonWhite'. We use $5,000$ randomly selected points for training.

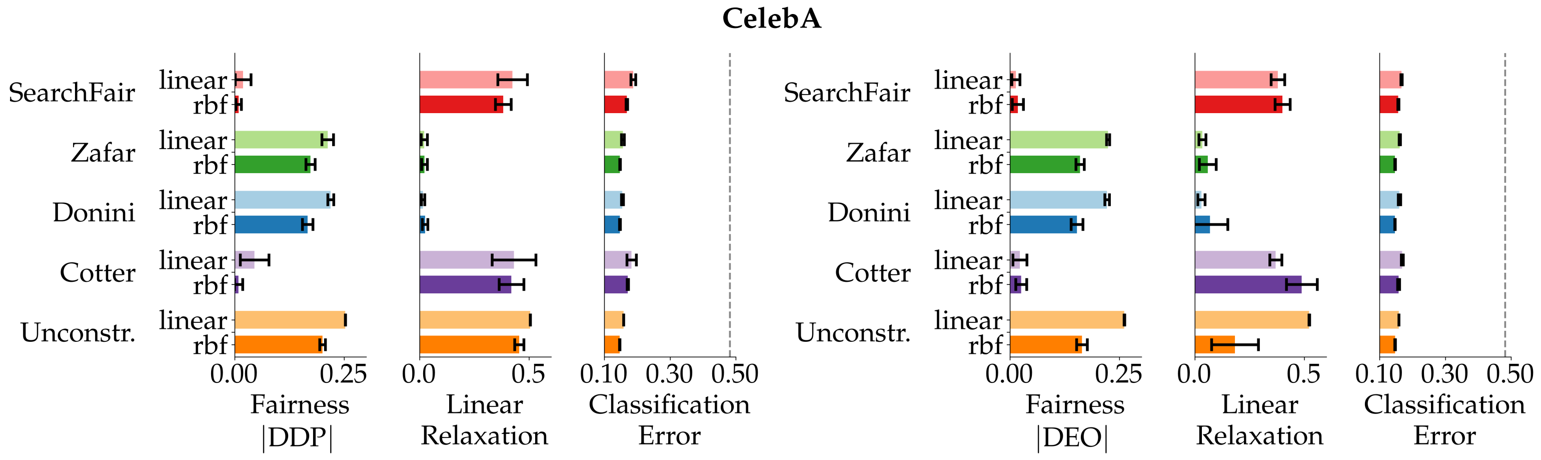
**Communities and Crime–Figure A.5.** This dataset includes socio-economic data of 1994 communities in the United States (Redmond and Baveja, 2002). It consists of 128 attributes, of which we drop the name of the state, county, and community, and features with missing values. Overall, we drop 29 features. We use the attribute RACEPCTWHITE to construct a binary protected attribute. A community with a percentage of white residents higher than the mean 0.75 obtains the protected label 1, otherwise the label is $-1$. The goal of this data set is to predict the number of violent crimes. We binarize the label by splitting VIOLENTCRIMESPERPOP at the mean of 0.24. We use $1,500$ randomly selected points for training.

**German Credit–Figure A.6.** There are 1000 records of german applicants for a credit with 20 attributes (Dua and Graff, 2017). The goal is to classify the applicants in creditworthy or not creditworthy. The categorical feature 'personal status' is changed into the binary feature sex. We use it as the protected attribute and use the other 19 features for training. We use 700 randomly selected points for training.

**Toy dataset–Figure 2.1.** The toy dataset set in Figure 2.1 consists of 600 points (for the sake of readability, we only plot a random subset of 400 examples). We draw the points from different Gaussian distributions. For the protected attribute (the dots), we sample 150 points with negative label from a Gaussian with mean $\mu_1 = [2, -2]$ and covariance matrix $\Sigma_1 = [[1, 0], [0, 1]]$, and another 150 points for the positive class from the mixture of two Gaussians, with $\mu_2 = [3, -1]$ and $\Sigma_2 = [[1, 0], [0, 1]]$ and $\mu_3 = [1, 4]$ and $\Sigma_3 = [[0.5, 0], [0, 0.5]]$. For the unprotected attribute (the crosses), we draw 150 points with positive label from a Gaussian with $\mu_4 = [2.5, 2.5]$ and $\Sigma_4 = [[1, 0], [0, 1]]$, and 150 points with negative label from a Gaussian with $\mu_5 = [4.5, -1.5]$ and $\Sigma_5 = [[1, 0], [0, 1]]$.

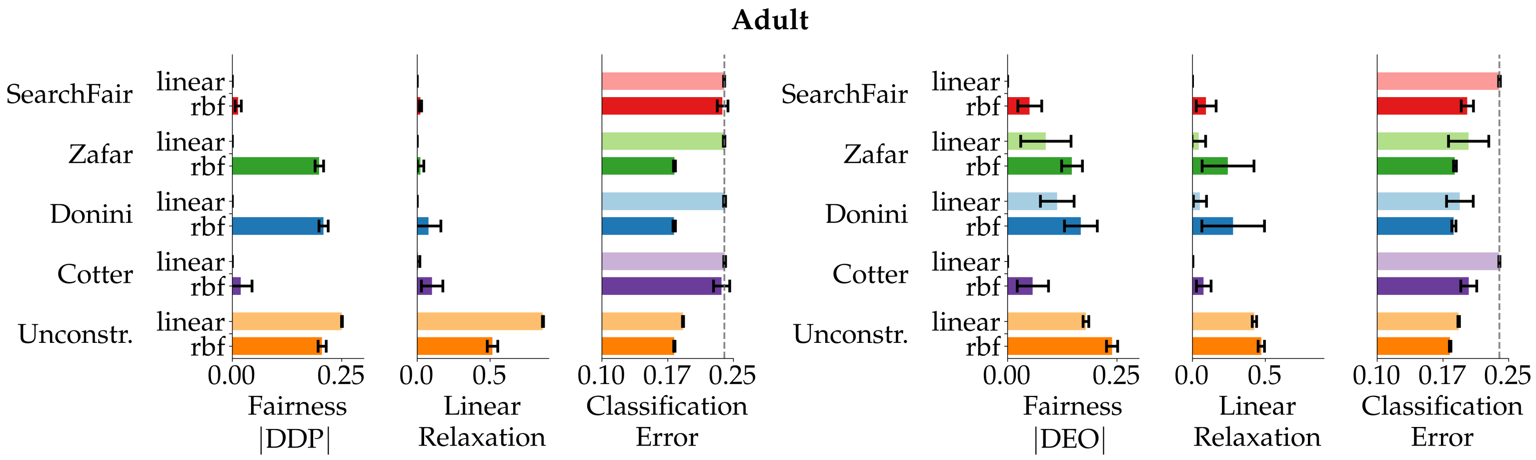
**Cross Validation Procedure.** We report the results for different cross validation procedures as discussed Chapter 2. In Figure A.7 we use a procedure called NVP proposed by Donini et al. (2018). In a first step, we exclude the hyperparameters with an accuracy score that is lower than 90% of the best accuracy score. Then, we choose the set of hyperparameters with the best average fairness score. Finally, we use these hyperparameters to train on the whole train set.

In Figure A.8 we report the results when we use a given fairness threshold. We shortlist all hyperparameters with an absolute fairness score lower than 0.05 and, among them, choose the hyperparameters with the highest accuracy score. We report average and standard deviation of classification error and absolute fairness scores DDP and DEO over 10 repetitions. Note that we also report results for the approach by Cotter et al. (2019) for comparison, even though the linear version does not tune any hyperparameters. Using the rbf kernel on the other hand, we need to tune the width of the kernel.



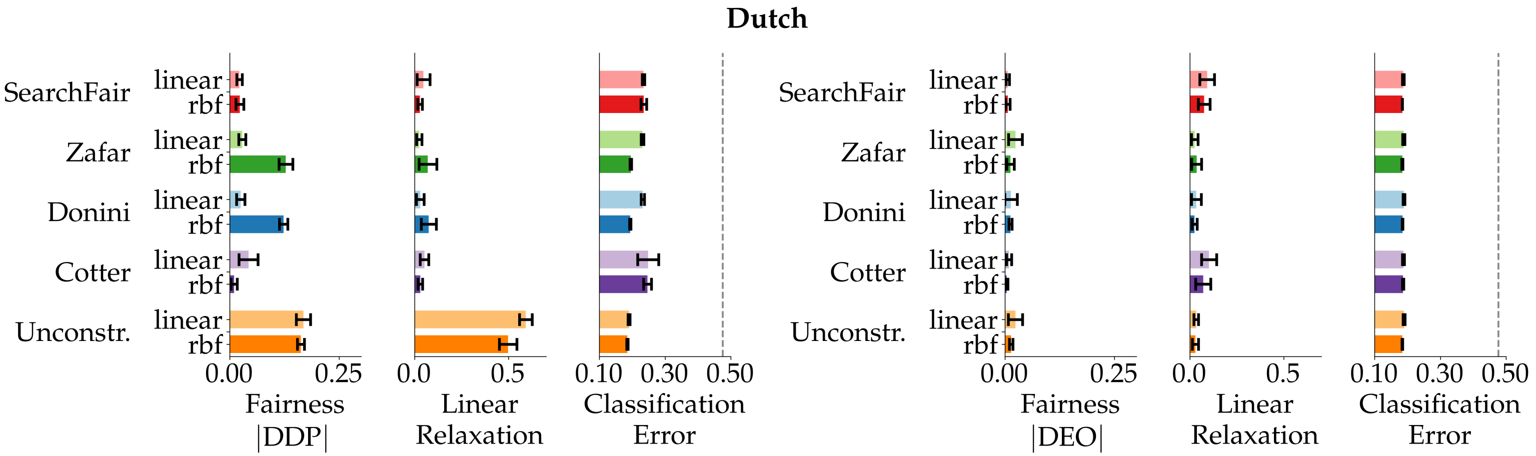
| FAIRNESS NOTION | | Demographic Parity | | | Equality of Opportunity | | |
|---|---|---|---|---|---|---|---|
| | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.02 ± 0.02 | 0.42 ± 0.07 | 0.19 ± 0.01 | 0.01 ± 0.01 | 0.38 ± 0.01 | 0.17 ± 0.00 |
| | rbf | 0.01 ± 0.01 | 0.38 ± 0.04 | 0.17 ± 0.00 | 0.02 ± 0.01 | 0.40 ± 0.04 | 0.16 ± 0.00 |
| Zafar | linear | 0.21 ± 0.01 | 0.02 ± 0.01 | 0.16 ± 0.00 | 0.22 ± 0.00 | 0.04 ± 0.02 | 0.16 ± 0.00 |
| | rbf | 0.17 ± 0.01 | 0.02 ± 0.01 | 0.15 ± 0.00 | 0.16 ± 0.01 | 0.06 ± 0.04 | 0.15 ± 0.00 |
| Donini | linear | 0.22 ± 0.01 | 0.02 ± 0.01 | 0.15 ± 0.00 | 0.22 ± 0.01 | 0.03 ± 0.02 | 0.16 ± 0.00 |
| | rbf | 0.17 ± 0.01 | 0.03 ± 0.01 | 0.15 ± 0.00 | 0.15 ± 0.01 | 0.07 ± 0.08 | 0.15 ± 0.00 |
| Cotter | linear | 0.05 ± 0.03 | 0.43 ± 0.10 | 0.18 ± 0.01 | 0.02 ± 0.02 | 0.37 ± 0.03 | 0.17 ± 0.00 |
| | rbf | 0.01 ± 0.01 | 0.42 ± 0.06 | 0.18 ± 0.00 | 0.03 ± 0.01 | 0.49 ± 0.07 | 0.16 ± 0.00 |
| Unconstrained | linear | 0.25 ± 0.00 | 0.51 ± 0.00 | 0.16 ± 0.00 | 0.26 ± 0.00 | 0.52 ± 0.00 | 0.16 ± 0.00 |
| | rbf | 0.20 ± 0.01 | 0.46 ± 0.02 | 0.15 ± 0.00 | 0.16 ± 0.01 | 0.18 ± 0.11 | 0.15 ± 0.00 |
| Constant | – | 0.00 ± 0.00 | – | 0.48 ± 0.00 | 0.00 ± 0.00 | – | 0.48 ± 0.00 |

Figure A.1: **CelebA.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. Overall, SearchFair is the only method with both low DDP and DEO, while linear Cotter is only slightly worse for DDP. Additionally, SearchFair and Cotter exhibit high values for the linear relaxations which might imply that this relaxation is not suitable here. This is confirmed by the fact that the competing methods have low relaxation values with high DDP and DEO values.



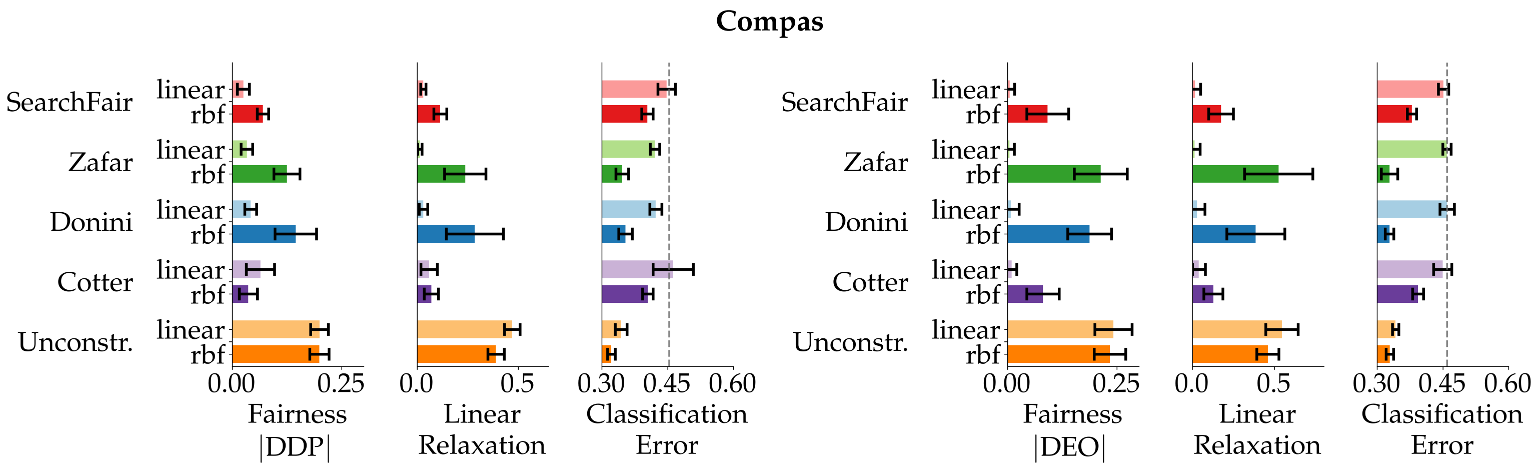
| FAIRNESS NOTION | | Demographic Parity | | | Equality of Opportunity | | |
|---|---|---|---|---|---|---|---|
| | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.24 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.24 ± 0.00 |
| | rbf | 0.01 ± 0.01 | 0.02 ± 0.01 | 0.24 ± 0.01 | 0.05 ± 0.03 | 0.09 ± 0.07 | 0.20 ± 0.01 |
| Zafar | linear | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.24 ± 0.00 | 0.09 ± 0.06 | 0.04 ± 0.05 | 0.20 ± 0.02 |
| | rbf | 0.20 ± 0.01 | 0.02 ± 0.02 | 0.18 ± 0.00 | 0.15 ± 0.02 | 0.24 ± 0.18 | 0.19 ± 0.00 |
| Donini | linear | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.24 ± 0.00 | 0.11 ± 0.04 | 0.05 ± 0.05 | 0.19 ± 0.02 |
| | rbf | 0.21 ± 0.01 | 0.08 ± 0.08 | 0.18 ± 0.00 | 0.17 ± 0.04 | 0.28 ± 0.21 | 0.19 ± 0.00 |
| Cotter | linear | 0.00 ± 0.00 | 0.01 ± 0.01 | 0.24 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.24 ± 0.00 |
| | rbf | 0.02 ± 0.03 | 0.10 ± 0.07 | 0.24 ± 0.01 | 0.06 ± 0.04 | 0.08 ± 0.05 | 0.20 ± 0.01 |
| Unconstrained | linear | 0.25 ± 0.00 | 0.86 ± 0.00 | 0.19 ± 0.00 | 0.18 ± 0.01 | 0.43 ± 0.02 | 0.19 ± 0.00 |
| | rbf | 0.21 ± 0.01 | 0.52 ± 0.04 | 0.18 ± 0.00 | 0.24 ± 0.01 | 0.47 ± 0.02 | 0.18 ± 0.00 |
| Constant | – | 0.00 ± 0.00 | – | 0.24 ± 0.00 | 0.00 ± 0.00 | – | 0.24 ± 0.00 |

Figure A.2: **Adult.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. All the methods tend to learn the constant classifier to obtain a DDP fair model with the linear kernel. With the rbf kernel, SearchFair and Cotter obtain low fairness scores (both for DDP and DEO) showing that the fairness of the model learned by the relaxation based baselines can be heavily linked to the complexity of the models. Note that, even though all the fairness methods learn classifiers with a low linear relaxation, their DDP scores vary widely. It confirms that there is no guarantee that a low relaxation value will lead to a fair classifier.



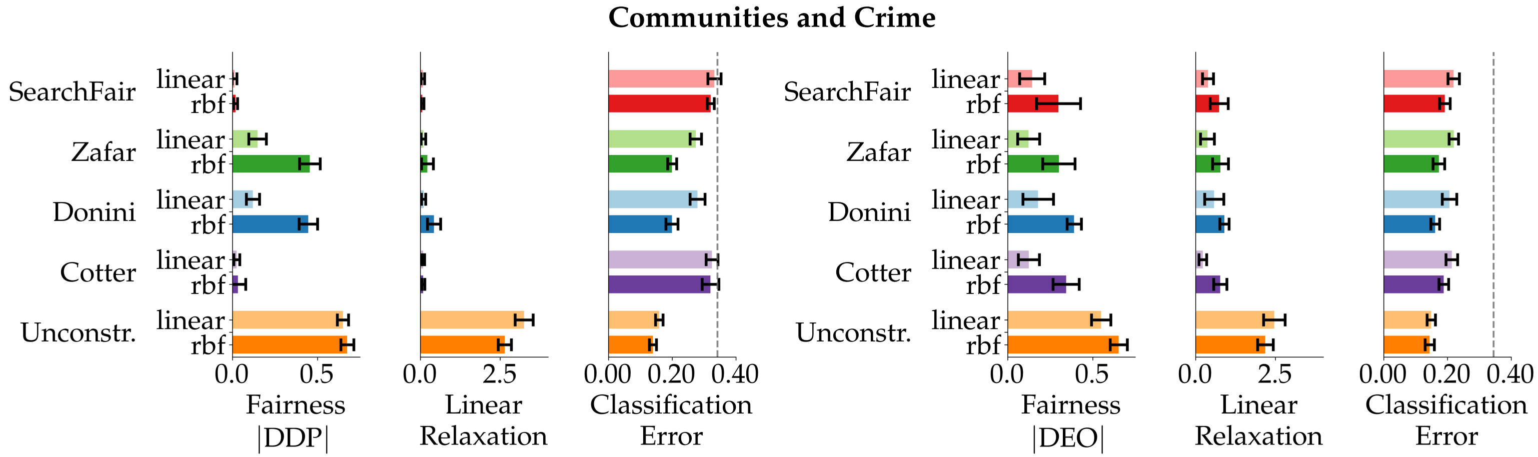
| FAIRNESS NOTION | | Demographic Parity | | | Equality of Opportunity | | |
|---|---|---|---|---|---|---|---|
| | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.02 ± 0.01 | 0.05 ± 0.03 | 0.23 ± 0.00 | 0.01 ± 0.00 | 0.09 ± 0.04 | 0.19 ± 0.00 |
| | rbf | 0.02 ± 0.01 | 0.03 ± 0.01 | 0.23 ± 0.01 | 0.01 ± 0.00 | 0.08 ± 0.03 | 0.18 ± 0.00 |
| Zafar | linear | 0.03 ± 0.01 | 0.03 ± 0.01 | 0.23 ± 0.00 | 0.02 ± 0.02 | 0.03 ± 0.02 | 0.19 ± 0.00 |
| | rbf | 0.13 ± 0.02 | 0.07 ± 0.05 | 0.20 ± 0.00 | 0.01 ± 0.01 | 0.04 ± 0.03 | 0.18 ± 0.00 |
| Donini | linear | 0.03 ± 0.01 | 0.03 ± 0.02 | 0.23 ± 0.00 | 0.01 ± 0.01 | 0.03 ± 0.03 | 0.19 ± 0.00 |
| | rbf | 0.12 ± 0.01 | 0.08 ± 0.04 | 0.19 ± 0.00 | 0.01 ± 0.00 | 0.03 ± 0.01 | 0.18 ± 0.00 |
| Cotter | linear | 0.04 ± 0.02 | 0.05 ± 0.02 | 0.25 ± 0.03 | 0.01 ± 0.01 | 0.10 ± 0.04 | 0.19 ± 0.00 |
| | rbf | 0.01 ± 0.01 | 0.03 ± 0.01 | 0.25 ± 0.01 | 0.00 ± 0.00 | 0.07 ± 0.04 | 0.19 ± 0.00 |
| Unconstrained | linear | 0.17 ± 0.02 | 0.59 ± 0.03 | 0.19 ± 0.00 | 0.02 ± 0.02 | 0.03 ± 0.01 | 0.19 ± 0.00 |
| | rbf | 0.16 ± 0.01 | 0.50 ± 0.05 | 0.19 ± 0.00 | 0.01 ± 0.00 | 0.03 ± 0.02 | 0.18 ± 0.00 |
| Constant | – | 0.00 ± 0.00 | – | 0.48 ± 0.00 | 0.00 ± 0.00 | – | 0.48 ± 0.00 |

Figure A.3: **Dutch.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. In terms of DEO all the methods perform equally well on this dataset as the Unconstrained classifier is already DEO fair. On the other hand, SearchFair and Cotter obtain a low DDP regardless of the complexity of the model. Once again, even though all the fairness methods learn classifiers with a low linear relaxation, their DDP scores vary widely. It confirms that there is no guarantee that a low relaxation value will lead to a fair classifier.



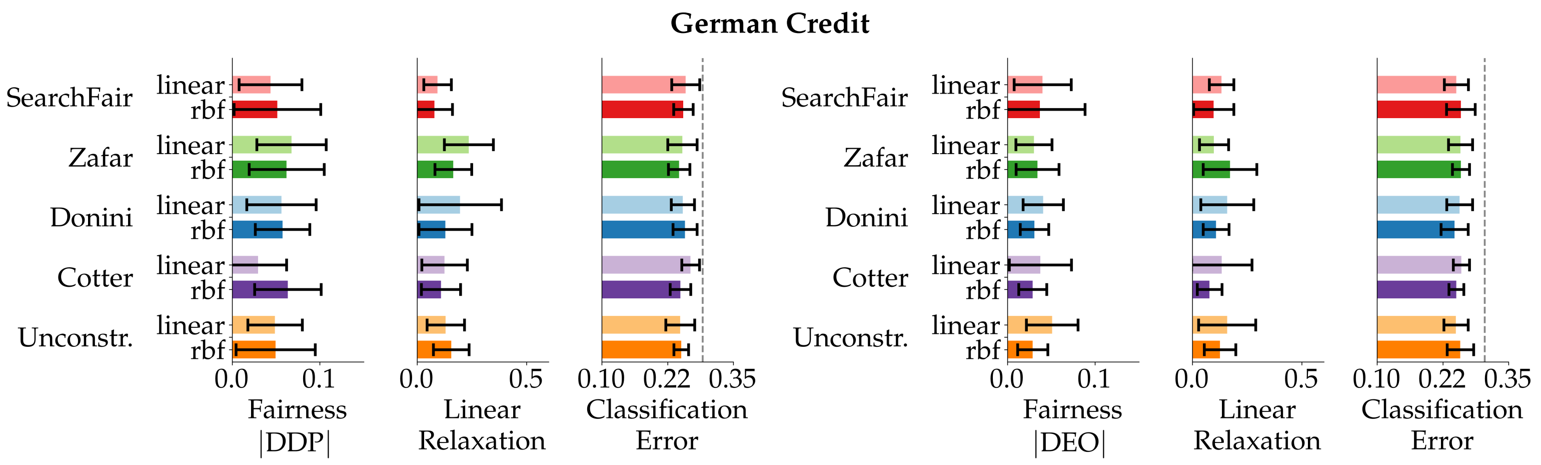
| FAIRNESS NOTION |  | Demographic Parity |  |  | Equality of Opportunity |  |  |
|---|---|---|---|---|---|---|---|
|  | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.03 ± 0.01 | 0.03 ± 0.01 | 0.45 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.03 | 0.45 ± 0.01 |
|  | rbf | 0.07 ± 0.01 | 0.11 ± 0.03 | 0.40 ± 0.01 | 0.09 ± 0.05 | 0.17 ± 0.08 | 0.38 ± 0.01 |
| Zafar | linear | 0.03 ± 0.01 | 0.01 ± 0.01 | 0.42 ± 0.01 | 0.00 ± 0.01 | 0.01 ± 0.03 | 0.46 ± 0.01 |
|  | rbf | 0.12 ± 0.03 | 0.24 ± 0.10 | 0.35 ± 0.01 | 0.21 ± 0.06 | 0.53 ± 0.21 | 0.33 ± 0.02 |
| Donini | linear | 0.04 ± 0.01 | 0.03 ± 0.02 | 0.42 ± 0.01 | 0.01 ± 0.02 | 0.03 ± 0.05 | 0.46 ± 0.02 |
|  | rbf | 0.14 ± 0.05 | 0.29 ± 0.14 | 0.35 ± 0.02 | 0.19 ± 0.05 | 0.39 ± 0.18 | 0.33 ± 0.01 |
| Cotter | linear | 0.06 ± 0.03 | 0.06 ± 0.04 | 0.46 ± 0.05 | 0.01 ± 0.01 | 0.04 ± 0.04 | 0.45 ± 0.02 |
|  | rbf | 0.04 ± 0.02 | 0.07 ± 0.04 | 0.40 ± 0.01 | 0.08 ± 0.04 | 0.13 ± 0.06 | 0.40 ± 0.01 |
| Unconstrained | linear | 0.20 ± 0.02 | 0.47 ± 0.04 | 0.34 ± 0.01 | 0.24 ± 0.04 | 0.55 ± 0.10 | 0.34 ± 0.01 |
|  | rbf | 0.20 ± 0.02 | 0.39 ± 0.04 | 0.32 ± 0.01 | 0.23 ± 0.04 | 0.46 ± 0.07 | 0.33 ± 0.01 |
| Constant | – | 0.00 ± 0.00 | – | 0.46 ± 0.01 | 0.00 ± 0.00 | – | 0.46 ± 0.01 |

Figure A.4: **Compas.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. On this dataset, Zafar and Donini with rbf kernel tend to have high values for the linear relaxation, which is probably due to an overfitting issue. Overall, SearchFair obtains good fairness scores, comparable to Cotter. For the DDP, SearchFair is slightly worse than Cotter with an rbf kernel, but better with a linear kernel. Surprisingly, both methods also have low relaxation values which hints that, on this dataset, this relaxation might be relevant if one could avoid overfitting.



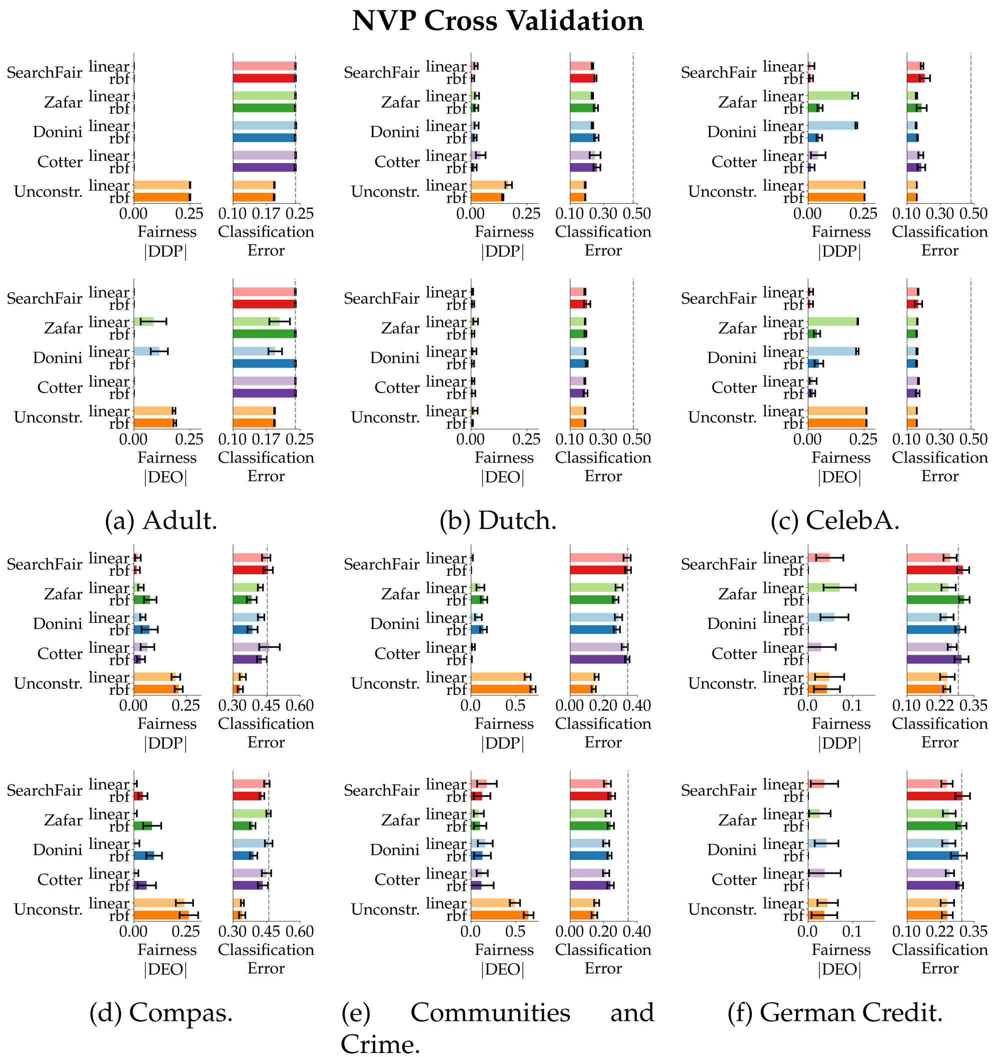
| FAIRNESS NOTION | | Demographic Parity | | | Equality of Opportunity | | |
|---|---|---|---|---|---|---|---|
| | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.01 ± 0.01 | 0.07 ± 0.05 | 0.33 ± 0.02 | 0.14 ± 0.07 | 0.38 ± 0.17 | 0.22 ± 0.02 |
| | rbf | 0.02 ± 0.01 | 0.06 ± 0.05 | 0.32 ± 0.01 | 0.30 ± 0.13 | 0.73 ± 0.28 | 0.19 ± 0.02 |
| Zafar | linear | 0.15 ± 0.05 | 0.09 ± 0.07 | 0.27 ± 0.02 | 0.12 ± 0.06 | 0.37 ± 0.22 | 0.22 ± 0.01 |
| | rbf | 0.45 ± 0.06 | 0.22 ± 0.19 | 0.20 ± 0.01 | 0.30 ± 0.10 | 0.78 ± 0.25 | 0.17 ± 0.02 |
| Donini | linear | 0.12 ± 0.04 | 0.10 ± 0.07 | 0.28 ± 0.02 | 0.18 ± 0.09 | 0.58 ± 0.30 | 0.21 ± 0.02 |
| | rbf | 0.45 ± 0.05 | 0.42 ± 0.21 | 0.20 ± 0.02 | 0.39 ± 0.04 | 0.90 ± 0.14 | 0.16 ± 0.01 |
| Cotter | linear | 0.02 ± 0.02 | 0.09 ± 0.04 | 0.32 ± 0.02 | 0.12 ± 0.06 | 0.22 ± 0.12 | 0.21 ± 0.02 |
| | rbf | 0.03 ± 0.05 | 0.09 ± 0.05 | 0.32 ± 0.03 | 0.34 ± 0.08 | 0.77 ± 0.21 | 0.19 ± 0.02 |
| Unconstrained | linear | 0.65 ± 0.03 | 3.25 ± 0.28 | 0.16 ± 0.01 | 0.55 ± 0.06 | 2.46 ± 0.34 | 0.15 ± 0.01 |
| | rbf | 0.67 ± 0.04 | 2.64 ± 0.21 | 0.14 ± 0.01 | 0.65 ± 0.05 | 2.19 ± 0.24 | 0.14 ± 0.01 |
| Constant | – | 0.00 ± 0.00 | – | 0.34 ± 0.01 | 0.00 ± 0.00 | – | 0.34 ± 0.01 |

Figure A.5: **Communities and Crime.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. Overall, all the fairness methods perform similarly well in terms of DEO. For DDP, only SearchFair and Cotter are able to learn a fair classifier for both the linear and rbf kernel. Once again, one can notice that a low linear relaxation might or might not imply a DDP fair classifier. Indeed, the DDP scores of Zafar, Donini, and SearchFair are very different while their linear relaxation scores are all close to 0.



| FAIRNESS NOTION | | Demographic Parity | | | Equality of Opportunity | | |
|---|---|---|---|---|---|---|---|
| | Kernel | \|DDP\| | \|LR\| | Error | \|DEO\| | \|LR\| | Error |
| SearchFair | linear | 0.04 ± 0.04 | 0.09 ± 0.06 | 0.26 ± 0.03 | 0.04 ± 0.03 | 0.13 ± 0.06 | 0.25 ± 0.02 |
| | rbf | 0.05 ± 0.05 | 0.08 ± 0.08 | 0.25 ± 0.02 | 0.04 ± 0.05 | 0.10 ± 0.09 | 0.26 ± 0.03 |
| Zafar | linear | 0.07 ± 0.04 | 0.24 ± 0.11 | 0.25 ± 0.03 | 0.03 ± 0.02 | 0.10 ± 0.07 | 0.26 ± 0.02 |
| | rbf | 0.06 ± 0.04 | 0.17 ± 0.08 | 0.25 ± 0.02 | 0.03 ± 0.02 | 0.17 ± 0.12 | 0.26 ± 0.02 |
| Donini | linear | 0.06 ± 0.04 | 0.20 ± 0.19 | 0.25 ± 0.02 | 0.04 ± 0.02 | 0.16 ± 0.12 | 0.26 ± 0.02 |
| | rbf | 0.06 ± 0.03 | 0.13 ± 0.12 | 0.26 ± 0.02 | 0.03 ± 0.02 | 0.11 ± 0.06 | 0.25 ± 0.03 |
| Cotter | linear | 0.03 ± 0.03 | 0.13 ± 0.10 | 0.27 ± 0.02 | 0.04 ± 0.04 | 0.13 ± 0.14 | 0.26 ± 0.02 |
| | rbf | 0.06 ± 0.04 | 0.11 ± 0.09 | 0.25 ± 0.02 | 0.03 ± 0.02 | 0.08 ± 0.06 | 0.25 ± 0.01 |
| Unconstrained | linear | 0.05 ± 0.03 | 0.13 ± 0.09 | 0.25 ± 0.03 | 0.05 ± 0.03 | 0.16 ± 0.13 | 0.25 ± 0.02 |
| | rbf | 0.05 ± 0.05 | 0.16 ± 0.08 | 0.25 ± 0.01 | 0.06 ± 0.03 | 0.13 ± 0.07 | 0.26 ± 0.03 |
| Constant | – | 0.00 ± 0.00 | – | 0.29 ± 0.02 | 0.00 ± 0.00 | – | 0.30 ± 0.02 |

Figure A.6: **German Credit.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. This is the smallest dataset out of the 6 with 700 training examples and 300 test examples. This explains the large standard deviations. For this particular dataset, SearchFair does not bring any significant improvement in terms of fairness compared to the baselines. We believe that it is due to a slight overfitting issue since the dataset is so small. Nevertheless, SearchFair is not worse that the other baselines as all the methods perform comparably.



(a) Adult.

(b) Dutch.

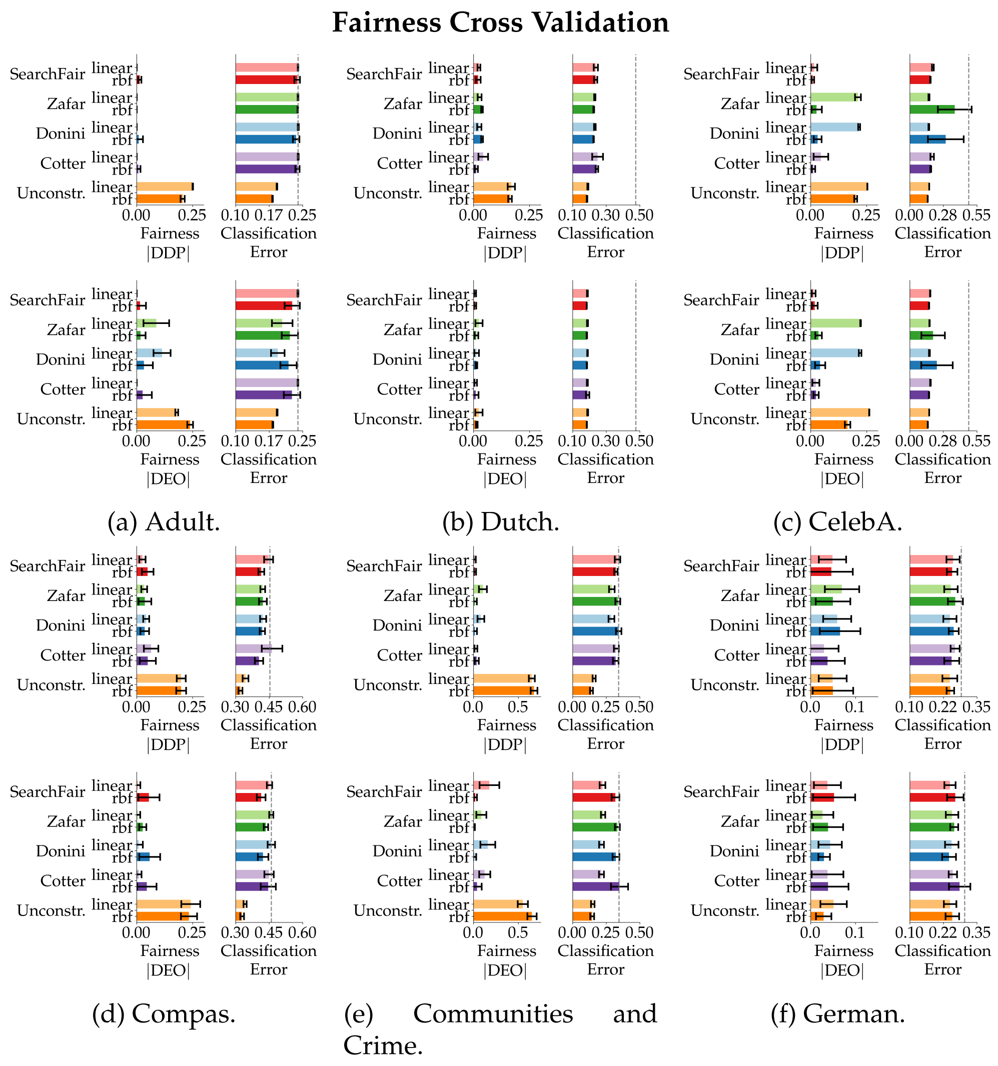
(c) CelebA.

(d) Compas.

(e) Communities and Crime.

(f) German Credit.

Figure A.7: We use a procedure called NVP (Donini et al., 2018), where we choose the set of hyperparameters with the best average fairness score while having an accuracy above a given threshold. Overall, using this procedure greatly improves the performances of the fairness baselines. Hence, on most datasets, they now obtain classifiers that are as fair as the ones learned by SearchFair and Cotter. Nevertheless, there is no guarantee that the method will succeed and it indeed fails for both DDP and DEO on CelebA (linear kernel), and for DEO on Adult (linear kernel). The fact that NVP succeeds for the rbf kernel and sometimes fails for the linear kernel hints that NVP is a good way to address the complexity issue of the linear relaxations but that it does not solve the other shortcommings. The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO.



(a) Adult.

(b) Dutch.

(c) CelebA.

(d) Compas.

(e) Communities and Crime.

(f) German.

Figure A.8: We use another cross validation procedure, where we shortlist all hyperparameters with an absolute fairness score lower than 0.05 and, among them, choose the hyperparameters with the highest accuracy score. The results are very similar to the ones of NVP presented in Figure A.7 and the same conclusions can be drawn. In particular, it seems to solve the complexity issue of linear relaxations with rbf kernel but can still fail when using the linear kernel (for both DDP and DEO on CelebA, and for DEO on Adult). The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO.

# Appendix B

# Disparate Treatment

In this appendix, we provide more detailed results that we omitted from main chapter. In Section B.1, we complement the results on protected attribute awareness in fair networks. In Section B.2, we apply our explicit approach to MobileNetV3-Small and compare to all other fairness approaches. Finally, in Section B.3 we extend our equivalence results to different target tasks, models, and fairness approaches.

## B.1 Protected Attribute Awareness

In this section, we report further results on protected attribute awareness in fair neural networks. We plot last-layer tSNE visualizations for another CelebA task in Figure B.1 and for a FairFace task in Figure B.1. Similar to Figure 3.1 in Chapter 3, gender is separated into two clusters when we regularize the model for demographic parity.

In Figure B.3, we plot the Kendall-tau correlations when using MobileNetV3-Small for both of the two presented regularizers. As with a ResNet50 model, we observe a strong association between fairness parameter and an increase in group awareness. However, for the $\widehat{\mathcal{R}}_{\text{DP}}^{\text{abs}}$ regularizer a positive association is less often significant than for $\widehat{\mathcal{R}}_{\text{DP}}$. In Figure B.4, we apply *Massaging* preprocessing with a varying fairness parameter. Results on FairFace are presented in Figures B.5 and B.6.

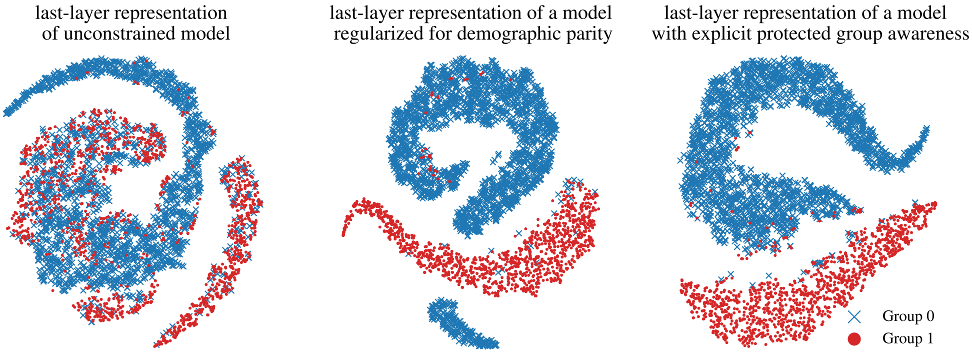


Figure B.1: **tSNE (van der Maaten and Hinton, 2008) visualization of feature representations of unconstrained (left), fairness-regularized with $\widehat{\mathcal{R}}_{\mathrm{DP}}$ (center), and group-aware (Section 3.3) (right) Resnet50 models.** Each point is colored according to the protected attribute MALE, and we aim to classify the binary label ATTRACTIVE. Similar to Figure 3.1 in Chapter 3, we observe that the fair model in the center and the group aware model on the right separate the genders.

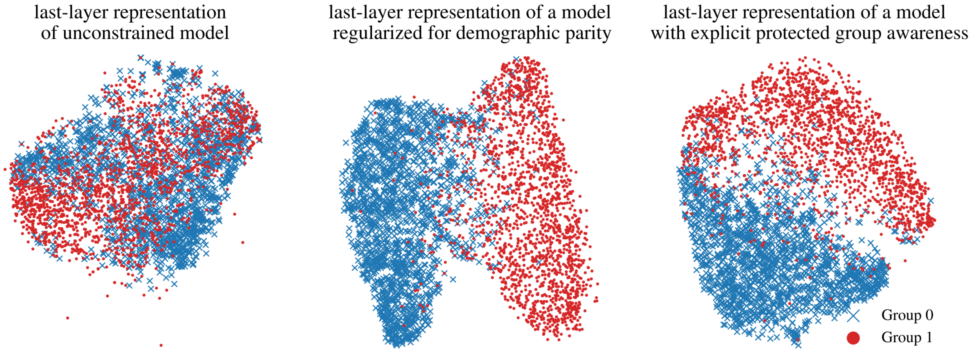
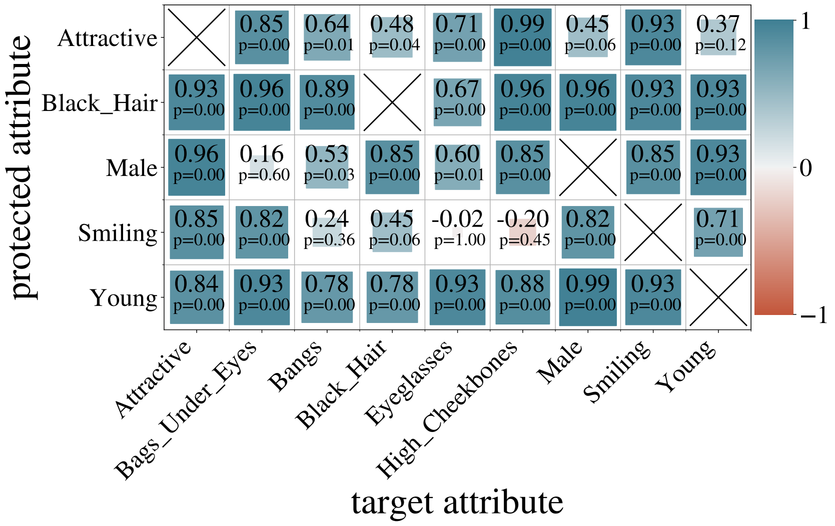
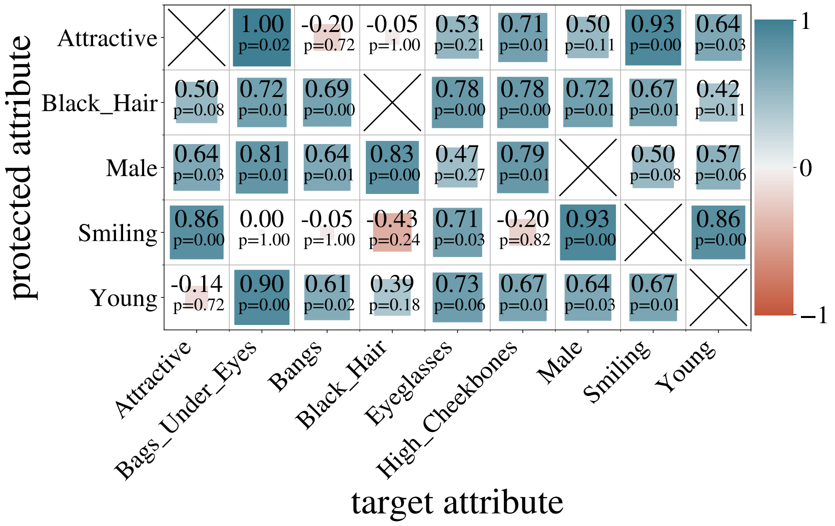


Figure B.2: **tSNE visualization of feature representations of unconstrained (left), fairness-regularized (center), and group-aware (right) Resnet50 models.** In this figure, we use the FairFace dataset. Each point is colored according to the protected attribute GENDER, and we classify the binary label BELOW_30. Similar to CelebA, we observe that gender is separated into disjoint clusters in fair and group aware models, whereas they were mixed in the unconstrained model.

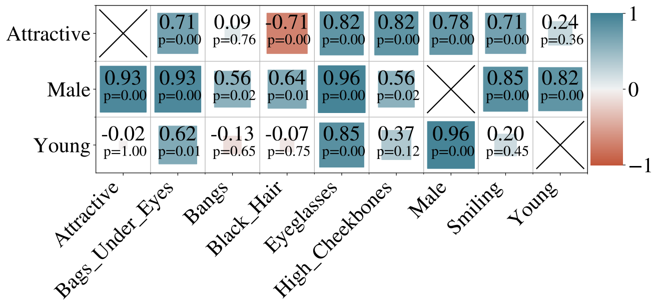


(a) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}_{\text{DP}}$ on CelebA.



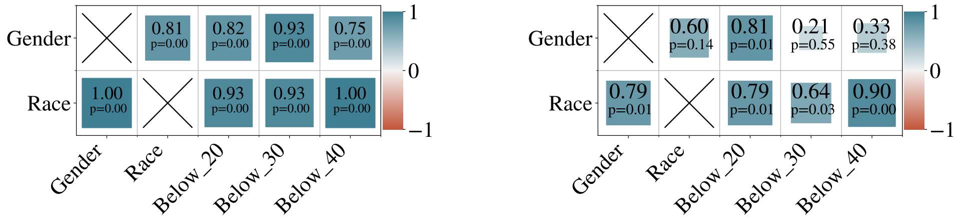
(b) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}^{\text{abs}}_{\text{DP}}$ on CelebA.

Figure B.3: **Kendall-tau correlation between fairness parameter and protected attribute accuracy.** Similar to the results in Chapter 3, where ResNet50 was used, we also find for MobileNetV3-Small that group awareness is increasing as the fairness parameter is increased. In (b) we evaluate the regularizer $\widehat{\mathcal{R}}^{\text{abs}}_{\text{DP}}$ and, although on fewer tasks, observe a similar behavior.



(a) MobileNetV3-Small with *Massaging* preprocessing on CelebA.

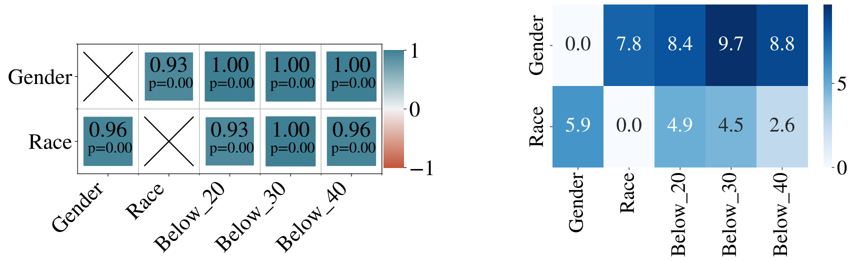
Figure B.4: **Kendall-tau correlation between fairness parameter $\lambda$ and protected attribute accuracy.** Similar to the regularized approaches, we find an increased group awareness for the *Massaging* preprocessing method, especially when the protected attribute is MALE.



(a) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ on FairFace.

(b) MobileNetV3-Small with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}^{\mathrm{abs}}$ on FairFace.

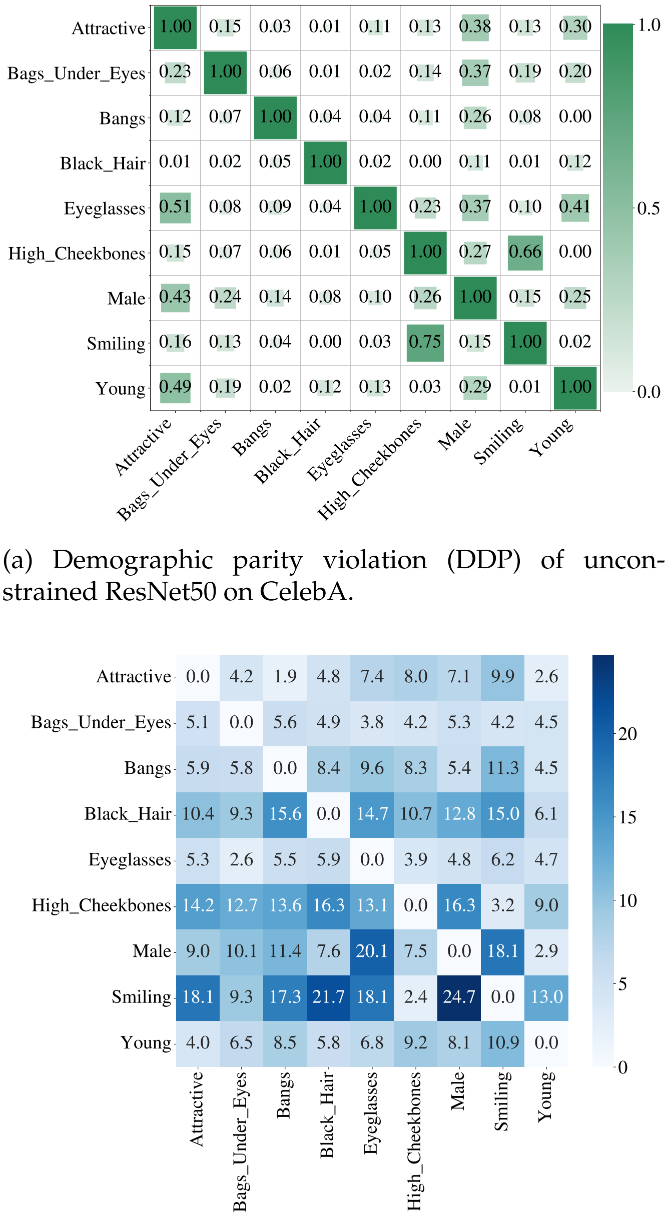
Figure B.5: **Kendall-tau correlation between fairness parameter $\lambda$ and protected attribute accuracy.** Similar to the findings on the CelebA dataset, we also find an increased group awareness on FairFace for the protected attributes RACE and GENDER.



(a) ResNet50 with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ on FairFace.

(b) ResNet50 with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ on FairFace.

Figure B.6: (Left) **Kendall-tau correlation between fairness parameter $\lambda$ and protected attribute accuracy.** (Right) **Increase of protected attribute accuracy** of the group classifier learned on the last layer of ResNet50.



(a) Demographic parity violation (DDP) of unconstrained ResNet50 on CelebA.

(b) Maximum increase of protected attribute accuracy when training ResNet50 with regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ on CelebA.

Figure B.7: (Top) **Demographic parity violation (DDP) of the unconstrained classifier.** The increase in group awareness is more moderate for those tasks where the unconstrained classifier is very unfair, for example for the task (column) MALE. However, this is not always the case as for protected attribute SMILING and target BANGS for example. (Bottom) **Maximum increase of protected attribute accuracy.** Compared to the unconstrained model, we show the highest difference to the second head accuracy of fair models. Even though the unconstrained model is fair, for example for a few target tasks when protected attribute is SMILING, the increase in second head accuracy can still be large.

## B.2 Explicit Two-headed Approach.

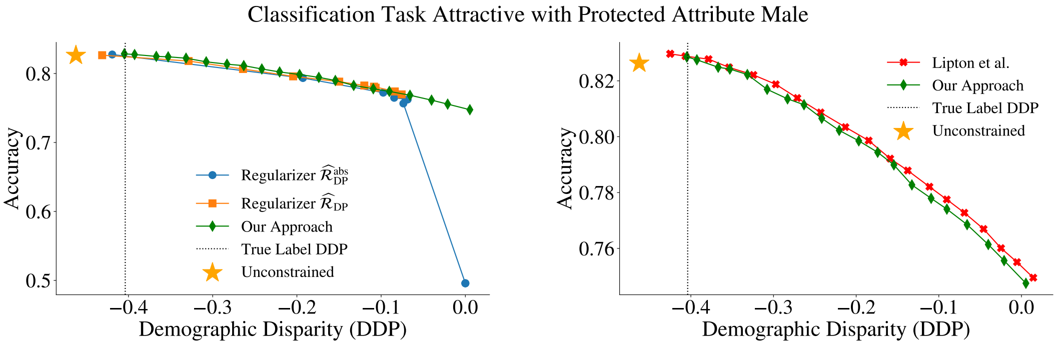


Figure B.8: **Comparison of different fairness approaches using the MobileNetV3-Small architecture.** We compare our group aware model to fairness-regularized models (*left plot*) and the approach of Lipton et al. (2018) (*right plot*) on when predicting the target ATTRACTIVE with respect to the protected attribute MALE. For all methods, we observe the typical trade-off: as the model becomes fairer (DDP is closer to 0), the target accuracy for ATTRACTIVE decreases. All methods obtain similar accuracy for a particular DDP value. However, the regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ is unable to achieve near perfect fairness and saturates around a DDP value of $-0.1$. The regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}^{\mathrm{abs}}$ collapses to a trivial fair solution. Note that Lipton et al. (2018) requires the protected attribute at test time, while we infer the protected attribute.

Table B.1: **Accuracy under strict fairness constraints.** The first block requires a reduction of the absolute value of the DDP of at least 50%, the second at least 80%. The protected attribute is MALE from CelebA dataset and we use the MobileNetV3-Small architecture. Crosses indicate that the method did not achieve the required fairness. Similar to the main paper, the regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ often fails to find sufficiently fair solution. The regularizer $\widehat{\mathcal{R}}^{\mathrm{abs}}_{\mathrm{DP}}$ always finds fair solutions, however, at high costs in accuracy, often resulting in trivial solutions. **Our explicit two-headed approach can always find a fair solution** and is comparable to Lipton, which, contrary to us, requires the true protected attribute.

| | Attractive | Bags_Under_Eyes | Bangs | Black_Hair | Eyeglasses | High_Cheekbones | Smiling | Young |
|---|---|---|---|---|---|---|---|---|
| | | | | 50% disparity reduction | | | | |
| Lipton | 0.8034 | 0.8441 | 0.9473 | 0.9001 | 0.9658 | 0.8609 | 0.9200 | 0.8773 |
| Our Approach | 0.8023 | 0.8439 | 0.9445 | 0.8930 | 0.9773 | 0.8613 | 0.9212 | 0.8762 |
| Massaging | 0.7986 | 0.8424 | 0.9396 | 0.9012 | 0.9661 | 0.8572 | 0.9135 | 0.8520 |
| Regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ | 0.7959 | 0.8446 | ✗ | 0.8993 | ✗ | 0.8603 | ✗ | 0.8710 |
| Regularizer $\widehat{\mathcal{R}}^{\mathrm{abs}}_{\mathrm{DP}}$ | 0.7935 | 0.8311 | 0.8443 | 0.8962 | 0.9545 | 0.8609 | 0.4997 | 0.8728 |
| | | | | 80% disparity reduction | | | | |
| Lipton | 0.7775 | 0.8352 | 0.9331 | 0.9001 | 0.9618 | 0.8403 | 0.9106 | 0.8606 |
| Our Approach | 0.7741 | 0.8347 | 0.9310 | 0.8921 | 0.9652 | 0.8443 | 0.9105 | 0.8580 |
| Massaging | 0.7612 | 0.8204 | ✗ | 0.8947 | ✗ | ✗ | ✗ | 0.8468 |
| Regularizer $\widehat{\mathcal{R}}_{\mathrm{DP}}$ | 0.7740 | 0.8294 | ✗ | 0.8989 | ✗ | ✗ | ✗ | 0.8562 |
| Regularizer $\widehat{\mathcal{R}}^{\mathrm{abs}}_{\mathrm{DP}}$ | 0.7693 | 0.8311 | 0.8443 | 0.8962 | 0.9407 | 0.5182 | 0.4997 | 0.8307 |

## B.3 Fair Networks Behave like Explicit Approach.

In this section, we conduct the experiments from Section 3.4.1 on other tasks and computer vision models. Predicting SMILING we use our two-headed approach to reconstruct fair models and recover the unconstrained model using a ResNet50 with $\widehat{\mathcal{R}}_{\mathrm{DP}}$ regularizer (Figure B.9), using a MobileNetV3-Small with $\widehat{\mathcal{R}}_{\mathrm{DP}}$ regularizer (Figure B.10), and using a MobileNetV3-Small with *Massaging* preprocessing (Figure B.11). In Figure B.12 and B.13, we recover the unconstrained model from fair ResNet50 and MobileNetV3-Small models predicting either ATTRACTIVE or YOUNG. Overall, we are able to replicate the behavior of fair models using both heads of our explicit approach and to recover the unconstrained model from a fair model with the group classifier head. Sometimes, as observed in Figure B.12 the unconstrained classifier cannot be recovered from the fairest models within the performance of the random baseline.

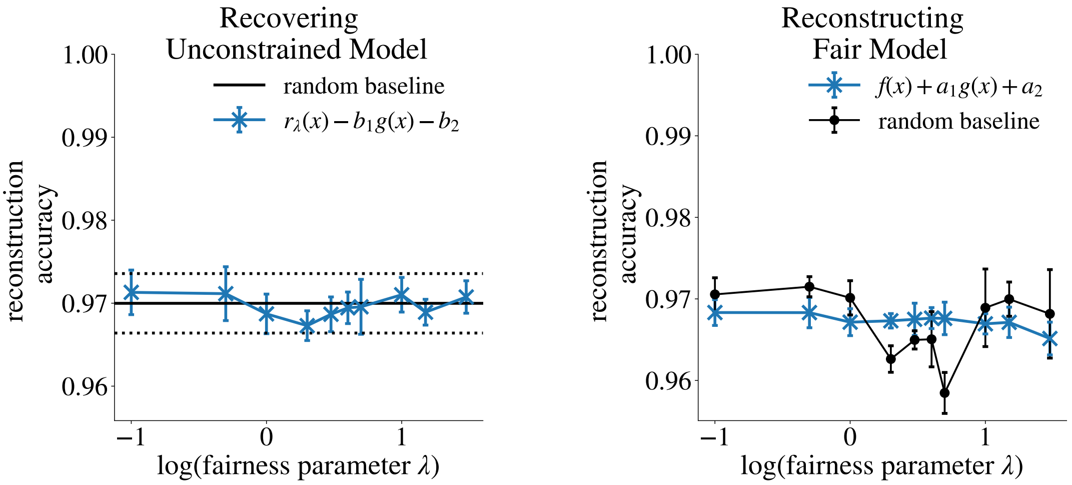


Figure B.9: **Recovering the unconstrained classifier and reconstructing fair classifiers.** We train ResNet50 models with the $\widehat{\mathcal{R}}_{\mathrm{DP}}$ regularizer for the target SMILING and protected attribute MALE. Again, we can reconstruct fair models with our two-headed approach and we can recover the unconstrained model by adding the second head to the fair model.

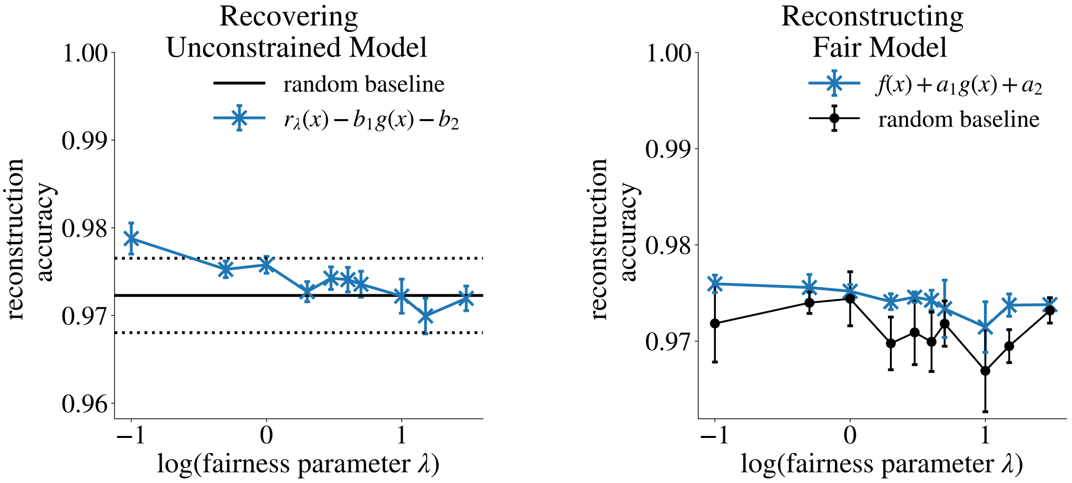


Figure B.10: **Recovering the unconstrained classifier and reconstructing fair classifiers.** We train MobileNetV3-Small models with the $\widehat{\mathcal{R}}_{\mathrm{DP}}$ regularizer for the target SMILING and protected attribute MALE. Similarly to the analysis above with a ResNet50, our observations hold for MobileNetV3-Small models as well.

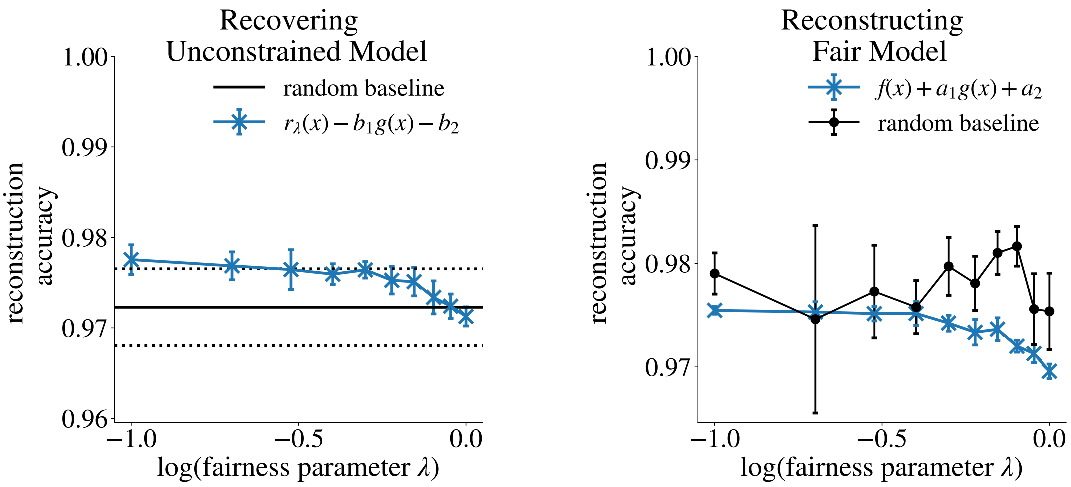
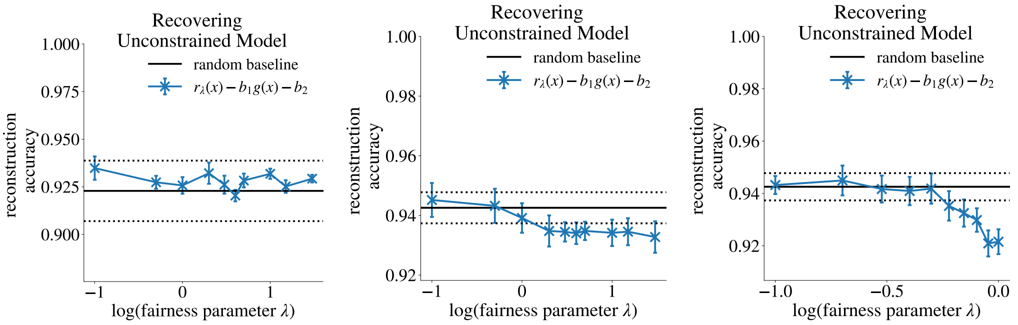


Figure B.11: **Recovering the unconstrained classifier and reconstructing fair classifiers.** We train MobileNetV3-Small models with the *Massaging* preprocessing for the target SMILING and protected attribute MALE. When using *Massaging*, we can reconstruct the resulting fair models with the two-headed approach. However, the two-headed approach reconstructs the most fair models slightly less accurately than a retrained fair model.

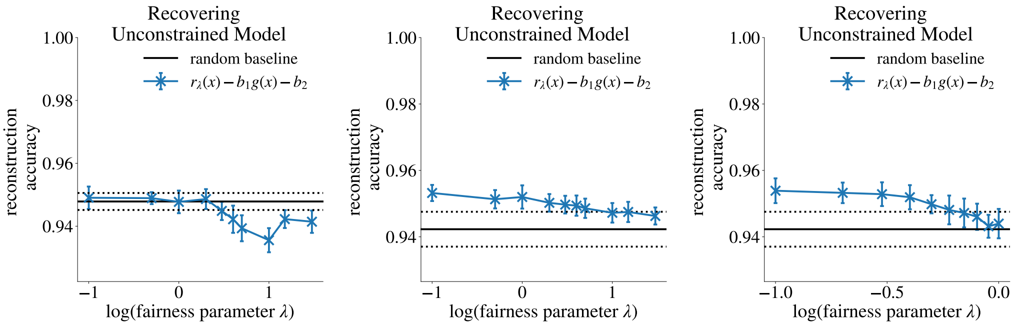


(a) ResNet50 with $\widehat{\mathcal{R}}_{\mathrm{DP}}$.

(b) MobileNetV3-Small with $\widehat{\mathcal{R}}_{\mathrm{DP}}$.

(c) MobileNetV3-Small with *Massaging*.

Figure B.12: **Recovering the unconstrained classifier.** For different models and fairness approaches for the target ATTRACTIVE and protected attribute MALE, we evaluate how our two-headed approach can reproduce the behavior of the fair model. From regularized ResNet50 models we can recover the unconstrained model well. From fair regularized or massaged MobileNetV3-Small models we recover the unconstrained model slightly worse than a retrained unconstrained model.



(a) ResNet50 with $\widehat{\mathcal{R}}_{\mathrm{DP}}$.

(b) MobileNetV3-Small with $\widehat{\mathcal{R}}_{\mathrm{DP}}$.

(c) MobileNetV3-Small with *Massaging*.

Figure B.13: **Evaluating reconstruction accuracy of our two-headed approach.** For different models and fairness approaches for the target YOUNG and protected attribute MALE, we evaluate how identical our two-headed approach is. In this example, we can recover the unconstrained model from fair models using the second head well. The accuracy for fair ResNet50 models is below the random baseline.